University of Chester

University of Chester

Proceedings of the 19th European Conference on Cyber Warfare and Security

a Virtual Conference hosted by
University of Chester
UK
25-26 June 2020

Edited by
Dr Thaddeus Eze, Dr Lee Speakman and
Dr Cyril Onwubiko

acpi

A conference managed by ACPI, UK

Proceedings of the

19th European Conference on Cyber Warfare and Security

ECCWS 2020

Hosted By

The University of Chester, UK

25-26 June, 2020

Edited by

Dr Thaddeus Eze, Dr Lee Speakman
Dr Cyril Onwubiko

Review Process

Papers submitted to this conference have been double-blind peer reviewed before final acceptance to the conference. Initially, abstracts were reviewed for relevance and accessibility and successful authors were invited to submit full papers. Many thanks to the reviewers who helped ensure the quality of all the submissions.

Ethics and Publication Malpractice Policy

ACPIL adheres to a strict ethics and publication malpractice policy for all publications – details of which can be found here: http://www.academic-conferences.org/policies/ethics-policy-for-publishing-in-the-conference-proceedings-of-academic-conferences-and-publishing-international-limited/

Self-Archiving and Paper Repositories

We actively encourage authors of papers in ACPIL conference proceedings and journals to upload their published papers to university repositories and research bodies such as ResearchGate and Academic.edu. Full reference to the original publication should be provided.

Conference Proceedings

The Conference Proceedings is a book published with an ISBN and ISSN. The proceedings have been submitted to a number of accreditation, citation and indexing bodies including Thomson ISI Web of Science and Elsevier Scopus.

Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

The Electronic version of the Conference Proceedings is available to download from DROPBOX https://tinyurl.com/eccws20 Select Download and then Direct Download to access the Pdf file. Free download is available for conference participants for a period of 2 weeks after the conference.

The Conference Proceedings for this year and previous years can be purchased from http://academic-bookshop.com

E-Book ISBN: 978-1-912764-62-4
E-Book ISSN: 2048-8610
Book version ISBN: 978-1-912764-61-7
Book Version ISSN: 2048-8602

Published by Academic Conferences and Publishing International Limited
Reading, UK
+44-118-972-4148
www.academic-conferences.org
info@academic-conferences.org

Contents

ECRM Preface

These proceedings represent the work of contributors to the 19th European Conference on Cyber Warfare and Security (ECCWS 2020), supported by University of Chester, UK on 25-26 June 2020. The Conference Co-chairs are Dr Thaddeus Eze and Dr Lee Speakman, both from University of Chester and the Programme Chair is Dr Cyril Onwubiko from IEEE and Director, Cyber Security Intelligence at Research Series Limited.

ECCWS is a well-established event on the academic research calendar and now in its 19th year the key aim remains the opportunity for participants to share ideas and meet. The conference was due to be held at University of Chester, UK, but due to the global Covid-19 pandemic it was moved online to be held as a virtual event. The scope of papers will ensure an interesting conference. The subjects covered illustrate the wide range of topics that fall into this important and ever-growing area of research.

The opening keynote presentation is given by Dr Cyril Onwubiko on the topic of A Comprehensive Cyber Recovery Operational Framework. The second day of the conference will open with an address Detective Sergeant Damian Ward of the Metropolitan Police speaking on his Experiences of Policing Cybercrime in the UK: Challenges and Successes.

With an initial submission of 123 abstracts, after the double blind, peer review process there are 50 Academic research papers, 13 PhD research papers, 1 Masters research paper and 2 work-in-progress papers published in these Conference Proceedings. These papers represent research from Algeria, Australia, Bahrain, Brazil, China, Czech Republic Estonia, Finland, Germany, Ireland, Italy, Lithuania, Malaysia, The Netherlands, Norway, Portugal, Romania, South Africa, Sweden, UAE, UK and USA.

We hope you enjoy the conference.

Dr. Thaddeus Eze
University of Chester
UK
June 2020

Conference Committee

Dr. Mohd Faizal Abdollah, University Technical Malaysia Melaka, Malaysia; Dr William ("Joe") Adams, Univ of Michigan/Merit Network, USA; Dr. Tariq Ahamad, Prince Sattam Bin Abdulaziz University, Saudi Arabia; Prof Hamid Alasadi, Basra University, Iraq; Dr. Kari Alenius, University of Oulu, Finland; Prof. Antonios Andreatos, Hellenic Air Force Academy, Greece; Dr. Olga Angelopoulou, University of Hertfordshire, UK; Dr. Leigh Armistead, Edith Cowan University, Australia; Johnnes Arreymbi, University of East London, UK; Dr. Hayretdin Bahsi, Tallinn University of Technology, Estonia; Dr. Darya Bazarkina, Sholokhov Moscow State Humanitarian University, Russia; Mr Robert Bird, Coventry University, UK; Prof. Matt Bishop, University of California at Davis, USA; Dr Radomir Bolgov, Saint Petersburg State University, Russia; Dr. Svet Braynov, University of Illinois at Springfield, USA; Prof. Larisa Breton, FullCircle Communications, LLC, USA; Dr Jim Chen, DoD National Defense University, USA; Bruce Christianson, University of Hertfordshire, UK; Dr. Maura Conway, Dublin City University, Ireland; Dr. Paul Crocker, Universidade de Beira Interior, Portugal; Prof. Tiago Cruz, University of Coimbra, Portugal; Dr. Christian Czosseck, CERT Bundeswehr (German Armed Forces CERT), Germany; Geoffrey Darnton, Bournemouth University, UK; Josef Demergis, University of Macedonia, Greece; Paul Dowland, Edith Cowan University, Australia; Marios Efthymiopoulos, Political Science Department University of Cyprus, Cyprus; Dr. Colin Egan, University of Hertfordshire, Hatfield, UK; Dr Ruben Elamiryan, Public Administration Academy of the Republic of Armenia, Armenia; Prof. Dr. Alptekin Erkollar, ETCOP, Austria; Dr Thaddeus Eze, University of Chester, UK; John Fawcett, University of Cambridge, UK; Prof. Eric Filiol, ENSIBS, Vannes, France & CNAM, Paris, France; Dr. Chris Flaherty, University of New South Wales, Australia; Prof. Steve Furnell, University of Plymouth, UK; Mr. Tushar Gokhale, Hewlett Packard Enterprise, USA; Dr. Michael Grimaila, Air Force Institute of Technology, USA; Prof. Stefanos Gritzalis, University of the Aegean, Greece; Dr. Mils Hills, Northampton Business School, Uk; Ulrike Hugl, University of Innsbruck, Austria; Aki Huhtinen, National Defence College, Finland; Bill Hutchinson, Edith Cowan University, Australia; Dr. Abhaya Induruwa, Canterbury Christ Church University, UK; Dr. Md Ruhul Islam, Sikkim Manipal Institute of Technology, India; Hamid Jahankhani, University of East London, UK; Nor Badrul Anuar Jumaat, University of Malaya, Malaysia; Maria Karyda, University of the Aegean, Greece; Ass. Prof. Vasilis Katos , Democritus University of Thrace, Greece; Dr. Anthony Keane, Technological University Dublin, Ireland; Jyri Kivimaa, Cooperative Cyber Defence and Centre of Excellence, Tallinn, Estonia; Prof. Ahmet Koltuksuz, Yasar University, Dept. of Comp. Eng, Turkey; Theodoros Kostis, Hellenic Army Academy, Greece; Prashant Krishnamurthy, University of Pittsburgh, USA; Mr. Peter Kunz, DoctorBox, Germany; Dr. Erikk Kurkinen, University of Jyväskylä, Finland; Takakazu Kurokawa, National Defence Acadamy, Japan; Rauno Kuusisto, Finnish Defence Force, Finland; Martti Lehto, National Defence University, Finland; Mr Trupil Limbasiya, NIIT University, Neemrana, Rajasthan, India; Dr Efstratios Livanis, University of Macedonia, Greece; Peeter Lorents, CCD COE, Tallinn, Estonia; James Malcolm, University of Hertfordshire, UK; Dr Mary Manjikian, Regent University, USA; Mario Marques Freire, University of Beira Interior, Covilhã, Portugal; Ioannis Mavridis , University of Macedonia, Greece; Rob McCusker, Teeside University, Middlesborough, UK; Dr Imran Memon, zhejiang university, china; Dr Shahzad Memon, University of Sindh, Pakistan; Jean-Pierre Molton Michel, Ministry of Agriculture, Haiti; Dr. Yonathan Mizrachi, University of Haifa, Israel; Dr Pardis Moslemzadeh Tehrani, University of Malaya, Malaysia; Evangelos Moustakas, Middlesex University, London, UK; Antonio Muñoz, University of Málaga , Spain; Daniel Ng, C-PISA/HTCIA, China; Dr. Funminiyi Olajide, Nottingham Trent University, UK; Dr Cyril Onwubiko, Cyber Security Intelligence at Research Series Limited,, UK; Rain Ottis, Tallinn University of Technology, Estonia; Dr Mahmut Ozcan, Webster University, USA; Prof Teresa Pereira, Instituto Politécnico de Viana do Castelo, Portugal; Michael Pilgermann, University of Glamorgan, UK; Dr Bernardi Pranggono, Sheffield Hallam University, UK; Prof Carlos Rabadão, Politechnic of Leiria, Portugal; Dr. Muttukrishnan Rajarajan, City University London, UK; Prof Saripalli Ramanamurthy, Pragati Engineering College, India; Dr Trishana Ramluckan, University of KwaZulu-Nata, South Africa; Dr Aunshul Rege, Temple University, United States; Dr. Neil Rowe, US Naval Postgraduate School, Monterey, USA; Prof Vitor Sa, Catholic University of Portugal, Portugal; Dr. Char Sample, Carnegie Mellon University/CERT, USA; Prof. Henrique Santos, University of Minho, Portugal; Dr Keith Scott, De Montfort University, UK; Prof. Dr. Richard Sethmann, University of Applied Sciences Bremen, Germany; Dr. Yilun Shang, Northumbria University, UK; Prof. Paulo Simoes, University of Coimbra, Portugal; Dr Umesh Kumar Singh, Vikram University, Ujjain, India; Prof. Jill Slay, University of South Australia, Australia; Dr Lee Speakman, University of Chester, UK; Dr Joseph Spring, , UK; Dr Hamed Taherdoost, Hamta Group | Hamta Business Solution Sdn Bhd, Malaysia; Unal Tatar, University at Albany - SUNY, USA; Dr. Selma Tekir, Izmir Institute of Technology, Turkey; Prof. Dr. Peter Trommler, Georg Simon Ohm University Nuremberg, Germany; Prof Tuna USLU, Istanbul Gedik University, Occupational Health and Safety Program, Türkiye; Craig Valli, Edith Cowan UniversitY, Australia; Dr Brett van Niekerk, University of KwaZulu-Natal & Transnet, South Africa; Richard Vaughan, General Dynamics UK Ltd, UK; Dr Namosha Veerasamy, Council for Scientific and Industrial Research, South Africa; Dr

Sangapu Venkata Appaji, KKR & KSR Institute of Technology and Sciences, India; Stilianos Vidalis, School of Computer Science, University of Hertfordshire, UK; Dr. Natarajan Vijayarangan, Tata Consultancy Services Ltd, India; Dr Khan Ferdous Wahid, Airbus Group, Germany; Prof Mat Warren, Deakin University, Australia, Australia; Dr. Santoso Wibowo, Central Queensland University, Australia; Prof. Trish Williams, Flinders University, Australia; Prof Richard Wilson, Towson University, USA; Simos Xenitellis, Royal Holloway University, London, UK.

Biographies

Conference co-Chairs

Dr Thaddeus Eze is a Senior Lecturer in Cyber Security at the University of Chester. He is the founder and convener of the IEEE UK & Ireland YP Postgraduate STEM Research Symposium, Vice Chair, IEEE UK & Ireland Young Professionals, a technical committee member for a number of international conferences (e.g., ICAS, CYBERWORLDS, EMERGING etc.), and currently runs the Computer Science departmental research seminar series. His research interests include Trustworthy Autonomics, MANET and Cyber Security (specifically, Return Oriented Programming, Policing the Cyber Threat and Cyber Education) and he has a number of publications in these areas.

Dr Lee Speakman is a Senior Lecturer and Programme Lead for Cybersecurity at the University of Chester. Lee has worked in Defence since 1999. He has gained experience in Intelligence, Strategy, Electronic Warfare, Communications, and Networking. He gained his PhD in MANETs from Niigata University, Japan, in 2009, and returned to Defence to work in Cyber and Information Security, and related areas. He joined the University of Chester in 2015 as the Programme Leader to develop and deliver Cybersecurity courses. His research interests are in software and system security.

Programme Chair and Keynote Speaker

Dr Cyril Onwubiko is the Secretary, IEEE UK & Ireland, Chair, IEEE UK & Ireland Blockchain Group, and Director, Cyber Security Intelligence at Research Series Limited, where he is responsible for directing strategy, IA governance and cyber security. Prior to Research Series, he had worked in the Financial Services, Telecommunication, Health sector and Government and Public services Sectors. He is a leading scholar in Cyber Situational Awareness (Cyber SA), Cyber Security, Security Information and Event Management (SIEM), Data Fusion & SOC; and interests in Blockchain and Machine Learning. He is the founder of the Centre for Multidisciplinary Research, Innovation & Collaboration (C-MRiC) https://www.c-mric.com. Detailed profile for Cyril can be found on https://www.c-mric.com/cyril

Keynote Speaker

Damian Ward is a Detective Sergeant with the Metropolitan Police Service. Damian has spent 21 years in law enforcement; principally as a Detective investigating homicide, counter terrorism, organised crime groups and cyber crime. Recently he has spent several years within national fraud and cyber investigations, on both reactive and prevention teams. In 2016 he completed a BSc at Canterbury Christchurch University, reviewing some of the challenges for law enforcement when responding to cyber crime. He has received commendations for his management of body identification at disasters such as the MH17 air crash in the Ukraine and the Grenfell Tower fire. Currently Damian works within counter terrorism investigations in the UK.

Mini-Track Chairs

Nasser S. Abouzakhar is a senior lecturer at the University of Hertfordshire, UK. Currently, his research area is mainly focused on critical infrastructure security, cloud security and applying machine learning solutions to various Internet and Web security and forensics related problems. He received PhD in Computer Sci Engg in 2004 from the University of Sheffield, UK. He is a technical studio guest to various BBC World Service Programmes such as Arabic 4Tech show, News-hour programme and Breakfast radio programme. Nasser is a BCS chartered IT professional (CITP), CEng and CSci and is a BCS assessor for the accreditation of Higher Education Institutions (HEIs) in the UK.

Dr. Sangapu Venkata Appaji is working as a Professor, at KKR and KSR Institute of Technology and Sciences, Guntur, India. He completed his doctorate in Computer science and Engineering in the area of Cryptography and Network Security from Jawaharlal Nehru Technological University Hyderabad, INDIA. He has more than 10 years of teaching experience in Computer Science and engineering and Information Technology department for graduate and post graduate students. He supervised IOT, Security related projects for graduate and postgraduate students.

Dr. Shahzad Memon is working as a Professor, at Institute of Information and Communication Technology at University of Sindh, Pakistan. He completed his doctorate in Biometrics fingerprint sensors technology from Brunel University, London, UK. Dr. Memon has published and supervised MPhil and PhD students in his main field of research Cyber Security, Privacy issues in IoT and Social Media, Smarts systems and ICT applications. He also granted funding for research from Higher Education commission and ICT Research and Development, Ministry of Information Technology, Pakistan.

Brett van Niekerk is a senior lecturer in computer science at the University of KwaZulu-Natal. He serves as secretary for the International Federation of Information Processing Working Group on ICT in Peace and War, and the co-Editor-in-Chief of the International Journal of Cyber Warfare and Terrorism. He has numerous years of information/cyber-security experience in both academia and industry, and has contributed to the ISO/IEC information security standards. In 2012 he graduated with his PhD focusing on information operations and critical infrastructure protection. He is also holds a MSC in electronic engineering and is CISM certified.

Trishana Ramluckan has been an academic and researcher in the IT and governance fields for the past 12 years. She is a Postdoctoral Researcher in the College of Law and Management Studies at the University of KwaZulu-Natal. She is a member of the IFIP working group on ICT Uses in Peace and War, the Institute of Information Technology Professionals South Africa and serves as an Academic Advocate for ISACA. In 2017 she graduated with a Doctor of Administration specialising in IT and Public Governance. Her current research areas include Cyber Law and Information Technology Governance

Dr. Char Sample is the Chief Cybersecurity Research Scientist for the Cybercore division at Idaho National Laboratory. Dr. Sample is a visiting academic at the University of Warwick, Coventry, UK and a guest lecturer at Bournemouth University, Rensselaer Polytechnic University and Royal Holloway University. Dr. Sample has over 20 years' experience in the information security industry. Dr. Sample's research focuses on deception, and the role of cultural values in cybersecurity events. More recently she has begun researching the relationship between human cognition and machines. Presently Dr. Sample is continuing research on modeling cyber behaviors by culture, other areas of research are data resilience, cyber-physical systems and industrial control systems.

Pardis Moslemzadeh Tehrani is a senior lecturer at the Faculty of Law, University of Malaya. Her research interests lie in the areas of cyberterrorism, cyberlaw, and international humanitarian law. Pardis's research has been widely published in peer-reviewed journals and she has presented papers at national and international level conferences. She is a member of the editorial review board in several journals. She is also an international scientific member of the Australian and New Zealand Society of International Law.

Richard L. Wilson teaches in the Philosophy and Computer and Information Sciences at Towson University in Towson, MD and is a Research Fellow in the Hoffberger Center for Professional Ethics at the University of Baltimore. Professor Wilson has taught a wide variety of Applied Ethics courses including Engineering Ethics, Computer Science Ethics, Medical Ethics, Environmental Ethics and Business Ethics, and has a wide variety of publications in all of these areas. Previous experience also includes being an Ethics for the Department of Justice for the United States Government

Author Biographies

Pascal Ahr is a Bachelor of Science and Masters student of Electrical and Computer Engineering in Embedded Systems at University of Kaiserslautern (TUK) Germany. He works as research assistant at the German Research Center for Artificial Intelligence (DFKI). His research focusses on the field of Hardware - Physical Layer Security (PhySec) and Physically Unclonable Functions (PUFs).

Khalil Akbariavaz is a phd student at the Faculty of Law, University of Malaya. Research interests include public international law, Human rights, International Humanitarian Law, Rights of civilians, Rights of Children and Women and Middle East Politics. Khalil has presented papers in national and international Conferences. He has translated and edited a number of articles and books from Arabic and English to Persian.

Francis Xavier Kofi Akotoye is a PhD student at the University of Pretoria, South Africa. He received his MEng in Computer Science and Technology from Hunan University, China in 2007. His research interest are in digital forensic, computer security and mobile computing.

Hisham Al-Assam is senior lecturer in Computer Science, the University of Buckingham. He teaches Introduction to Computer Systems, Web Applications Development and Information Security,Web Technologies and Applications, and Information Security in Communications Key research interest is machine learning in security and healthcare. Security applications include deep learning for biometric recognition, dynamic signatures, privacy-aware biometric template security, and multi-factor remote authentication.

Ossama Al-Maliki is a Ph.D. research student at the University of Buckingham. Research focuses on the EMV contactless card specifications and improving the security of the EMV payment protocol. His M.Sc. degree was finished at the Information Technology college at the UNITEN University, Malaysia in 2012. His bachelor's degree was done at the computer science college, university of Basra, Iraq in 2008.

Victor Emmanuell BADEA has graduated the Faculty of Aerospace Engineering – Propulsion Systems in 2015. In 2017 he has graduated the Aerospace Propulsion and Environmental Protection – Master Program at the same Faculty of Aerospace Engineering. In 2018 he joined the staff of the National Institute for Research and Development in Informatics, as aerospace engineer.

Olimzhon Baimuratov, Associate Professor of Suleyman Demirel University, Doctor PhD. Research interests are mainly in the area of Control Systems, Cyber Security, Education, Medicine and mathematical modeling. Also participates as an expert in the Project management and Development. Today has over 40 journal publications, 35 conference papers, and four book chapters.

Jorge Augusto Castro Neves Barbosa, PhD & MSc degree in Electrical and Computer Engineering, Technical University of Lisbon, Portugal; Former President of Coimbra Polytechnic – ISEC; Coordinator Professor, Coimbra Polytechnic - SEC; Auditor of the National Defense Course, National Defense Institute (IDN); Auditor of the Cybersecurity Course and Crisis Management in Cyberspace, IDN; Auditor of the Civil Crisis Management Course, IDN

Stacey Omeleze Baror obtained her B.sc Hon and Masters Computer Sci., from the Department of Computer Science, University of Pretoria. She is currently studying towards a Doctorate with the Computer Science, University of Pretoria. Her research areas are focused on Cloud computing, Digital forensic readiness, Cloud forensics, Cybercrime, Machine learning, Robotics, and Natural language processing techniques for cybercrime and Information Security.

Vijay Bhuse is an Assistant Professor of Computer Science at the Grand Valley State University. He is received his Ph.D. from Western Michigan University in 2007 and completed his postdoctoral fellowship from the Dartmouth College. He worked in industry before returning to academia. His research interests are Network Security, Wireless Sensor Networks and Secure Coding.

Steve Blenkin is a recent graduate of the University of South Wales where he obtained an MSc in Cyber Security. Steve has a keen interest in digital forensics and security.

Brett van Niekerk is a senior lecturer in computer science at the University of KwaZulu-Natal. He serves as chair for the IFIP Working Group on ICT in Peace and War, and the co-Editor-in-Chief of the International Journal of Cyber Warfare and Terrorism. He holds a PhD focusing on information operations and critical infrastructure protection.

Ian Bryant is and Adjunct Professor in the Cyber Security Centre (CSC) at the University of Warwick. In his other roles he is a Professional Engineer focusing on Information and Info-Cyber Systems (ICyS) protection, is heavily involved with various Standards Development Organisations (SDO), and is the Secretary of the UK’s Advisory Committee on Trustworthy Systems (ACTS).

Prof. Ladislav BURITA, University of Defence, Brno, Czech Republic. PhD. from 1985 and professor from 2003. He worked in NATO/MIP team; was head of the faculty research program and defense project MENTAL. His area of interest are information and knowledge systems. He published hundred papers at conferences and in journals.

Ivan Burke is a senior cyber security research for the Council for Scientific and Industrial Research (CSIR) in South Africa. His research topics include network security and building defensive sensors for national cyber infrastructure. Ivan is currently pursuing his Masters in Computer Security at Rhodes University.

Peter Chan is a cybersecurity engineer at the Council for Scientific and Industrial Research (CSIR) in South Africa. He focuses in enterprise security, computer incident response and ontology engineering. He is a certified security practitioner (CASP+) and team member of a FIRST accredited Computer Incident Response Team (CSIRT).

Tiago Cruz received the Ph.D. degree in informatics engineering from the University of Coimbra in 2012, where he has been Assistant Professor with the Department of Informatics Engineering, since 2013. Research interests include management systems for communications infrastructures and services, critical infrastructure security, broadband access network device and service management, Internet of Things, software defined networking, and network function virtualization.

Dr Didier Danet is a senior lecturer in Military Academy of Saint-Cyr (France), "Mutation of Conflicts" Research Department. Didier DANET is Head of Post-Master Degree In Cyber Operations and Crisis Management.

Lisa Davidson is a Signals Officer within the Australian Army. Lisa is undertaking a Ph.D at the University of New South Wales. Her research focus is designing a holistic Cyber workforce for the Australian Army based on a Middle Power approach.

Gareth Davies is a Senior Lecturer in Digital Forensics and Cyber Security at the University of South Wales. As Chairman of The First Forensic Forum (F3), Gareth has organised international conferences and regular yearly workshops to demonstrate new digital forensics research and cutting edge technologies to the law enforcement community.

Arno de Coning is a Systems Engineer in Facilities management at the University of Pretoria. He received his M.Eng in 2013 for Computer and Electronic engineering and is working towards his PhD in the same field. His main research is in optimising processes by implementing data driven system designs.

Arnold Dupuy is an Assistant Professor at the Virginia Polytechnic Institute and State University in Blacksburg, VA. He is also the Chair of the Energy Security in the Era of Hybrid Warfare project of the NATO Science & Technology Organization.

Dr. Petrus Duvenage served as an officer in the South African armed forces and in the intelligence services. He holds a Senior Research Fellowship at the Academy for Computer Science and Software Engineering, University of Johannesburg. He has PhD from the University of Pretoria and a PhD (D.Com) from the University of Johannesburg.

Peter Eden is a Lecturer in Digital Forensics and Cyber Security at the University of South Wales. He is Course Leader for BSc/MSc Computer Security courses. He has provided digital forensic consultancy services to UK law enforcement agencies and police forces and has a number of publications within SCADA forensics.

Dr. Jan Fesl is professor assistant at the Institute of Applied Informatics (University of South Bohemia in České Budějovice, Czech Republic). He received his PhD in informatics from Czech Technical University in Prague in 2018. His main areas of research are computer networks, computer networks security and application of artificial intelligence methods for computer networks, and distributed systems.

Tim Grant is retired but an active researcher (Professor emeritus, Netherlands Defence Academy). Tim has a BSc in Aeronautical Engineering (Bristol University), a Masters-level Defence Fellowship (Brunel University), and a PhD in Artificial Intelligence (Maastricht University). Tim's research focuses on offensive cyber operations and on Command & Control and Emergency Management systems. More details can be found at https://www.linkedin.com/in/tim-grant-r-bar/.

Ferdinand J. Haberl is a doctoral candidate at the Department of Oriental Studies at the University of Vienna, where he focuses on jihadism worldwide, cyberterrorism, and jihadi (counter-) intelligence activities. He he has furthermore engaged in terrorism research in the Middle East, North and East Africa and Southeast Asia.

Saïd Haddad, Ph.D in Political science (René Descartes University, Paris), is Senior lecturer in Sociology and member of the research team, Conflits in Mutation of the Saint-Cyr Research Center at the Saint-Cyr Military Academy, France. His current research focuses on the construction of cyber as a French national priority and the sociology of "cyber warriors".

Clay Hampton is a Research associate and adjunct professor at IUPUI. With a background in networking and Information security. I enjoy continuously learning about new technologies and how cybersecurity is an important field to secure systems and information to ensure we are able to stay safe.

Jakub Harašta is an Assistant Professor at the Institute of Law and Technology at the Faculty of Law, Masaryk University, Brno, CZ. He is the Editor-in-Chief of the Masaryk University Journal of Law and Technology. Jakub holds master's degree in law (2013), Ph.D. in ICT law (2018) and master's degree in security studies (2020).

Michael Hegarty is a lecturer in the Technological University – Dublin. Michael is based in the Blanchardstown Campus where he delivers modules in Digital Forensics and Business. He holds an MSc in Strategic Management of ICT (Network Security) and is currently completing a PhD in the area of Digital Steganalysis and Artificial Intelligence. Michael has published a book "Steganography, The World of Secret Communications".

Jukka Heikkonen has been a professor of computer science of University of Turku, Finland, since 2009. His current research as the head of the Algorithms and Computational Intelligent (ACI) research group is related to data analytics, machine learning and autonomous systems. He has worked at top level research laboratories and Center of Excellences in Finland and international organizations (European Commission, Japan) and has led many international and national research projects. He has authored more than 150 scientific articles.

Malcolm Hodgson studied his bachelors of commerce degree at Stellenbosch University in South Africa. Following that he did his postgraduate degree in Information Systems Management at Stellenbosch University. Malcolm currently works at BDO in Cape Town as a financial services technology junior analyst and has a passion for cyber security and business process improvement.

Aki-Mauri Huhtinen, (LTC (GS), PhD) is a military professor at the Finnish National Defence University in the Department of Leadership and Military Pedagogy. His areas of expertise are military leadership, command and control, information warfare, the philosophy of science in military organizational research and the philosophy of war. He has published peer-reviewed journal articles, a book chapter and books on information warfare and non-kinetic influence in the battle space.

Aarne Hummelholm, PhD in Information Technology (University of Jyväskylä, 2019) has been involved in the design, development of architectures` of authorities` telecommunications networks and information systems. Key themes in his work have been critical service availability, cyber security and preparedness issues.

Gazmend Huskaj is a doctoral student in Cyberspace Operations at the Swedish Defence University. Was Director Intelligence in the Swedish Armed Forces on cyber-related issues. He graduated from Harvard Kennedy School in Cybersecurity: The Intersection of Policy and Technology, has an MSc in Information Security from Stockholm University and an MSc in Security and Risk Management from the University of Leicester. He is ISACA Certified Information Security Manager (CISM).

Professor Bill Hutchinson was Foundation IBM Chair in Information Security at Edith Cowan University, in Western Australia. He was Director of SECAU (Security Research Centre) and was coordinator of the Information Operations and Security programmes. From 2000 to 2010, he was the Chief Editor and founder of the Journal of Information Warfare.

Mika Huttunen, Lt. Col, Doctor of Military Sciences is working in the Finnish Defence Research Agency, Concepts and Doctrines Division. Dr. Huttunen has made his career in the Finnish Defence since 1986, and defended his doctoral thesis on the different theories of warfare in different countries in 2010.

Louis Hyman studied his bachelors of commerce degree at Stellenbosch University in South Africa. Following that he did his postgraduate degree in Information Systems Management at Stellenbosch University. Louis works at Entelect in Johannesburg as a junior software engineer and is passionate about development and security.

Ion Iftimie is the Eisenhower PhD Fellow at the NATO Defense College in Rome, Italy and Senior Advisor at the European Union Research Center of the George Washington School of Business in Washington, D.C.

Eduardo Arthur Izycki International Relations M.A. Student at the University of Brasília (UnB) and public servant. Eduardo Izycki worked on developing solutions for risk assessments in the cycle of major events in Brazil (2012/2016). He currently works in the Critical Infrastructure Protection Coordination of the Brazilian Institutional Security Office (GSI).

Petri Jääskeläinen is a masters student from the Faculty of Information Technology and Communication Sciences in Tampere University. He also works as a freelance journalist specialized in radicalized movements, disinformation, conspiracy theories and technology.

Yahlieel Jafta a Software Developer and has been working professionally since 2012. Areas of interest include Software Design patterns and Test Driven Development with a keen interest in Domain Driven Development and Artificial Intelligence. He holds a BSc in Computer Science and is currently completing his Computer Science Honours degree at the University of the Western Cape.

Dr. Victor J Jaquire has been within the field of cyber and information security for over 20 years within government and private sector focusing on strategy, performance management and operations. He holds d PhD in Informatics from the University of Johannesburg - specialising in strategies for cyber counterintelligence maturity and the security of cyberspace. His professional certifications include CISSP, CISM and CCISO.

Jiri Jelinek Ph.D. (Czech Technical University in Prague). He worked a long time at the University of Economics, Prague. Since 2011 he is a professor assistant in the Institute of Applied Informatics of Faculty of Science at the University of South Bohemia in České Budějovice. Professional interests include distributed artificial intelligence, neural networks, multi-agent systems, social networks, and simulation models.

Juozapavičius holds a PhD in theoretical physics from KTH Royal Institute of Technology, Sweden. He leads the Department of Defence Technologies at General Jonas Žemaitis Military Academy of Lithuania. His research interests are cybersecurity and computer modelling. He participates in EU-funded cybersecurity-related projects, and he is responsible for the cybersecurity specialisation of the study programs

Dr. Connie Justice has over 30 years' experience in cybersecurity, computer, and systems engineering. She designed courses in cybersecurity curriculum to NSA/DHS Center of Academic Excellence and NIST National Initiative for Cybersecurity Education standards. Research areas include: misinformation, industrial controls risk, experiential learning, information and security risk management, digital forensics.

Lt.Col Harry Kantola conducts research at the Finnish National Defence University, Helsinki. Also currently appointed to the Finnish Army Signal School as commandant.. He served in various capacities in the Finnish Navy, Armoured Signal Coy, and Armoured Brigade from 1991. Also served as a researcher at the NATO Cooperative Cyber Defence Centre of Excellence,Tallinn, Estonia and Chief of Cyber Division in C5 Agency.

PhD Martti J Kari is university teacher of cyber security, hybrid threats and strategic intelligence in Jyväskylä University, Finland. He retired as colonel from Finnish Defense Intelligence in the end of year 2017. His last post in military was Assistant Chief of Defense Intelligence. He has MA in Russian language (1993) and literature, MA in cyber security (2017), and PhD in cyber security 2019 in Jyväskylä University. The topic of his PhD thesis was Russian cyber threat perception.

Mr. Antti Kariluoto is a data science and an artificial intelligence enthusiast who researches smart buildings, cybersecurity, and blockchain in the University of Jyväskylä.

Henrik Karlzén is a researcher at the Swedish Defence Research Agency and got his master's in cyber security in 2009 at Chalmers University of Technology, Gothenburg, Sweden. His research focuses on cyberwar, risk management, culture and behaviour.

Dr. Kiviharju works as a principal scientist in the Finnish Defence Research Agency in Cyber Defence, for 16 years now. Kiviharju wrote his PhD in cryptology, and specializes in the technological aspects of the cyberspace.

Thorsten Kodalle LTC (General Staff) lectures on security policy (Command and Staff College of the German Armed Forces) with a particular focus on NATO, Critical Infrastructure and Cyber. A member of the NATO research task group "Gamification of Cyber Defense/Resilience", an experienced facilitator of manual wargaming on the operational level for courses of action analysis, for operational analysis, operations research, serious gaming and especially for matrix wargaming.

Michal Konopa, MSc. is professor assistant in the Institute of Applied Informatics of Faculty of Science at the University of South Bohemia in České Budějovice. received his MSc in Informatics (Charles University in Prague, 2010). His main research area is machine learning and its application to computer network security problems. Other professional interests are algorithms, programming, and SAT solving.

Arturs Lavrenovs is a security researcher at NATO CCD COE focusing on the web and network technologies while teaching security courses, performing applied and academical research, and contributing to cyber exercises. Arturs has taught web technology and IT security courses at the University of Latvia where currently he is a PhD candidate.

Rudi Le Roux is currently completing his Bachelor of Arts degree at Stellenbosch University whilst residing in Stellenbosch. Rudi has a passion for IT security and early intrusion detection. Rudi spends his free time honing his hacking skills and familiarizing himself with the Kali operating system.

Louise Leenen's specializations are Artificial Intelligence applications in Cyber Defence and mathematical modelling. Currently an Associate Professor at the University of the Western Cape in South Africa, she was Chair of the International Federation for Information Processing's Working Group 9.10 on ICT Uses in Peace and War from 2014-2019. Holds a PhD in Computer Science from the University of Wollongong in Australia.

Dr. Martti Lehto, (Military Sciences), Col (GS) (ret.) works as a Professor (Cyber security) in the University of Jyväskylä. He has over 40 years' experience in C4ISR Systems in Finnish Defence Forces. Now he is a Cyber security and Cyber defence researcher and teacher and the pedagogical director of the Cyber Security MSc. program. He is also Adjunct professor in National Defence University in Air and Cyber Warfare. He has over 130 publications on the areas of C4ISR systems, cyber security and defence, information warfare, artificial intelligence, air power and defence policy.

Christoph Lipps is a researcher and Ph.D. candidate at the German Research Center for Artificial Intelligence's (DFKI) Intelligent Networks research group. He graduated in Electrical and Computer Engineering from the University of Kaiserslautern, where he lectures. Research focuses on Physical Layer Security (PhySec), Physically Unclonable Functions (PUFs) and the identification and authentication of various entities, including biometric authentication of humans, participating in the communication and network environment.

Sin Ming Loo is a professor at Boise State University with interests in cyber-physical systems security. He is responsible for Hartman Systems Integration and Cyber Lab for Industrial Control Systems laboratories. He holds a joint appointment with Idaho National Lab. He is a member of IEEE/CS, ISSA, Tau Beta Pi, and amateur radio (KI4AKS).

Andre Lopes is a student studying computer science. He obtained his undergraduate degree at Monash South Africa and is currently studying for his master's degree at the university of Cape Town, South Africa.

Dr. Tim Lynar is a Lecturer at the University of New South Wales. Tim is an expert in machine learning and agent-based modelling. Previously, Tim worked at IBM research for 7 and a half years, in that time he worked on a variety of projects cultivating substantial industry experience with a focus on machine learning and an emerging focus on Cyber security.

Dr. Leandros A. Maglaras is a Senior Lecturer in the School of Computer Science and Informatics of De Montfort University conducting research in the Cyber Security Centre. He is an author of more than 120 papers in scientific magazines and conferences and is a senior member of IEEE.

Isabel Martins, holds a Ph.D. in Organizational Behavior. She is currently an Associate Professor in Organizational Behavior at the University of KwaZulu-Natal, School of Management, IT & Governance, South Africa. Her scholarship spans across different countries including, South Africa, UK, Germany Portugal, Hong Kong, Middle East as well as the Gulf Countries.

Angela Mison is a Research Student at the University of South Wales. Her current interests are in cyber security of the road transportation system. She is the winner of the Automotive Electronics Innovation Cyber Student of the Year in Automotive.

Harmi Armira is a security analyst from Cyber Security Malaysia. She received her Master in Information Security (MIS) from University Putra Malaysia (UPM) in 2018. She now is an active security assessor and also conducting research on malware forensics, threat modelling, and web security.

Dr Panu Moilanen is a senior lecturer and a head of MDP in security and strategic intelligence at the University of Jyväskylä, Finland. He is interested in the technology use in everyday life and its effects on societies. Currently, he is especially interested in the role of technology in the comprehensive security of modern societies.

Michael B. Motlhabi received his BSc CS Honours (Cum Laude 2011) and his Master of Science degree in 2014 from the University of the Western Cape (UWC). He worked in the Telecommunications industry as a Packet Optical Transport Network engineer for 6 years. He is currently a Senior Cybersecurity Engineer at the Council for Scientific and Industrial Research (CSIR).

Dr. Francois Mouton is an Associate Professor in Cyber Security at Noroff University College based in Oslo, Norway. His fields of expertise are social engineering, penetration testing and digital forensics. His research has had a significant impact within the field of social engineering and he currently has an h-index and an i10-index of 9.

Julie Murphy is a PhD researcher with interests in cybersecurity and cyberpsychology. Julie works in IBM and has over 10 years' experience in the telecommunications industry. She holds an MSc in Cybersecurity, and a BA in MIS. Julie is actively involved in community efforts to promote cybersecurity, advocating education as a primary defense.

Erja Mustonen-Ollila is currently a doctoral student in the University of Jyväskylä, Finland. Her thesis covers the areas of hybrid information environment (HIE), hybrid warfare, information influence, information operations, defence strategies and actors. She has previously published in high quality journals and conferences in the areas of Information Systems Science, Software Engineering and Knowledge Management.

Dr Teija Norri-Sederholm is an adjunct professor at the Finnish National Defence University's Department of Leadership and Military Pedagogy. Her main research areas are situational awareness, inter-organisational communication in command centres and hybrid environments, national security, and the dark side of social media.

Daniel Nussbaum is a faculty member in both the Operations Research Department and the Business School at Naval Postgraduate School (NPS). He chairs the NPS Energy Academic Group and provides leadership to SECNAV Executive Energy Education program.

Christopher O'Flaherty studied a BCOM at Stellenbosch University, South Africa. Following that, completed his postgraduate degree in Information Systems Management at Stellenbosch University. Christopher currently works at BDO in Cape Town as a financial services technology junior analyst and has a passion for cyber security, technology, while also being an avid runner.

Olav Opedal has worked as an information security professional since 2001 after leaving the US Army. Fourteen years with Microsoft, three years at T-Mobile, and three years as a psychotherapist. Olav Opedal graduated as a Ph.D. in psychology in November 2019 and holds an MS in clinical psychology and a BS in computer science.

Toyosi Oyinloye is a 2002 graduate of University of Ilorin, Nigeria, with BSc in Computer Science. She obtained her MSc (with distinction) in Cyber Security from the University of Chester in 2019. She is currently undertaking PHD research in the area of Software Protection Methods and Techniques including Control Flow Integrity.

Stefan Pickl is full Professor and Chair for Operations Research at the University of the Bundeswehr in Munich. Previously, he was scientific assistant and project manager at the Center for Applied Computer Science Cologne (ZAIK).

Dr Heloise Pieterse is currently employed as a senior researcher within the Cyber Warfare research group at the Council of Scientific and Industrial Research. She completed her PhD Computer Science degree in 2019, with a focus on identifying the authenticity of smartphone data. Her interests include digital forensics, mobile device security and cyber security.

Paul POCATILU graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 1998. He achieved the PhD in Economics in 2003 with thesis on Software Testing Cost Assessment Models. He is professor at the Department of Economic Informatics and Cybernetics within the Bucharest University of Economic Studies, Bucharest.

Mr. Jouni Pöyhönen, MSc. (Industrial Development and Management), Col (ret.) is PhD-student in Cybersecurity at the Faculty of Information Technology (University of Jyväskylä) with over 30 years' experience in C4ISR Systems in Finnish Air Forces. He is a project researcher of cyber security programs in university. He has more than ten research reports and articles on the areas of cyber security.

Jyri Rajamäki is Principal Lecturer in Information Technology at Laurea University of Applied Sciences and Adjunct Professor of Critical Infrastructure Protection and Cyber Security at the University of Jyväskylä, Finland. He holds D.Sc. degrees in electrical and communications engineering from Helsinki University of Technology, and a PhD in mathematical information technology from University of Jyväskylä

Dr Trishana Ramluckan a Post- Doctoral Researcher in International Cyber Law, College of Law and Management Studies, UKZN. Her areas of research include IT and Governance. She is a member of the IFIP working group on ICT Uses in Peace and War and is an Academic Advocate for ISACA.

Dr. Huw O.L. Read is Professor of Digital Forensics and the Director of the Centre of Cybersecurity and Forensics Education and Research (CyFER) at Norwich University, Vermont, USA. For over 15 years, he has taught in several countries, authored over 20 peer-reviewed publications in the field and has been awarded numerous competitive grants.

Dr. Aunshul Rege is an Associate Professor with the Criminal Justice department at Temple University. Her cybercrime/security research on adversarial decision-making and adaptation, organizational and operational dynamics, and proactive cybersecurity is funded by several National Science Foundation grants. She has also developed hands-on cybersecurity education projects that have been mapped to the NICE Cybersecurity Framework.

MSc Reetta Riikonen is a researcher at the National Defence University of Finland. She is also a PhD student at Tampere University, Finland. She majors in social psychology. Her main research areas are the dark side of social media, intergroup relations and national identities.

Dr Marius Rogobete is associate professor for "Computer System Architecture" and for "Cryptography" at Titus Maiorescu University, Bucharest, Romania from 2012. He is System Architect for automotive at Harman Romania and received his PhD in Electronics from Romanian Technical Military Academy (2014). He is reviewer for DSP-Elsevier and other journals and involved in organising SEA-CONF international conferences.

Mr. Dimitrios Sarris earned his Master's degree in Computer Systems Security from the University of Glamorgan, UK. He has worked exclusively in the Cybersecurity sector in several countries including Greece, Kuwait and is presently Director of Risk Assessment at Digital14. He specialises in security compliance monitoring, data recovery, forensics investigations, and information assurance standards assessments.

Janine Schmoldt studied International Relations at the University of Erfurt, Germany. Afterwards, she completed her Master at the Vrije Universiteit Amsterdam where she studied Law and Politics of International Security. She is currently a PhD student at the University of Erfurt.

Matthias Schulze researcher at the security division of the German Institute for International and Security Affairs – SWP. His research covers numerous topics within the field of international cyber-security since 2010, like the strategic use of cyber-capabilities in international relations, such as cyber-conflicts, cyber-espionage, information operations, and cyber-crime.

Paulo Simões received the Doctoral degree from the Department of Informatics Engineering, University of Coimbra, Portugal, in 2002, where he is Assistant Professor. He leads industry-funded technology transfer projects for companies e.g. telecommunications operators and energy utilities. He was founding partner of two technological spin-off companies. Research interests include network and infrastructure management, security, critical infrastructure protection, and virtualization of networking and computing resources.

Ms Thenjiwe Sithole is a PhD student at the University of Johannesburg. She holds a Masters in Information Technology (Information Systems) from the University of Pretoria and a Master of Engineering Sciences with a focus on Electronics and Telecommunications from the University of Stellenbosch. She also has Certificate in Cyber Security from the University of Johannesburg.

Professor Iain Sutherland is Professor of Digital Forensics at Noroff University College in Kristiansand, Norway. He is a recognised expert in the area of computer forensics and data recovery. He has authored numerous articles in forensics practice and procedure and network security. He has consulted on Cyberesecurity issues for UK police forces and commercial organisations

Syed Ahmed Ali is doing PhD in field of Cloud Forensics at Institute of Information and Communication Technology, University of Sindh, Pakistan. He did his MS in the virtualization from Information Technology Centre, Sindh Agricultural University, Tando Jam, Pakistan. His research interests are Software Engineering, Information Security, Digital Forensics and Cloud System.

Eleanor Taylor is the Program Manager for Cybercore's Workforce Development and University Partnerships at Idaho National Laboratory (INL). She is responsible for leading initiatives designed to develop the interdisciplinary talent pipelines to attract and retain industrial control systems workforce. Before joining INL, Taylor held leadership positions at the University of Chicago, Argonne National Laboratory and SAS Institute.

Bruna Toso de Alcântara is a Ph.D. student at the Federal University of Rio Grande do Sul. She holds a Master's Degree in International Strategic Studies and a bachelor's degree in International Relations. Currently, she is a fellow at the Alexander von Humboldt Institute for Internet and Society and a Ph.D. exchange student at Humboldt University in Berlin.

Dr. Zouheir Trabelsi is a Professor of Information Security at the College of Information Technology, UAE University. He received his Ph.D. in Computer Science from Tokyo University of Technology and Agriculture, Japan. His research interests include: Network security, Intrusion systems, Firewalls, Covert channels, Information security education and curriculum development.

Gulfarida Tulemissova, associate professor of Suleyman Demirel University, Doctor PhD.Research interests are mainly in the area of Artificial Intelligence, Multiservice Networks, Software Defined Networks, Internet of Things, FPGA, Embedded systems. Today has over 45 journal publications, more than 20 conference papers (11 in Scopus and others) and three book.

Ms Eda Tumer is a Master's of Science student studying Software Engineering at De Montfort University, coming from a background of Bachelor's of Science in Computing which she studied at Cardiff Metropolitan University.

Maija Turunen is a PhD Student at the Finnish National Defense University. Her main research areas consist of cyber warfare, Russia and strategic communication.

Mr. Petri Vähäkainu is a cybersecurity PhD student (MSc., BSc.) in Faculty of Information Technology at the University of Jyväskylä in Finland, and a researcher in Finnish Defence Research Agency (FDRA). His main research areas in the University of Jyväskylä have been utilization of Artificial Intelligence in cybersecurity, health care and Structural Health Monitoring.

Carien van 't Wout is an Industrial Psychologist, working in cyber operations research at CSIR. Her career includes 25 years in the military. She is a subject matter expert in military psychological operations. Carien holds an Honours degree in Clinical Psychology and a Master of Commerce degree in Industrial and Organizational Psychology.

Brett van Niekerk is a senior lecturer in computer science at the University of KwaZulu-Natal. He serves as chair for the IFIP Working Group on ICT in Peace and War, and the co-Editor-in-Chief of the International Journal of Cyber Warfare and Terrorism. He holds a PhD focusing on information operations and critical infrastructure protection.

Federico Variola is a PhD candidate at Fudan University of Shanghai, researching cyber conflict and the intersection between conventional components of national security and contemporary technologies. Variola has working and academic experience in Europe, Russia, and China, where he currently operates as an independent consultant for a cryptocurrency derivatives exchange, Phemex.

Daniel Ventre, PhD (Political Science), is a researcher at CESDIP Laboratory (CNRS, University of Versailles, Ministry of Justice, University of Cergy Pontoise), Guyancourt, France. His research focusses on cybersecurity national strategies and cyberdefense military doctrines. He has published over 10 books about information warfare, cyberwar, cybersecurity, cyberconflict. He coordinates the Cybersecurity Series at ISTE Publishing.

Suné von Solms is an Associate Professor at the Faculty of Engineering and the Built Environment at the University of Johannesburg, South Africa. Her research interests include networks and communication, cybersecurity, engineering education and the social and human aspects of engineering. She is actively involved in engineering and community engagement projects within rural communities.

Prof. SH (Basie) von Solms is Research Professor in the Academy for Computer Science and Software Engineering at the University of Johannesburg where he is also the Director of the Centre for Cyber Security. He is Associate Director of the Global Cybersecurity Capacity Centre of the University of Oxford, serves on the Word Economic Forum's Global Council on Cybersecurity and is Past President of the International Federation for Information Processing.

Major Janne Voutilainen is a Ph.D. student at the University of Jyväskylä, Finland. He has completed Maters Degree of military sciences from pilot programme at 2005 and worked since in the Finnish Air Force at various positions. His main research areas are cybersecurity, artificial intelligence, big data analytics and military strategy.

Ambrósio Patrício Vumo is a PhD student in the ProSeM project between the Technical University of Dresden, Germany and Eduardo Mondlane University in Mozambique. He received his master's degree in Network Engineering and Communication Services from the University of Minho, Portugal. He is the author of several articles. His main research areas are: cloud computing security, network security and advanced IP network.

Rami Abu Wadi is an Assistant Professor of Accounting and Economics at Ahlia University. He has extensive experience in teaching several Accounting and Economics courses as well as supervising research dissertations for both undergraduate and graduate students. He has published many articles in international scientific Journals. He is an official reviewer in many international scientific journals.

Eduardo Wallier Vianna Ph.D. in Information Science from the University of Brasília (2019). He works with public and private sectors in the areas of digital protection and strategic knowledge. He advised the Brazilian Ministry of Defense during the cycle of major events in Brazil (2012/2016) and the Cyber Defense Command (2018/2019). Acts as Researcher at the Superior School of War and Special Advisor to the Public Ministry

Murdoch Watney is a professor at the University of Johannesburg, South Africa. She is Head of the Department: Private Law. Murdoch is an NRF rated established researcher. She contributed to four textbooks and has published on the law of criminal law, and cyber law and has delivered peer-reviewed papers at national and international conferences.

Dr Claude WEBER is a senior lecturer in Military Academy of Saint-Cyr (France), "Mutation of Conflicts" Research Department. Claude WEBER is Head of Sociology Department.

JACK WHITTER-JONES received his Master of Computing within the field of Computer Security from the University of South Wales, Wales, in 2017 and is continuing his study to achieve his Ph.D. degree. He also pursues his teaching interests at the university, focusing towards Secure Web Programming. Current research interests include network security, application security, incident response and secure programming.

Richard L. Wilson is a Professor of Philosophy at Towson University in Towson, MD. Teaching Ethics in the Philosophy and Computer and Information Sciences departments and Senior Research Fellow in the Hoffberger Center for Professional Ethics at the University of Baltimore. Professor Wilson specializes in Applied Ethics teaching a wide variety of Applied Ethics Classes.

Harry Wingo is an Assistant Professor at the National Defense University (NDU), USA. He is a member of NDU's College of Information and Cyberspace, and focuses on cyber law and emerging technologies like IoT, 5G, and artificial intelligence. He is a graduate of Yale Law School and the U.S. Naval Academy.

Ashley Wood is a Visiting Lecturer and current PhD student at the University of Chester, with a First class BSc(Hons) degree in Cybersecurity and an MSc in Advanced Computer Science with Distinction. Ashley has keen interests in digital forensics, cybercrime investigation, systems/network security and malware analysis.

Prof. Dr. Konstantinos Xynos has a strong interest in embedded devices, IoT and games consoles. Not only does he observe a device's security aspects but also the potential forensic value. He continues to pursue an active research role investigating hardware and software challenges that encompass these devices and their technological advances.

Alin ZAMFIROIU finished his PhD research in Economic Informatics at the Bucharest University of Economic Studies in 2014. Currently he works as a Lecturer at the Bucharest University of Economic Studies and as a Senior Researcher at "National Institute for Research & Development in Informatics, Bucharest".

Dr Linda Zeghache is a permanent researcher at Research Centre for Scientific and Technical Information (CERIST), Algiers, ALGERIA. She received her PhD in computer sciences from the University of Sciences and Technology Houari Boumediene in 2014. Her main research areas are application and network security, internet of things and distributed systems.

Najam ul Zia is an early career researcher. He is PhD scholar at Tomas Bata University in Zlin, Czech Republic. His broader research interests include Industry 4.0 and big data management.

Prof. Virginia Greiman is an international scholar and expert in the fields of International Law and Development, National Security, Megaproject Investment, and Cyber Law and Governance. She teaches and conducts research at Boston University and has academic appointments at Harvard University Law School, and at the Harvard Kennedy School of Government. Her research and publications are extensive and cover such topics as international cyber strategies and security, innovation megaprojects, and a global perspective on artificial intelligence. Prof. Greiman has served in several high level appointments in the U.S. Department of Justice including as U.S. Trustee on some of the nation's largest reorganizations including the Bank of New England and the Public Service Company of New Hampshire Nuclear Energy projects among other federal and state agencies. She also served as a diplomatic official to the U.S. Department of State and USAID in Eastern and Central Europe, Asia and Africa.

On the Security of the EMV Authentication Methods of Contactless Cards

Ossama Al-Maliki and Hisham Al-Assam
School of Computing, The University of Buckingham, Buckingham, UK
Ossama.al-maliki@bucingham.ac.uk
Hisham.al-assam@buckingham.ac.uk

DOI: 10.34190/EWS.20.091

Abstract: The Europay, MasterCard, and Visa (EMV) is the global standard for a chip-based card transaction. EMV payment specifications support a variety of payment methods such as chip & PIN, chip & signature, contactless, and mobile payment transactions. This paper is mainly concerned with the security analysis and evaluation of the EMV card authentication methods of contactless cards as well as investigating their vulnerabilities to sniffing and replay attacks. It provides a detailed evaluation of the three EMV card authentication methods: Static Data Authentication (SDA), Dynamic Data Authentication (DDA) and Combine Data Authentication (CDA). To demonstrate that, the paper presents a Java framework to simulate all the stages of an EMV contactless transaction under the three authentication methods. It also simulates the two attacks using Java contactless cards including a counterfeit card to show the feasibility of the replay attack. Our security analysis reveals that the authentication methods have a different level of robustness against the attacks. The paper shows that SDA has the lowest level of security with vulnerabilities to both attacks. The paper also demonstrates that although DDA and CDA are perceived to be secure, they are both vulnerable to sniffing attacks that can be easily launched to steal credit/debit card details. We argue that the card authentication methods share a fundamental flaw related to the trust model between the point of sale and the contactless card. While the Point of Sale (POS) enforces strict rules to ensure the authenticity of the card, the contactless cards release sensitive information to anyone with a card reader without any checks whatsoever.

Keywords. EMV, card, authentication methods, replay attack, sniffing attack, transaction

1. Introduction

EMV protocol stands for Europay MasterCard and Visa is the most popular chip-based payment protocol. The EMV chip-based cards provide cryptographic functions for the more secure transaction than the traditional magnetic stripe cards. The EMV payment protocol owned by the EMVCO provides all the required specifications for each supported payment methods. The EMVCO reported that 85.5 % of the cards in Europe including the UK were EMV cards in the last quarter of 2018.

Moreover, the same source reported that 97.3% of the transactions were EMV transactions (EMVCO, 2018). EMV specifications support different kinds of payment methods such as Chip & PIN, Chip & Signature, Contactless card, Mobile, and magnetic stripe transactions where each payment method has its own EMV specifications(EMVCo, 2016). EMV payment protocol consists of three phases to complete any EMV transaction. The phases are card authentication, cardholder verification and transaction authorization (Breekel *et al.*, 2016). This paper is mainly focused on the security of EMV contactless cards and, therefore the first phase of the EMV payment protocol i.e. card authentication. As the name suggests, the aim behind the card authentication phase is to authenticate an EMV card to a POS or any Automated Teller Machine (ATM). The paper investigates the feasibility and effectiveness of launching sniffing and replay attacks against the three EMV card authentication methods explained in section 2.

The UK finance reported in 2018 that %56 of the total frauds in the UK bank system were payment card frauds. The remote purchase and Card Not Present (CNP) frauds represent %76 of the total card frauds in the UK with a total amount of £506.4 million. This number has been increased by %24 with a comparison with the previous year (UK Finance, 2019). These enormous fraud numbers are possible due to many vulnerabilities in the EMV payment protocol. Therefore, this paper details one of these vulnerabilities that make these frauds are possible.

The paper argues that sniffing attack is easy to perform on the EMV cards especially in case of the EMV contactless cards due to the wireless connectivity of such cards. It has been shown that attackers can sniff most of the EMV contactless card's sensitive information such as the Primary Account Number (PAN), cardholder name and the expiry date of the card without the knowledge of the genuine cardholder (Madhoun, Bertin and Pujolle, 2018). This can be done using off-the-shelf software and hardware such as an NFC reader or even an Android phone(Emms and van Moorsel, 2011). Once such information is stolen, it can be used to launch a "card not present" attack such as online shopping as some e-commerce websites do not require the Card Verification Code (CVC), i.e., the 3-digits written on the back of the card (Mehrnezhad *et al.*, 2016). The second attack that this paper focuses on is the replay attack in which an attacker records the static information of the contactless card and replay the information to get some spoofed transactions authorized.

This paper starts with a brief introduction to the EMV payment protocol and the three phases of the protocol as detailed in section 1. Then section 2 details briefly the EMV authentication methods and the three types. While section 3 shows our analysis and break down of the three types of EMV card authentication methods. Next, section 4 details our implementations and evaluations of two different attacks which are sniffing and replay attacks. Finally, section 5 details our conclusion and future work.

2. EMV Card Authentication Methods & Literature Review

The EMV card authentication methods are used to authenticate an EMV card to a POS and ATM (EMVCo, 2011a) as shown in Figure 1. The EMV specifications support the following three methods:

1. Static Data Authentication (SDA)
2. Dynamic Data Authentication (DDA)
3. Combine Data Authentication (CDA)

Each EMV card must support at least one of the above authentication methods. This is typically decided by the Issuer bank (IB) and is declared by the Application Interchange Profile (AIP) sent by the EMV card to the POS during the transaction(EMVCo, 2016). After requesting the AIP using the Get Processing Option commands (GPO), the POS chooses the authentication method of the highest priority where CDA has the highest followed by DDA and then SDA. The rationale behind these different priorities is explained in section 4.

The paper argues that the EMV specifications do not provide any authentication methods to authenticate the POS to the EMV card which means that under the current specifications the card is revealing sensitive information to any POS or card readers without any form of authentication. Also, the EMV card authentication methods offer one-way authentication methods that offer to the POS the ability to authenticate the EMV cards while the reverse is not happening.

Many proposed solutions have been proposed a mutual authentication protocol for the EMV mobile payment in order to provide a two-way authentication method between both the NFC mobile payment and the POS and ensure the confidentiality of the sensitive information of the EMV mobile payment.

Many researchers proposed the use of a cloud-based secure authentication protocol for the EMV mobile payment due to the ability of the mobile to be connected to the cloud at the transaction time. For example,(Madhoun, Guenane and Pujolle, 2015) proposed the use of asymmetric cryptography to ensure mutual authentication and encryption of the payment information during the transaction between NFC mobile payment and POS. At the time of the transaction, the NFC mobile requests the cloud infrastructure to verify the POS as a genuine one before starting to send any sensitive payment information back to the POS. Also, the cloud generates a session key if the POS was authenticated as a genuine one by the cloud authenticator. This session key is used by both NFC mobile and POS to encrypt the transaction details to ensure the integrity and confidentiality of the transaction data. The second example was called KerNees protocol where it is based on Needham-Schroeder based protocol to ensure mutual authentication and confidentiality between the NFC mobile payment and POS(Ceipidor *et al.*, 2012). KerNees proposed an authentication server to be used to verify both NFC mobile and POS to each other and provide them with a session key to be used in order to encrypt the TD during the EMV mobile transaction.

However, the work that was proposed by (Martin Emms, Budi Arief, 2013) to design a mutual authentication protocol for the EMV contactless cards was targeting the EMV contactless cards by proposing the use of a revised version of the EMV card authentication methods. When the POS should have its own pair of RSA public/private key along with the POS public key certificate (EMVCo, 2011). The POS needs to sign the EMV contactless card nonce by its own RSA private key and send it back along with its own public key certificate to the EMV card. Then, the card retrieves the POS public key to verify its own nonce and authenticate the POS.

However, there is a massive limitation in this proposal related to the size of the APDU commands. Where the maximum size of single APDU command is 256 bytes and in the proposal the POS needs to send two certificates to the EMV contactless card which is going to overdo the maximum size of the APDU. Therefore, the authors of this proposal suggested the use of Elliptic Curve Cryptography (ECC) instead of the RSA keys in order to avoid this problem.

Both previous proposals for the EMV mobile payment depend on the capability of the NFC mobile phone to connect with either the cloud and the authentication server to authenticate the POS. Nevertheless, this NFC mobile's capability is not available in the EMV contactless cards as this kind of card does not have any connectivity without side world apart of the POS. As a result, a reasonable mutual authentication protocol for the EMV contactless cards should depend on the EMV cards themselves without the need for any sort of cloud or even authentication server. Also, one of the considerations that we need to put it in our mind when designing a mutual authentication for the EMV contactless card is the size of the APDU commands and responses. Moreover, a crucial point that needs to be considered is the EMV contactless card transaction time. The EMV specifications allow up to 500 milliseconds to process the contactless card transaction (Jain and Dahiya, 2015). However, some of the EMV contactless card transactions could take a slightly longer time than 500 ms due to various reasons (Lacmanović, Radulović and Lacmanović, 2010). One of these reasons is the way to place the contactless card on the POS as it makes a significant difference when placing the card directly on the POS or placing it close to the POS. This affects the transaction time between both the POS and the EMV contactless card. The second reason is the difference between hardware and software of both POS and the contactless chip that might be different from Bank to others. Therefore, to propose a mutual authentication for the EMV contactless card, the transaction time must be considered.

In the next Section, we provide the technical details of the three EMV authentication methods that are needed to build the Java framework for simulation purposes. After that, we evaluate and analyze the security of each of the methods.

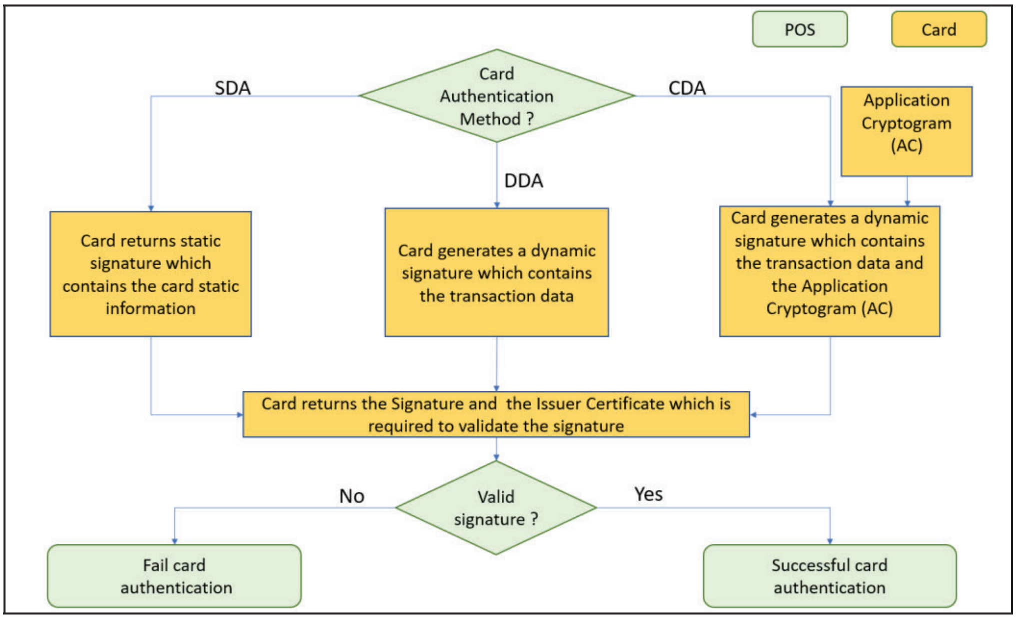

POS
Card
Card Authentication Method ?
SDA
DDA
CDA
Application Cryptogram (AC)
Card returns static signature which contains the card static information
Card generates a dynamic signature which contains the transaction data
Card generates a dynamic signature which contains the transaction data and the Application Cryptogram (AC)
Card returns the Signature and the Issuer Certificate which is required to validate the signature
Valid signature ?
No
Yes
Fail card authentication
Successful card authentication

Figure 1: EMV Card Authentication Methods

3. Analysis of the EMV Card Authentication Methods

Each of the three EMV authentication methods consists of five stages that are the keys distribution, IB public key certificate, card personalization, transaction, and authentication. The first three stages are done before issuing the EMV card to its cardholder as shown in Figure 2 while the last two stages are done during the transaction as shown in both Figures 3 and 4.

3.1 EMV Card Authentication at the Keys Distribution Stage

At this stage, all the required keys are shared between all the involved parties as shown in the next three steps. This stage applies to all three types of EMV card authentication methods namely SDA, DDA, and CDA as shown in Figure 2.

1. The Certificate Authority (CA) sends its public key (CApk) in a secure environment to the IB.

   (1) CA→ IB: CApk

2. The CA sends its CApk in a secure environment to the Acquirer Bank (AB).

   (2) CA→ AB: CApk

3. The AB upload the CApk into the POS which belongs to AB.

   (3) AB→ POS: CApk

3.2 EMV Card Authentication at the IB Public Key Certificate Stage

At the second stage of the EMV card authentication, the IB certificate is generated as shown in the next two steps. This stage applies to all three types of EMV card authentication methods as shown in Figure 2.

1. The IB sends its IBpk to the CA.

   (4) IB→ CA: IBpk

2. The CA uses its Private key (CAsk) to sign the IBpk and then send it back to the IB.

   (5) CA→ IB: {IBpk}CAsk

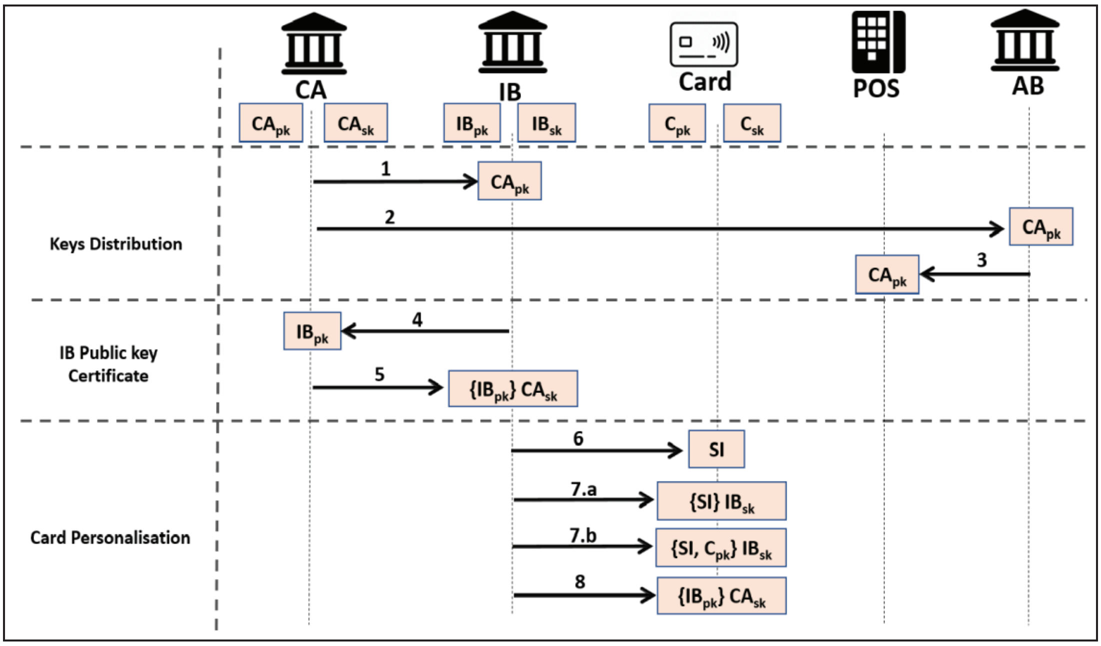


Figure 2: Keys Distribution, IB public key certificate and Personalization processing in the EMV card authentication methods

3.3 EMV Card Authentication at the Card Personalisation Stage

At the third stage of the EMV authentication methods, the EMV card is uploaded with all its Static Information (SI) and the IB public key certificate. This stage applies to all three types of EMV card authentication methods as shown in Figure 2.

1. The IB stores the SI for each EMV card. These SI include the cardholder name, PAN and the expiry date of the card.

   (6) IB→ Card: SI

2. The IB signs the SI by its IBsk and stores into the card in case of the SDA as shown in Figure 2 step (7.a). While in the case of the DDA and CDA, the IB signs both SI and the Cpk by the IBsk as shown in Figures 2 step (7.b). Equation 7.a represents the process in case of the SDA while equation 7.b represents the process in case of both the DDA and CDA.

   (7.a) IB→ Card: {SI'}IBsk

   (7.b) IB→ Card: {SI', Cpk}IBsk

3. The IB stores the IBpk certificate which was generated at step 5 into the card.

(8) IB→ Card: {IBpk}CAsk

3.4 EMV Card Authentication at the Transaction Stage

At the fourth stage of the EMV authentication methods, both POS and EMV cards start to communicate with each other to process a transaction. Steps 9, 10 and 11 apply for all the three types of EMV card authentication methods. While steps 12 and 13 apply just for both DDA and CDA and step 14 applies just in case of the CDA as shown in Figures 3.

1. The card sends its own SI in plaintext (the IB stored that at step 6) to the POS in the response of read record command.

   (9) Card→ POS: SI

2. The card sends the signed SI (that was generated and stored at step 7.a in case of the SDA) to the POS in response to the read record command as shown at the equation 10.a. while in case of both DDA and CDA, the card sends both of the signed SI and the Cpk (which were generated and stored at step 7.b in case of both DDA and CDA) to the POS in response to the read record command as shown at the equation 10.b.

   (10.a) Card→ POS: {SI'}IBsk

   (10.b) Card→ POS: {SI', Cpk}IBsk

3. The card sends the IBpk certificate (which was generated at step 5 and stored at step 8) to the POS in response to the read record command.

   (11) Card→ POS: {IBpk}CAsk

4. POS sends the Transaction Data (TD) to the card during the Generate Application Cryptogram command. This TD includes the transaction amount, the transaction date, transaction time, transaction currency and Unpredictable Number (nonce).

   (12) POS→ Card: TD

5. In case of the DDA, the card signs the TD which was sent by the POS at step (12) by its Private key (Csk) and sends it back to the POS during the Generate AC response as shown in the equation 13.a. while in case of the CDA, the card signs the TD together with the Application Cryptogram (AC) by its Csk and send it back to the POS during the Generate AC response as shown in the equation 13.b.

   (13.a) Card→ POS: {TD'}Csk

   (13.a) Card→ POS: {TD', AC'}Csk

6. The card sends its own AC in plaintext to the POS. this step is valid just in case of the CDA.

   (14) Card→ POS: AC

3.5 EMV Card Authentication at the Authentication Stage

At the fifth and last stage of the EMV authentication methods, the POS retrieves the IB_{pk} in case of SDA, DDA and CDA and the C_{pk} in case of both DDA and CDA. These retrieved keys help the POS to obtain all the required information and decide whether the EMV card is a genuine card or not. Steps 15 and 16 apply for all the three types of EMV authentication methods. While steps 17, 18 and 19 apply for both DDA and CDA as shown in Figure 4.

1. POS uses the CApk (which was uploaded by the AB at step 3) to obtain the IBpk from the message number (11). This step applies to the three types of EMV card authentication methods.
2. In the case of the SDA, the POS uses the IBpk (which was obtained at step 15) to verify the message number (10.a) and obtain the SI' which belongs to the EMV card. Then, the POS compares both SI data and all the static card information which send to the POS in plaintext during the read record command at step 9. If both are matched. Then, the POS successfully authenticates the EMV card as a genuine one and it should set the SDA Failed bit to 0 in the first byte of the Terminal Verification Results (TVR) [9]. Else, the authentication process is failed, and the POS should set the SDA Failed bit to 1 in the first byte of the TVR. While in the case of both DDA and CDA, the POS uses the IBpk (which was obtained at step 15) to decrypt message number (10.b) to obtain the SI' which belongs to the EMV card and the Cpk.

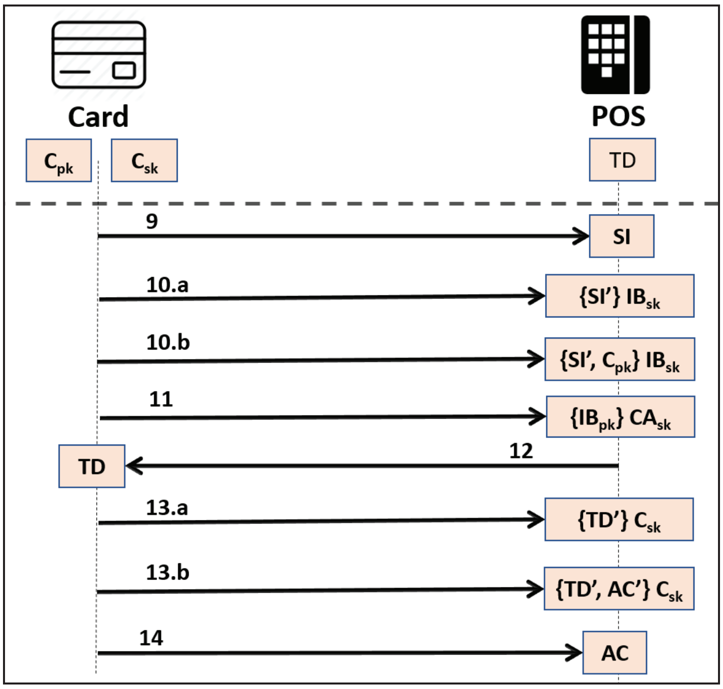


Figure 3: Transaction stage of the EMV card authentication methods

1. In the case of the DDA, the POS uses the Cpk which was obtained at step (16) to verify the message number (13.a) and obtain the TD. While in the case of the CDA, the POS uses the Cpk which was obtained at step (16) to verify message number (13.b) and obtain both TD' and the AC'.
2. In the case of the DDA, the POS compares two components which are the SI and TD to make its decision whether the card is a genuine EMV card or not. Firstly, the POS matches between both the signed version of the SI' (which was obtained at step 16) and all the static card information send to the POS in plaintext during the read record command at step 9. Secondly, the POS compares between its own TD (which was sent to the card at step (12) and the TD' which was signed by the card' Csk at step (13) and obtained at step (17). If both two comparisons are true. Then, the POS successfully authenticates the EMV card as a genuine one and it should set the DDA Failed bit to 0 in the first byte of the TVR. Else, the authentication process is failed, and the POS should set the DDA Failed bit to 1 in the first byte of the TVR.
3. In the case of the CDA, the POS compares three components which are the SI, TD and the AC to make its decision whether the card is a genuine EMV card or not. Firstly, the POS matches between both the signed SI' (which was obtained at step 16) and all the static card information send to the POS in plaintext during the read record command at step 9. Secondly, the POS compares between its own TD (which was sent to the card at step 12) and the TD' which was signed by the card' Csk at step (13) and obtained at step (17). Thirdly, the POS matches between the signed AC' (which was obtained at step 17) and the AC which was sent by the card as a response of the generate AC command at step (14). If all the three comparisons are true. Then, the POS successfully authenticates the EMV card as a genuine one and it should set the CDA Failed bit to 0 in the first byte of the TVR. Else, the authentication process is failed, and the POS should set the CDA Failed bit to 1 in the first byte of the TVR.

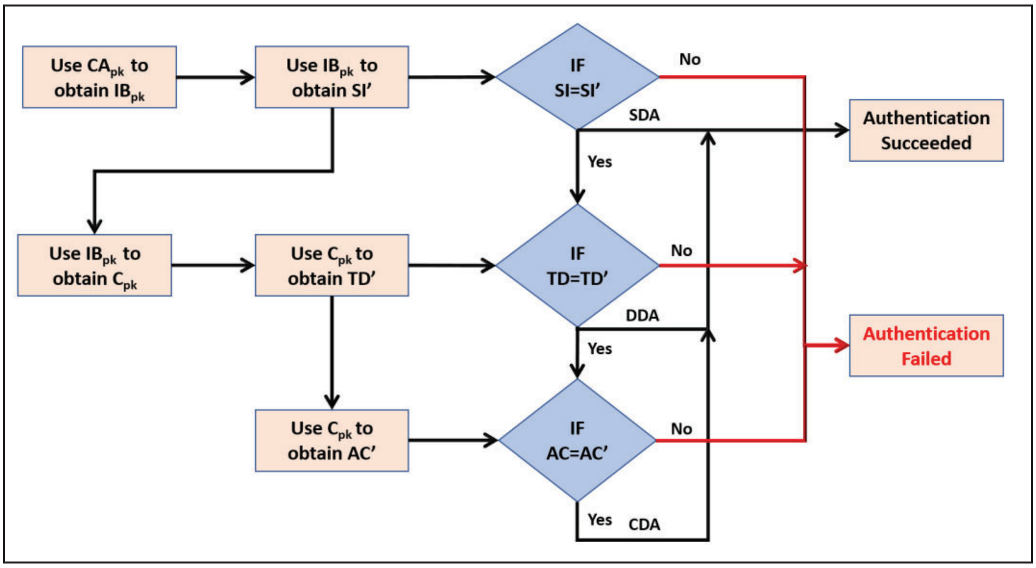

Use CA$_{pk}$ to obtain IB$_{pk}$
Use IB$_{pk}$ to obtain SI'
IF SI=SI'
No
SDA
Yes
Authentication Succeeded
Use IB$_{pk}$ to obtain C$_{pk}$
Use C$_{pk}$ to obtain TD'
IF TD=TD'
No
DDA
Yes
Authentication Failed
Use C$_{pk}$ to obtain AC'
IF AC=AC'
No
Yes
CDA

Figure 4: Authentication stage of the EMV card authentication methods

4. Implementation & Evaluation

This section details our implementations for the EMV contactless cards payment protocol and the three authentication methods. Our main aim behind the implementation is to implement the three EMV card authentication methods and how these methods react to both sniffing and replay attacks. We have done our implementation by using both hardware and software as details in the next subsections. Also, this section discusses how the three EMV authentication methods react against two attacks which are the sniffing and replay attacks.

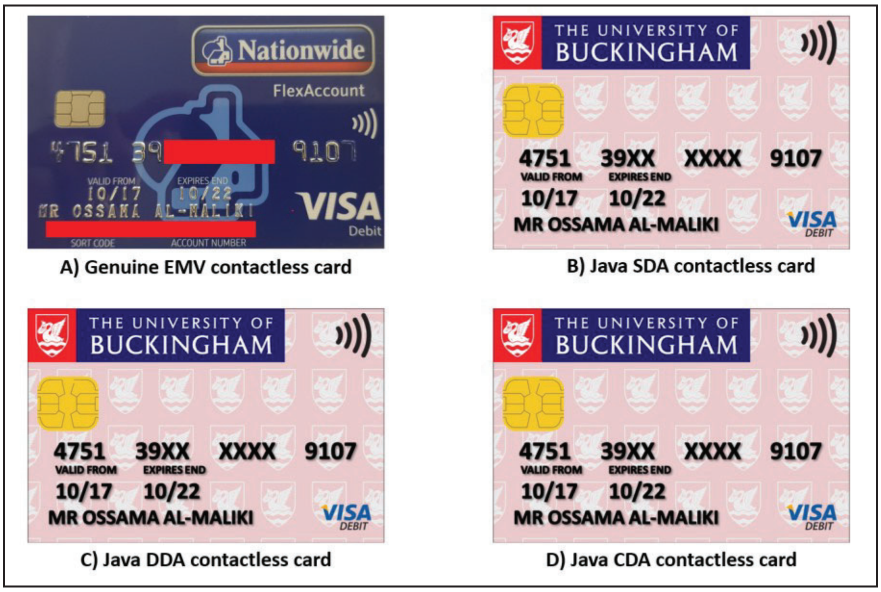
4.1 Hardware & Software

We have used the NFC ACR 122U reader to represent the POS in our implementation(Ltd., 2017). This NFC reader sends and receives all the required commands and responses according to the EMV specifications. Also, we have used Java contactless cards to represent the EMV contactless cards and the counterfeit cards in case of the replay attack implementation. While we have used Java Card Integrated Development Environment (JCIDE) as our software to develop our applets to represent EMV contactless cards(Attali *et al.*, 2001). As we have developed three applets each one represents one of the three types of EMV authentication methods namely SDA, DDA, and CDA. Besides, PyApduTool software was used to send and receive the Application Protocol Data Unit (APDU) commands and responses. Moreover, this software was used to download and install our applets to the Java contactless cards.

4.2 EMV Contactless Card Applets

We have developed three EMV applets using the JCIDE software to represent the three types of EMV authentication methods. The source code of the applets has been publicly available to other researchers to re-use [1]. The applets act as an EMV contactless card by responding to the POS commands according to the EMV specifications. However, all these three applets have been developed with the same keys (CA$_{sk}$, CA$_{pk}$, IB$_{sk}$, IB$_{pk}$, C$_{sk,}$ and C$_{pk}$) and the same card static information (cardholder name, PAN and expiry date) to serve the purpose of this paper and show how the SDA, DDA, and CDA react to each of sniffing and replay attacks. Figure 5 shows our four testing contactless cards which represent genuine EMV contactless card (A), Java SDA Contactless card (B), Java DDA contactless card (C), and Java CDA contactless card (D). Where we built our Java applets to duplicate the genuine EMV contactless card (A) while all of Java cards (B, C and D) represent the SDA, DDA and CDA respectively.

[1] https://www.buckingham.ac.uk/directory/dr-hisham-al-assam/



A) Genuine EMV contactless card

B) Java SDA contactless card

C) Java DDA contactless card

D) Java CDA contactless card

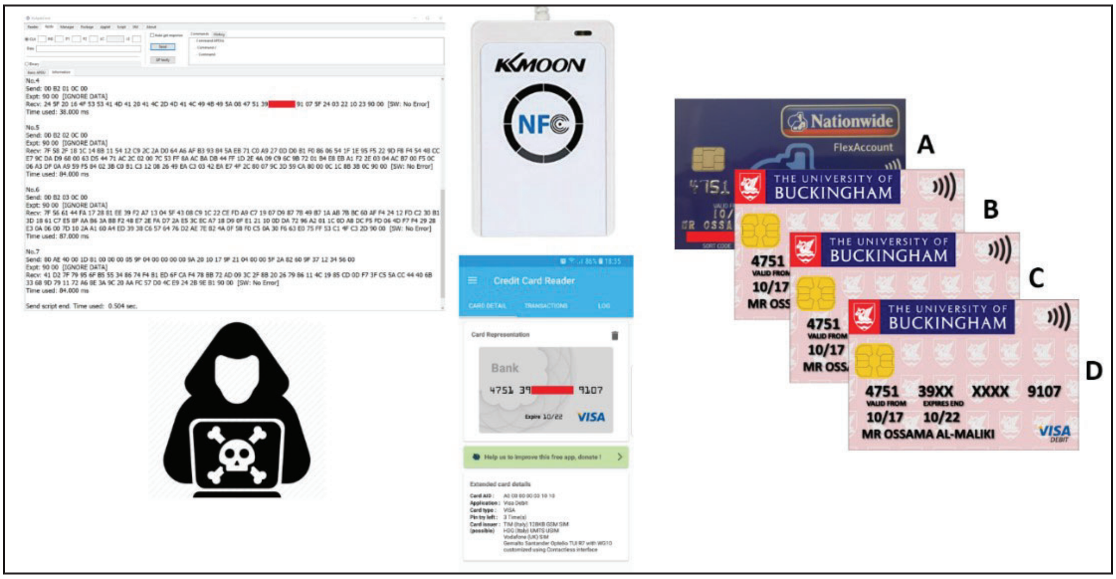
Figure 5: The Used Contactless Cards for testing the EMV contactless card authentication methods

4.3 Sniffing Attack

The EMV contactless cards can be read by any POS or even by any NFC enabled smartphones due to the wireless connectivity of the cards, which means that fraudsters can easily obtain most of the card's sensitive information such as the PAN, expiry date, and even the cardholder name using on-the-shelf hardware and software. We have successfully obtained all these data using the NFC ACR 122U reader as POS and the PyApduTool to send a set of commands to a genuine EMV contactless card and our Java contactless cards. Moreover, we have used a free off-the-shelf Android application called "Credit Card Reader" to read most of the contactless card information even though the card still inside the cardholder wallet, which can be easily done without the cardholder's knowledge as shown in Figure 6.

All three EMV authentication methods (SDA, DDA, and CDA) fall to the sniffing attack. That is because the EMV card sends its own SI in plaintext to the POS to enable the POS to verify the signed version of the SI either $\{SI\}IB_{sk}$ (step 10.a Figure 3) or $\{SI', C_{pk}\}IB_{sk}$ (step 10.b in Figure 3) with the plaintext version of the SI (step 9 in Figure 3). The stolen data could be eventually used by the attacker to achieve Card Not Present (CNP) attacks such as Online shopping as some of the online websites do not require the Card Verification Code (CVC) as shown in Figure 7. Moreover, the sniffed data could be uploaded into a counterfeit card to launch a replay attack as explained in the next subsection.

Many countermeasures have been proposed by researchers to prevent sniffing attacks on the EMV contactless card. One of the first ideas was to use a unique RFID-Block wallet to prevent these cards from being read by untheorized NFC readers/smartphones (Ghosh *et al.*, 2015). Also, one of the earlier proposals was making the EMV contactless cards as an active card with a built-in battery instead of the original passive cards. The first method was by using cards with the activation button where the cardholder needs to push the activation button to activate the card in order to allow the EMV card to perform a contactless transaction and the cardholder needs to deactivate the card after the transaction is done(Hutter, Schmidt and Plos, 2008). The second mothed was to propose the use of a built-in light sensor on the EMV contactless cards. Once the card is exposed to the light (outside the cardholder's wallet), the light's sensor activates the card while if the card inside the cardholder's wallet, the card is always deactivated. When a genuine cardholder needs to do a transaction, the sensor activates the card and allows it to communicate with the POS(Yum *et al.*, 2010). In both methods, skimming and relay attacks are prevented as the card is deactivated in case of not pushing the activation button by the cardholder or in case of the card is inside the cardholder's wallet. However, the main limitation of these two methods is the battery's life span where the card's battery is dead then the battery needs to be replaced or recharged. Another approach was proposed by MasterCard in October 2014 to use a built-in fingerprint sensor to prevent both skimming and relay attacks by verifying the cardholder each time a contactless card transaction is processed(Vats, Ruhl and Aghili, 2016). The proposed EMV contactless cards offer secure biometric authentication technology that stores the cardholder's fingerprint bitmap at the personalization phase. Then when the cardholder needs to do a contactless card transaction, he should place his fingerprint on the built-in sensor and place the card next to the POS. Next, the POS powers up the EMV card and the fresh fingerprint data is compared with the stored one. If both fingerprints are matched then the EMV contactless card is activated, else the card is not activated.



```
No.4
Send: 00 B2 01 0C 00
Expt: 90 00  [IGNORE DATA]
Recv: 24 5F 20 16 4F 53 53 41 4D 41 20 41 4C 2D 4D 41 4C 49 4B 49 5A 08 47 51 39[redacted]91 07 5F 24 03 22 10 23 90 00  [SW: No Error]
Time used: 38.000 ms

No.5
Send: 00 B2 02 0C 00
Expt: 90 00  [IGNORE DATA]
Recv: 7F 58 2F 18 1C 14 BB 11 54 12 C9 2C 2A D0 64 A6 AF B3 93 84 5A EB 71 C0 A9 27 0D D0 B1 F0 86 06 54 1F 1E 95 F5 22 9D F8 F4 54 48 CC
E7 9C DA D9 68 00 63 D5 44 71 AC 2C 02 00 7C 53 FF 8A AC BA DB 44 FF 1D 2E 4A 09 C9 6C 9B 72 01 B4 E8 EB A1 F2 2E 03 04 AC B7 00 F5 0C
06 A3 DF 0A A9 59 F5 84 02 3B C0 B1 C3 12 08 26 49 EA C3 03 42 EA E7 4F 2C 80 07 9C 3D 59 CA 80 00 0C 1C 8B 38 0C 90 00  [SW: No Error]
Time used: 84.000 ms

No.6
Send: 00 B2 03 0C 00
Expt: 90 00  [IGNORE DATA]
Recv: 7F 56 61 44 FA 17 28 81 EE 39 F2 A7 13 04 5F 43 08 C9 1C 22 CE FD A9 C7 19 07 D9 87 78 49 B7 1A AB 7B 8C 60 AF F4 24 12 FD C2 30 B1
3D 18 61 C7 E5 8F AA B6 3A BB F2 48 E7 2E FA D7 2A E5 3C EC A7 18 D9 0F E1 21 10 0D DA 72 96 A2 01 1C 0D A8 DC F5 FD 06 4D F7 F4 29 28
E3 0A 06 00 7D 10 2A A1 60 A4 ED 39 38 C6 57 64 76 D2 AE 7E 82 4A 0F 58 F0 C5 0A 30 F6 63 E0 75 FF 53 C1 4F C3 2D 90 00  [SW: No Error]
Time used: 87.000 ms

No.7
Send: 80 AE 40 00 1D 81 00 00 00 05 9F 04 00 00 00 00 9A 20 10 17 9F 21 04 00 00 5F 2A 82 60 9F 37 12 34 56 00
Expt: 90 00  [IGNORE DATA]
Recv: 41 D2 7F 79 95 6F 85 55 34 86 74 F4 B1 ED 6F CA F4 78 8B 72 AD 09 3C 2F 8B 20 26 79 86 11 4C 19 85 CD 0D F7 3F C5 5A CC 44 40 6B
33 68 9D 79 11 72 A6 8E 3A 9C 20 AA FC 57 D0 4C E9 24 2B 9E B1 90 00  [SW: No Error]
Time used: 84.000 ms

Send script end. Time used:  0.504 sec.
```

Credit Card Reader

CARD DETAIL TRANSACTIONS LOG

Card Representation

Bank

4751 39[redacted]9107

Expire 10/22 VISA

Help us to improve this free app, donate !

Extended card details

Card AID : A0 00 00 00 03 10 10
Application : Visa Debit
Card type : VISA
Pin try left : 3 Time(s)
Card issuer (possible) : TIM (Italy) 128KB GSM SIM
HDG (Italy) UMTS USIM
Vodafone (UK) SIM
Gemalto Santander Optelio TUI R7 with WG10 customized using Contactless interface

Figure 6: Sniffing Attacks on our testing cards

4.4 Replay Attack

After demonstrating the ease of launching a sniffing attack, this section highlights the vulnerability of EMV contactless cards to the replay attack. We show that an attacker can use the sniffed data to create their own counterfeit card to launch the replay attack. To do so, we loaded a Java contactless cards with the following sniffed information: SI, signed SI', and the IB public certificate (in case of the SDA) so that the Java card can act as a counterfeit card. As illustrated in Figure 8, the attack can be successful if the attacker tries to pay for goods or services using the counterfeit card when the transaction is processed offline where the POS and the card approve or decline the transaction without the IB. In fact, some retail prefers to use an offline contactless transaction instead of the online for several reasons such as the cost of the connectivity and even the availability of the connectivity. For example, most of the POS at the London Underground Stations are using an offline contactless transaction(The Uk Cards Association, 2017). In this case, the attacker's card can fool the POS and act as a genuine EMV contactless card in case of the SDA. This is mainly because POS gets a positive outcome when upon verifying the signed version of SI using the plaintext version of the SI and the IB public certificate, which leads to approving the offline transaction.

It is essential to highlight that after the attacker uses the services or collects the goods, the POS sends later the transaction information to the IB through the AB to process the transaction. At that point, the IB declines the transaction because the AC is not signed by the C_{sk} as shown in Figure 8. However, this might be too late.

On the other hand, the details of DDA and CDA in section 3 shows that the above replay attack has no chance of success even in offline transactions due to the negative output that the POS gets when attempting to verify the signed version of the TD in case of the DDA and the signed version of TD and AC in case of the CDA as the attacker does not have the C_{sk} to sign the TD and AC. Therefore, the POS rejects the transaction initiated by the attacker using the counterfeit card as shown in Figure 8.

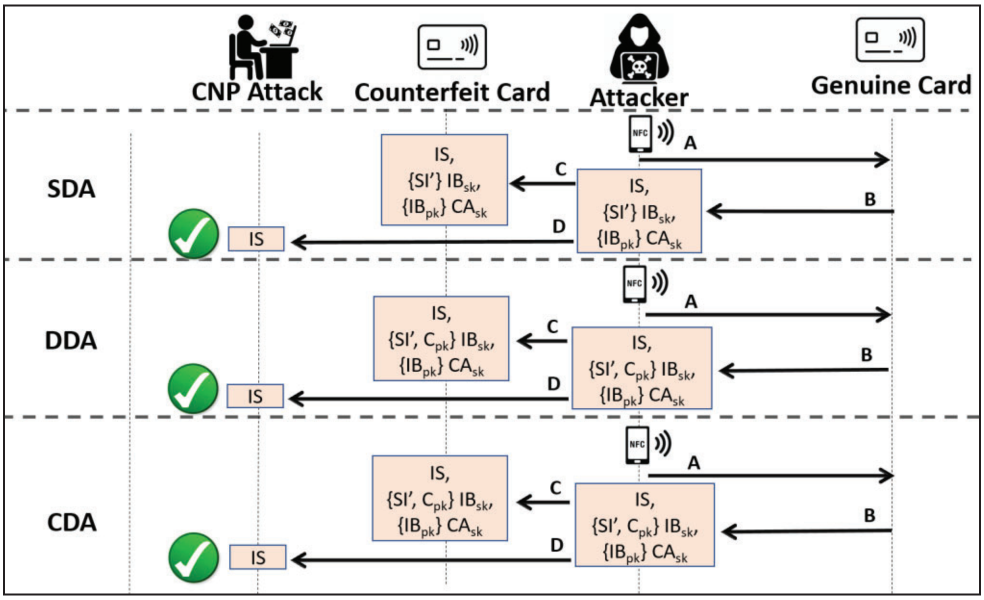

CNP Attack
Counterfeit Card
Attacker
Genuine Card
SDA
DDA
CDA
A
B
C
D
IS
IS, {SI'} IB_sk, {IB_pk} CA_sk
IS, {SI', C_pk} IB_sk, {IB_pk} CA_sk
NFC

Figure 7: Sniffing Attacks on the EMV card authenticating methods

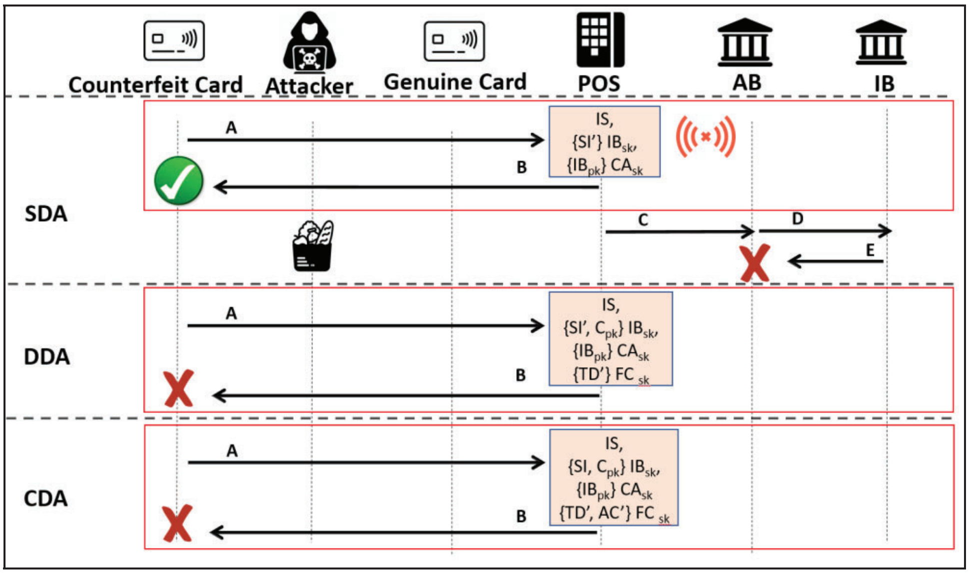

Counterfeit Card
Attacker
Genuine Card
POS
AB
IB
SDA
DDA
CDA
A
B
C
D
E
IS, {SI'} IB_sk, {IB_pk} CA_sk
IS, {SI', C_pk} IB_sk, {IB_pk} CA_sk {TD'} FC sk
IS, {SI, C_pk} IB_sk, {IB_pk} CA_sk {TD', AC'} FC sk

Figure 8: Replay Attacks on the EMV card authentication methods

5. Conclusions, Discussion & Future Work

This paper is an attempt to unwrap the EMV authentication methods and evaluate the security of the three EMV authentication methods. To underpin the exact source of vulnerabilities, the paper explains the five stages of the authentication methods, namely the keys distribution, Generation of IB public key certificate, card personalization, transaction and authentication stages. Moreover, we presented a Java framework and made it publicly accessible to other researchers to simulate all the stages of an EMV contactless transaction under the three authentication methods as well as simulating the sniffing and replay attacks using Java contactless cards.

Section 4 showed that the SDA, which usually has the lowest priority amount of the three EMV authentication methods, is vulnerable to the two attacks. However, the SDA still provides strong evidence to the POS that the card's static information is genuine and guarantees its integrity after the Personalization phase.

The DDA, on the other hand, has a higher level of security than the SDA as it provides a countermeasure against the replay attack via the unique signature for each transaction as shown in Figure 3 (step 13.a). However, the DDA is still vulnerable to the sniffing attack for the same reasons the SDA does. In addition to providing strong evidence about the authenticity and integrity of the EMV card's static data to the POS, it provides the EMV card approval of the TD by signing the TD using C_{sk} as shown in Figure 3 (step 13.a).

The CDA provides another layer of security and thus it often given the highest priority among the three EMV authentication methods. In addition to guaranteeing the integrity and authenticity of the static data, and signing the TD, the CDA signs its own transaction decision by signing the AC together with the TD using C_{sk} as showing in Figure 3 (step 13.b) to give another level of assurance to the POS that the card has indeed approved the transaction and prevent the possibility of non-repudiation attacks. Although the CDA is perceived to be highly secure, the paper demonstrated its vulnerability to sniffing attacks that can be easily launched to steal credit/debit card details. Table 1 summarises the vulnerabilities of the three EMV authentication methods against the sniffing and replay attacks.

Table 1: EMV Authentication Methods VS. Attacks

Attacks	**EMV Card Authentication Methods**		
	SDA	**DDA**	**CDA**
Sniffing Attack	Not Protect	Not Protect	Not Protect
Replay Attack	Not Protect	Protect	Protect

It can be argued that the card authentication methods all share a fundamental flaw related to the trust model between the Point of Sale and the contactless card. We showed that the POS enforces strict rules to ensure the authenticity of the card, the integrity of its information, and even getting the card to sign the TD and the final decision. In contrast, the paper showed that the contactless cards release sensitive information to anyone with a card reader without any checks whatsoever. Our simulations showed that all what it takes is to send the commands in the correct format as per the EMV publicly-available specifications. Therefore, future research direction could be focussing on developing practical solutions to provide mutual authentication between the contactless card and the POS to ensure that the card does not leak any sensitive information unless it gets the right level of assurance that the POS is a genuine one.

Acknowledgment

The first author would like to thank the Iraqi Ministry of Higher Education and Scientific Research, and the University of Basrah for sponsoring this PhD study. Also, the first author would like to thank Prof. Ihsan Lami for his valuable guidance throughout the first stage of his PhD.

References

Attali, I. *et al.* (2001) 'An integrated development environment for Java Card', *Computer Networks*. doi: 10.1016/S1389-1286(01)00162-1.

Breekel, J. Van Den *et al.* (2016) 'EMV in a nutshell', *Technical Report*, pp. 1–37.

Ceipidor, U. B. *et al.* (2012) 'A protocol for mutual authentication between NFC phones and POS terminals for secure', pp. 115–120.

Emms, M. and van Moorsel, A. (2011) 'Practical Attack on Contactless Payment Cards', in *HCI2011 Workshop - Heath, Wealth and Identity Theft*.

EMVCo (2011a) *Book 2: Security and Key Management*, *EMV Integrated Circuit Card Specifications for Payment Systems*. doi: 10.7591/9780801469121-014.

EMVCo (2011b) *Book 4: Cardholder, Attendant, and Acquirer Interface Requirements*.

EMVCo (2016a) 'Book A Architecture and General Requirements. Contactless Specifications for Payment Systems', 2.6(March).

EMVCo (2016b) 'Book B: Entry Point Specification. EMV Contactless Communication Protocol Specification', 2.6(July).

EMVCO (2018) '2018: a Year in Review', 392(10165), pp. 2669–2670. doi: 10.1016/s0140-6736(18)33244-6.

Ghosh, S. *et al.* (2015) 'Issues in NFC as a form of contactless communication: A comprehensive survey', in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings*. doi: 10.1109/ICSTM.2015.7225422.

Hutter, M., Schmidt, J. M. and Plos, T. (2008) 'RFID and its vulnerability to faults', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-540-85053-3_23.

Jain, G. and Dahiya, S. (2015) 'NFC: Advantages, Limits and Future Scope', *International Journal on Cybernetics & Informatics*, 4(4), pp. 1–12. doi: 10.5121/ijci.2015.4401.

Lacmanović, I., Radulović, B. and Lacmanović, D. (2010) 'Contactless payment systems based on RFID technology', *MIPRO 2010 - 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, Proceedings*, pp. 1114–1119.

Ltd., A. C. S. (2017) *ACR 122U : USB NFC Reader*.

Madhoun, N. El, Bertin, E. and Pujolle, G. (2018) 'An overview of the EMV protocol and its security vulnerabilities', *2018 4th International Conference on Mobile and Secure Services, MOBISECSERV 2018*, 2018-Febru, pp. 1–5. doi: 10.1109/MOBISECSERV.2018.8311444.

Madhoun, N. El, Guenane, F. and Pujolle, G. (2015) 'A cloud-based secure authentication protocol for contactless-NFC payment', *2015 IEEE 4th International Conference on Cloud Networking, CloudNet 2015*, pp. 328–330. doi: 10.1109/CloudNet.2015.7335332.

Martin Emms, Budi Arief, J. H. and A. van M. (2013) 'POS Terminal Authentication Protocol to Protect EMV Contactless', *TECHNICAL REPORT SERIES: Newcastle University.*, No. CS-TR-(2), pp. 2–9. doi: 10.1145/335527.335528.

Mehrnezhad, M. *et al.* (2016) 'Nfc payment spy: A privacy attack on contactless payments', in *International Conference on Research in Security Standardisation*, pp. 92–111. doi: 10.1007/978-3-319-49100-4_4.

The Uk Cards Association (2017) 'Guide for retailers: Accepting contactless and higher value contactless payments'.

UK Finance (2019) '2019 Half Year Fraud Report'.

Vats, H., Ruhl, R. and Aghili, S. (2016) 'Fingerprint security for protecting EMV payment cards', in *2015 10th International Conference for Internet Technology and Secured Transactions, ICITST 2015*. doi: 10.1109/ICITST.2015.7412065.

Yum, J. W. *et al.* (2010) 'Smart card with an integrated electrical switch for secure operation', in *International Workshop on Antenna Technology: Small Antennas, Innovative Structures and Materials*. doi: 10.1109/IWAT.2010.5464687.

Online Fraud & Money Laundry in E-Commerce

Hasan Alsaibai[1], Shooq Waheed[2], Fatima Alaali[3], and Rami Abu Wadi[4]
[1]Associate Consultant, Cyber Security, Penetration Tester, KPMJ Fakhro, Manama, Kingdom of Bahrain
[2]Freelance Researcher
[3]American University of Bahrain, Manama, Kingdom of Bahrain
[4]Ahlia University, Department of Accounting and Economics, Manama, Kingdom of Bahrain
alsaibaih@gmail.com
shouqi. w7@hotmail.com
Fatema.alaali@aubh.edu.bh
rwadi@ahlia.edu.bh
DOI: 10.34190/EWS.20.045

Abstract: The Internet has allowed us access to information in the most efficient ways. However, it's not all rainbows and butterflies. The case of individuals who attempt to utilize this technology has been brief yet quick to escalate during this short period of time. This paper is aimed at providing insight on and to better understand how “Fraudsters” use tools provided by e-commerce organizations to attain legitimate withdrawals from financial institutions. This paper explains “Cyber Laundering” and presents the different methods that criminals utilize to launder money through online payment. Information provided in this study is accumulated through analysis of various academic as well as non-academic sources. Efforts to combat and propose countermeasures of cyber laundering and online fraud are served at the end of this study.

Keywords: Money Launderers, Cyber Laundering, Online Payment, Online Fraud

1. Introduction

The whole world has been dealing with financial cybercrimes for a while now and no country has an effective counter to dealing with all the problems that online connectivity poses. The middle east in recent years has been booming when it comes to internet activity. The Kingdom of Bahrain has most recently announced that it’s trying to be a financial technology center (FIN-TECH) in the Middle East.[1] During the past couple of years, the banking sector within Bahrain has evolved to embrace technology through the introduction of multiple electronic payment methods such as “BWallet”, “BenefitPay”, and the “Fawri” service that is provided by retail banks. The development of this sector started back in the 1970s, which after the conclusion of multiple wars in the region, took advantage from the flow of funds that were targeting oil booms and other business opportunities that are present in the region due to its geographic location. While the Bahraini Financial Sector has developed well in the recent years and has been able to turn into a regional banking center, it has failed to advance its legislative authorities and stay ahead of the curve when it comes to combating online fraud and money laundry. The current stance that the Central Bank of Bahrain is taking on tackling such issues is passive. This study examines several online fraud techniques that may be practiced within the kingdom of Bahrain and proposes several methods that align with Bahrain’s vision of becoming a Fin-Tech hub. We are examining one of the successful models of e-commerce systems which allows customer to customer (C2C) payments and examine how fraud takes place through it. This study will also examine the gateways that process online payments. For those kinds of services, one of the cornerstones that attract people is the ease and practicality of which an individual can start an auction. Intermediaries that provide the auctioning platform for customers bears the responsibility of securing payments between those individuals involved in the process. Protecting both seller and buyer is a burden that grows with each transaction that is made and this burden has been transferred in this case over to PayPal. To tackle this issue, PayPal validates the users’ account information by charging the bank account of the consumer. The amount is then sent back to the account as a verification step. The account is then created, allowing the customer to send or receive money by giving their email address and placing a sum of money as a payment or a receipt depending if they are selling or buying the auctioned item. Therefore, this study aims to answer the following question: what are the methods of online fraud that are accessible in Bahrain’s financial system?

[1] http://bahrainedb.com/business-opportunities/financial-services/fintech-and-payments/

2. Literature Review

The effectiveness of online payments has been a significant topic of increasing interests in the current years due to the increase in cash flow between individuals and corporations. The increased concerns regarding the security of those payment solutions must be reported before scandals occur. The main aim of this is to assess and evaluate those payments systems and the existing loopholes within them and provide solutions that can be turned into regulations that can further support the industry. This section produces a coherent background regarding the topic of this study.

At present, most of the studies are conducted to resolve the issue of anti-electron money laundering using a qualitative method. Whereas, a smaller number of studies examine this issue from a technical perspective or quota perspective. On one hand, Weibing (2011) put forward some proposals to the anti-electron money laundering's work. On the other hand, Yun et, al (2008) conduct the research specifically to handle the anti-electron money laundering from a technical perspective. They focus on the detection methods and supervision measures for the money laundering taking advantage of e-commerce based on third party online payment. Jun (2005) discussed the main problems of current anti-money laundering data reporting regulations, including enormous data, high false rate, lack of self-adaptive and easy evasion. Ping (2007) looked at measures on combating money laundering and terrorism financing in the People's Republic of China. Li-fang and Ping (2004) assessed Chinese financial institution campaign against money laundering.

Nowadays, Information and Communications Technology (ICT) is diffused and common and the trend towards digitization is developing (Gercke, 2011). Due to the fundamental parts of banks within the development and financial improvement of any country, it's very important to protect these companies and institutions from the tricks of fraudsters. In any case, it is the same ICT frameworks utilized by the banks which are adversely utilized by fraudsters. The expanded use of ICT such as computers, versatile phones, web and other related innovations are the courses which gave rise solutions as well as problems. The problems referred to are 'electronic crimes' which incorporates spamming, credit card fraud, ATM fraud, money laundering, identity theft, phishing and more (Siddique & Rehman, 2011; Bamrara et al., 2013).

In the subject of financial markets - specifically banking and finance – fraud can be identified as the false presentation of financial information and the misrepresentation of information given to a certain entity (Fligstein and Roehrkasse,2016). Financial information is of vital importance to every transaction that is done in the market. A trade of value may occur at any given point based on the information supplied by the market as well as the individuals participating in the transaction and any falsification of information within this transaction will cause a loss to one side of the transaction. Monitoring the supply of information in the market must be gatekept in order to protect the integrity. Authorities must issue regulatory rules that facilitate the pass through of information between entities that conduct transaction within a governate. Another segment of protection must be done through protection of market participants who are deemed to not have the knowledge to navigate the available information. The level of ICT awareness that has been on the rise and contributed positively towards the economic development. Banks play a major role in the development of any country economy. Hence, the protection of those organizations is of deep importance. In the same system that those banks function there is a force that destroys what is being brought up through the economic development. That is what fraud and embezzlement is. The destroying force hosts activities such as spamming, credit fraud and money laundering activities (Siddique & Rehman, 2011). The fraud conducted over ICT is referred to as electronics fraud. It is separated into two distinctive categories. Direct fraud which covers credit and debit card fraud as well as money laundering. Indirect fraud covers hacking and ransomware as well as methods that induce users willingly or unwillingly to pay ransoms and/or utilize their hardware resources illegally. Card fraud involves identity theft as it requires personal information such as a person name to complete a fraudulent transaction (Arthur, 2007). Further in this study we will examine how to tackle this issue. Auction sites today are in an uncomfortable position compared to where they started. Individuals in the Arab world have a preference of face to face transactions instead of virtual ones that require no contact. When they expose themselves to new technologies, they only explore them as part of their self-innovation and not because they are interested in the method. It's up to the intermediary to create and build that trust to push financial technology in the market (Rouibah et al., 2016).

An important part of this study goes over a new market which is the electronic gaming industry. The industry is currently at 500 million dollars and is expected to reach 1.65 billion by 2021. [2] The growth of fraud and money laundering within this industry has been going under the radar. This is because the international criminalization recommendations only took into account cash-generating techniques which were the traditional way of laundering money. Through that convention, conventional methods of money laundry were put in a conceptual manner so that they can be combated. However, the exploitation of techniques that are on the level of another social group who are gamers in this case was not given its due when it comes to analysis and study. This research covers that area and displays how fraud and money laundry occurs through online gaming.

2.1 Definition of Main Concepts

2.1.1 Money Laundering:

The term "money laundering" was first used in 1973 during the Watergate scandal and therefore there is no original legal definition but a colloquial paraphrase describing the process of transforming illegal into legal assets (Schneider and Windischbauer, 2008). Traditionally, money laundering has been considered to be a process by which criminals attempt to hide the origins and ownership of the proceeds of their criminal activities (Hopton, 2005). The aim is to enable them to retain control over the proceeds and to provide, ultimately, a cover for their income and wealth. Hopton (2005) further says that there is no one way of laundering money or property. There are ways ranging from simple methods of using it in the form in which it is acquired to highly complex schemes involving a web of international business investments. The process of money laundering is usually divided into three stages: placement, layering and integration. According to Haynes (1993), the first and most difficult step involves illegal funds in both financial and non-financial systems. For instance, this may be achieved through creating a series of small cash deposits, each of which is below the minimum level for reporting under money laundering regulations. Alternatively, the funds may be deposited in an account supported by an apparently legitimate business transaction. In the non-financial system, funds may be placed through real estate purchases and related transactions. Regardless of the method used, the purpose is to place funds in the economy in a way that does not arouse suspicion and thus minimize the risks of detection. Clearly, money launderers who do better at exploiting legitimate financial and non-financial transactions are less likely to be exposed (Hampton and Levi, 1999).

2.1.2 Liability

The following example demonstrates the liability for the C2C processing of funds and protection of the buyer and seller. When customer A decides to sell child pornography to Customer B using PayPal as an intermediary. Customer B sends Customer A an email payment that signals Customer B to forward the said content to the buyer. The intermediary institution will have no way of identifying the content of the purchase. In another scenario, if Customer A attempts fraud by purchasing an item using a stolen credit card and after that transaction is executed and the item is sent over to the owner of the stolen card charges back the fee and the payment is refunded from the account. PayPal, in this scenario, is liable. Therefore, intermediary organizations take measures to verify that the user of the submitted payment method is the actual owner.

2.1.3 Liability for C2C Scammers

In accordance to Consumer Credit Act (CCA), PayPal can be sued as the creditor. The debtor can only sue the creditor if there is a credit agreement between the creditor and the item provider. In such cases, PayPal will have to pay back the debtor if there is a misrepresentation of goods or a fault in the product that the manufacturer is found to be responsible for. PayPal in this respect seems to be ahead of such scams or errors as it has introduced a protection system for people who opt to purchase through its system. This preventive step protects buyers up to 350USD in selected purchases. This preventive measure also covers goods that never arrived or that do not match their supplier's description. Such measure places good faith in the consumer. However, occasional chargebacks and abuse of this system have been an ongoing issue.

2.1.4 Legal Status of C2C Payment Intermediaries

The business operation model of those organizations clearly displays how this kind of entities are creditors as well as being depository institutions. As local Bahraini law proposes as well as how major markets deal with this kind of financial units is by considering them - for regulatory purposes - banks. However, those entities tend to highlight the fact that they will hold funds in a separate manner from its own corporate funds. This highlights

[2] https://www.statista.com/statistics/490522/global-esports-market-revenue/

the fact that in case of bankruptcy such organizations will not use users' funds as a payment method to issued debt payments. It's important to highlight here that there is no law forcing those organizations not to deposit those funds into actual banks to generate a return. The US does not consider those institutions that function in a similar manner to PayPal to be banks. This is because according to the US regulations, an entity must physically handle the funds of the individuals in order to be a bank. This is also in addition to the fact that PayPal lacks a bank charter.

After examining how liability affects those entities, we will turn to the other side of the coin. What attracts fraudsters to e-commerce and the typology of online fraud as well as brief look at typical techniques in fraudulent transactions to launder generated money (Harrington, 2017).

2.2 Why Launderers choose E-Commerce

Companies that conduct their businesses through e-commerce solutions face several factors according to our study.

2.2.1 Anonymity & Security

The contributions that we have collected through our study regarding internet collectively agree that the internet is a haven for anonymity. Users can choose to opt for services that protect privacy. Starting off with installing a virtual private network (VPN). The IP addresses are cloaked in a manner that displays another address that is provided by the service provider. This blocks website and internet services from taping into individual search preferences and location.

Using VPN along with an encryption service, which is usually delivered by the same service provider, provides users to secure their transferred data and block interceptors of information in public places. This acts as a secondary firewall even if the public hotspot manages to pin the offensive device in the network. A third firewall level can also be implemented cheaply by using services such as "Net Filter Project". This provides any user with server level firewall that blocks any type of traffic that could be used by cybersecurity experts to ping a suspicious device when tracking down criminals.

2.2.2 Depersonalization of Payments

When conducting an online transaction from the moment of receiving the invoice to paying it off, the lack of human presence fully or partially within every transaction only incentivize criminals to find new ways of identity theft. Even with biometric technology booming, authentication of payments rarely requires such technology and even with it implemented there is still room for fraud. To further elaborate on this, biometrics create a formula around your physical print. Each fingerprint is simulated as a unique formula within the electronic device. If the formula matches, you are through. The danger behind such type of authentication is that biometrics are permanent. Changing such traits is either impossible or requires drastic measures. Therefore, it is only attractive to Money Launderers to utilize e-commerce since the lack of human interaction lessens the probability of a transaction looking fraudulent.

2.2.3 Expedite of Transactions

E-Commerce has enabled Money Launderers to transfer money faster over longer distances. This has crippled the law enforcement ability to efficiently track such transactions. Moving funds whether within one country or over national borders is faster than ever. This at its core makes laundering money easier but making it harder to trace.

2.3 Anonymous Payments

"Blockchain" and "Liberty Reserve" both function in a similar fashion. Payments barely require any personal information. Individuals can cash out even with Fake IDs. If a criminal decides to send USD500 to your account, your account will be given a random key which is like "961H535A". No party within that transaction can know who the other end is. This is done in addition to the encryption that is done to the transaction to protect any information related to the buyer, seller or the goods that are being traded (Lee et al., 2010).Examples of peer to peer money exchange services that currently support credit card companies: Perfect Money, OkPay, WebMoney, EGOPay, Payza, SolidTrust Pay, PayPal, Entro Money, PayNAT.

It has been three years since the Founder of Liberty Reserve Guilty of Laundering more than USD250 Million. The closing of Liberty Reserve was an important milestone in combating online money laundry and fraud.

However, it has moved the criminal activity to legitimate sites who do not want to affiliate with such criminal behavior. To dive in further on methods of micro money laundry we will tackle a few examples of how it is done.

2.4 Methods of Online Laundry & Fraud

We will elaborate on a couple of methods that money laundering has been thriving in.

- **Online Gaming:** Massive Multiplayer Online Games (MMORPGs) are a major player in this formula. They provide numerous methods to generate funds out of the online activity. Most of the fraudulent schemes require multiple accounts to move money.
- **Micro Money Laundering:** Sites like PayPal that act as intermediaries for C2C payments are hosting an increasing volume of micro laundering. Criminals use advertisements for jobs, entering private auctions and private event invitations as tools to move those funds. Micro Laundering works by dividing huge fund amounts into smaller pieces that are sent in a random manner to reduce the detection chance. Criminals tend to use multiple accounts as well. Relying on the same account is a red flag. The traditional source of this funds is a scammed bank account. A transaction is then done to move funds to a virtual card which later which takes the funds through a laundering process.

3. Tackling Online Money-Laundering

Since the Bahrain commercial sector applies UK based regulations, we will examine the Third EU Directive that deals with money laundry. To be more concise, this directive only targets casinos and gambling within the gaming industry. It also specifies what the code of conduct of these games should include highlighting criminal behavior. Businesses within this sector have introduced a wide variety of business models to face off with the risk associated with introducing games that could potentially facilitate this kind of criminal behavior. These corporations have decided to tackle risks and issues in a corresponding manner. They have deployed a hybrid anti-fraud model that utilizes multiple methods. The following methods are not the only deployed solutions. Advanced solutions that include artificial intelligence learning are also used in this sector but in a much narrower spread of organizations due to their cost. According to our respondents the following methods are some of the possible ways to prevent this issue:

1. **Manual:** Employees monitor transactions through a screen that reports the ongoing activity. Suspicious or unusual activity is flagged for further investigation.
2. **Third Party Data:** Data miners could pile up sources that could highlight popular online sites that lay down the corner stones of such criminal activity. Acquiring such data is essential to blocking such criminal activity spread. Information related to a website that provides support to such activity, prominent means and methods of fraud and laundry activity could contribute to the creation of new preventive measures.
3. **Pre-Configuration**: Depending on the activity of the organization, for example, auctions. The platform could require participants to have made previous confirmed purchases with multiple sellers before opening a dispute with a new seller. This could be a preventive measure to prevent fraud on those auction websites.
4. **Statistical Based:** Through monitoring, customer activity outliers can be detected using regression analysis of data that the customer inputs by using the method of payment.
5. **Artificial Intelligence:** Utilizing a neural-based network architecture that works on the analysis of transactional behavior in the whole system to detect fraudulent behavior trends.

3.1 Preventing Auction Fraud

Auction Fraud can be prevented by aggregating data between financial companies. E-Commerce companies and banks must open their functions and procedures to each other allowing their application interfaces to interact directly with each other. This allows better verification of information before allowing payments to pass. It will also block fraudsters from escaping. For example, if the PayPal account cases and status is shared directly with the bank, the negative balance that occurs after the auction fraud will carry over to the bank account. Using this method will block funds from being withdrawn from the automated teller machine and will also notify the bank.

3.2 Big Data for Bahrain Fin-Tech Companies

The local Fin-Tech industry in Bahrain needs to start utilizing their resources in order to combat fraud by aggregating data. The current systems are still exposed to the fraudulent methods that are presented in this study. Investing in big data will allow organizations to secure their business models. The volume of funds that is being lost because of e-commerce fraud will haunt those financial services and will force them to implement off-shelf prevention methods. The investment in fraud detection and fraud prevention systems in financial institutions will combat the growth of fraud that is estimated to be USD5.55 Billion according to the U.S Department of Justice. Implementation of such strategies will increase the confidence in fraud prevention mechanisms in the Kingdom. The Fin-Tech industry has to start the process of fighting back and they can do so through the following procedure:

1. Identifying existing problems and assigning research and development resources to solve them.
2. Enhancing data collection to ensure that the data analysis process is working with high-quality data.
3. Working with the regulatory institutions within the country to implement customer privacy laws that will enhance customers trust in those data collection procedures.
4. Ensuring that Finance and Technology institutions are properly coordinating with each other to implement big data without the disruption of entity activities. This is to make sure that business models are adapting to Big Data.

Overcoming the challenges that will face Fin-Tech organizations in the process of fraud prevention will have a resounding impact on company's customer services and will contribute to lower fraud loss.

4. Conclusion and recommendations:

As Bahrain is transitioning more towards online banking services through the introduction of B Wallet, Benefit Pay, and the use of Benefit gateway to process Visa and Mastercard payments through the internet. It is important to look at how fraud occurs in those services and how money laundry assists in wiping any trace of those fraudulent transactions. Our Fin-Tech sector must evolve to be on par with the rest of the world. The U.K has already passed legislation that defines how those organizations need to cooperate to prevent criminal activity. Bahrain needs to step-up its efforts if it wants to be a pioneer of Financial Technology in the region. The solutions proposed within this study require backend adjustments and updates with minimal disruption to customers. The GCC is implementing a new VAT system that utilizes block chain mechanisms in order to verify payments and gather information about transactions, with that kind of implementation we believe that the Central Bank of Bahrain (CBB) is capable of regulating e-commerce effectively. In light of the foraging results on the research subject matter, the study recommends the following:

1. Timely implementation of system recommendations must be put in place to avoid leakage of VAT in the upcoming period along with those fraudulent transactions in sensitive areas such as prepaid cards and vouchers.
2. A fast-moving industry such as financial technology requires fast moving legislators. More needs to be done from a legislative standpoint to enforce proper practices.
3. Enhancing data collection in the Kingdom to ensure that the data analysis process is working with high-quality data.
4. A cloud-based system such as AWS to aggregate data in the same way that the Unified GCC VAT agreement suggests is optimal to close information dead zones in the Fin-Tech industry.

References

Arthur Williams, D.: Credit card fraud in Trinidad and Tobago. *Journal of financial crime*, *14*(3), 340-359 (2007).

Bamrara, D., Singh, G., & Bhatt, M.: Cyber Attacks and Defense Strategies in India: An Empirical Assessment of Banking Sector. *International Journal of Cyber Criminology*, 7(1), 49-61 (2013).

Fligstein, N., & Roehrkasse, A. F.: The Causes of Fraud in the Financial Crisis of 2007 to 2009: Evidence from the Mortgage-Backed Securities Industry. *American Sociological Review*, *81*(4), 617-643 (2016).

Gercke, M.: Understanding cybercrime: a guide for developing countries. *International Telecommunication Union (Draft)*, *89*, 93 (2011).

Hampton, M. P., & Levi, M.: Fast spinning into oblivion? Recent developments in money-laundering policies and offshore finance centres. *Third World Quarterly*, *20*(3), 645-656 (1999).

Harrington, C.: Why Fintech Could Lead to More Financial Crime. *CFA Institute Magazine*, *28*(3) (2017).

Haynes, A.: Money Laundering and Changes in International Banking Regulation. *Journal of International Banking*, 456 (1993).

Hopton, D.: Money Laundering: A Concise Guide for All Business, Gower, Aldershot (2005).

Jun, T. A. N. G.: An Anti-Money Laundering Data Monitoring and Analytical System Based on Customer Behavior Pattern Recognition [J]. *Journal of Zhongnan University of Finance and Economics*, *4* (2005).

Lee, B., Cho, H., Chae, M., & Shim, S.: Empirical analysis of online auction fraud: Credit card phantom transactions. *Expert Systems with Applications*, *37*(4), 2991-2999 (2010).

Li-fang, Z., & Ping, H.: The Chinese financial institution campaign against money laundering. *Journal of Money Laundering Control*, *7*(2), 145-152 (2004).

Ping, H.: Comments on the law of the People's Republic of China on anti-money laundering. *Journal of Money Laundering Control*, *10*(4), 438-448 (2007).

Rouibah, K., Lowry, P. B., & Hwang, Y.: The effects of perceived enjoyment and perceived risks on trust formation and intentions to use online payment systems: New perspectives from an Arab country. *Electronic Commerce Research and Applications*, *19*, 33-43 (2016).

Schneider, F., & Windischbauer, U.: Money laundering: some facts. *European Journal of Law and Economics*, *26*(3), 387-404 (2008).

Siddique, M. I., & Rehman, S.: Impact of Electronic Crime in Indian Banking Sector: an overview. *International Journal of Business and Information Technology*, *1*(2), 159-164 (2011).

Weibing, P.: Research on money laundering crime under electronic payment background. *Special Issue: Research Findings in Computer Science-Technology and Applications Guest Editors: Qihai Zhou, Min-bo Li, and Zhongjun Li*, *6*(1), 147 (2011).

Yun, F. E. N. G., Chang, Y. A. N., Dong-mei, Y. A. N. G., & Jing-jing, Z. H. A. N. G.: Detection and supervision for money laundering in e-commerce based on third party online payment [J]. *Systems Engineering-Theory & Practice*, *12*, 006 (2008).

Cyber Humanity in Cyber War

Jorge Barbosa
Coimbra Polytechnique - ISEC, Coimbra, Portugal
jorge.barbosa@isec.pt

DOI: 10.34190/EWS.20.116

Abstract: The consideration of humanitarian aspects in conventional, kinetic wars is a subject widely considered and studied, and there is even regulation in order to protect people in times of armed conflicts. In particular, the so-called Geneva Conventions are a series of treaties formulated in Geneva, Switzerland, defining the rules for international laws concerning so-called international humanitarian law. In addition to the Geneva Conventions, there are also the so-called Hague Conventions of 1899 and 1907. They established at the First and Second Peace Conferences in the city of The Hague, Netherlands. They officially called the Convention on the Peaceful Resolution of International Disputes, with the initial version of 1899 existing and the 1907 version subsequently appearing. Together, these two Conventions constitute the first international treaties on laws and war crimes. The appearance of these conventions and other regulations resulted essentially from the visual testimony of the horrors practiced in war actions essentially against the civilian population and the consequent human suffering caused by such actions. The concern that leads us to consider a study on the humanitarian aspects related to cyber war actions and its prevention and regulation is the fact that, contrary to the actions of kinetic warfare, the actions of cyber war apparently does not produce visible damage. These not visible effects, can lead to the not consideration of the need to establish conventions, in the humanitarian perspective, for this type of war. The fact that cyber war actions may not cause so much visible physical damage and through these damages or by the direct action of the actions performed cause great human suffering, does not mean that this suffering does not exist in an equal or perhaps even greater degree than the one caused by the actions of kinetic war. These damages and sufferings may simply not be as visible or may be masked. This "physical" non-visibility of these damages can lead to a lack of awareness of them and the related suffering and, as such, not to raise awareness for their regulation and mitigation, especially with regard to humanitarian aspects related to the protection of civilians populations. Therefore, this is the concern underlying this article and the possible need to establish a priori similar humanitarian regulations for cyber war actions. Such humanitarian regulation necessarily must be close linked to the legal regulation of these cyber war actions, but it should not necessarily be replaced or deprecated by such legal regulation. It must have its own space because it derives from a different framework, a framework with a view to the humanitarian well-being of civilian populations affected by cyber wars.

Keywords: Humanity, Cyber Humanity, Cyber War, Cyber Conflicts

1. Introduction

In the literature, there are several publications regarding ethical and legal aspects related to cyber warfare. However, the approaches taken within the aforementioned framework do not necessarily cover the due and necessary humanitarian framework. There must be a framework that reflects the concerns related to the suffering and deprivation that civilian populations may experience in the event of cyber warfare and their necessary consideration, mitigation and help in such situations.

The humanitarian framework related to cyber actions should not necessarily be a very different framework of the legal and ethical frameworks, but should not be lessened, not subjugated, and not neglected by the other important ethical-legal aspects. In this way, we believe that there should also be a concern with the establishment of this framework and in that sense a Universal Humanitarian Convention related to the effects of cyber actions, especially cyber wars, should be analyzed and implemented.

In addition to the difficulties in establishing international legal and ethical legislation for the regulation of this type of cyber actions, which we will analyze in the next sections, the establishment of a universal human convention will also suffer from these same or similar difficulties. These difficulties may also be due to specific conditions related to humanitarian issues.

So the aim of this paper is not the legal aspects of cyber war, namely the persecution of the attackers, the legal convention, but the consequences of the not perception of the cyber actions in the civil populations!

2. Ethical and Legal Regulation of Cyber Actions

In the absence of a clear international legal framework regarding the so-called cyber war actions, most of studies on this subject seek to make, by parallelism, an adaptation and extension of the existing legal provisions for conventional kinetic wars to cyber war, thus seeking norms and similar laws for cyber war. See in (Casimiro, 2018) (Evans, 2005) (Joyner & Lotrionte, 2001) (Delibasis, 2002) (Liaropoulos, 2010) (Mariarosaria Taddeo, 2017) (Rauscher, 2013) (San Remo Manual, 12 June 1994) (Seviş & Seker, 2016) (The Program on Humanitarian Policy and Conflict Research at Harvard University, 2013).

Thus, in practice and in reality, there is still no real international law for the regulation of acts practiced in cyber wars, as is the case for conventional kinetic wars. The various existing initiatives aimed at establishing an international convention of general and universal application, have encountered several difficulties.

Some of these difficulties related to technical issues underlying the cyber war itself, for example to the fact that they take place in cyberspace. The lack of physical boundaries in this environment, the absence of geographic distances, the possibility that attacks can almost launched from any side and in a more or less hidden way, brings problems that are not easy to be legally framed. One of the main technical aspects is the difficulty in identifying the aggressors. This difficulty makes it very difficult to prosecute and punish the offenders of an eventual international convention.

Issues of a diverse political nature, to which the fact of trying to establish this legal framework without being than in a post-cyber war scenario makes it difficult to do so. It should be remembered that most of the current international legislation applicable to conventional kinetic war, appeared, in its first versions, because a result of major wars, regional or world. In this context, there were political conditions for its approval or establishment, either by the populations , and consequently the governments, to be sensitized in the face of the horrors experienced, either because there were predominantly winning countries and vanquished countries and this context was favorable then to the establishment of rules and in particular to the establishment of international rules.
The concepts of *just war theory*, namely *jus ad bellum*, *jus in bello* and jus post bellum, which are behind the establishment of regulations for conventional kinetic warfare, also tend to be applied in the specific case of frameworks for cyber war actions, (Evans, 2005). However, even more limited conventions or regulations in their scope, such as the 2001 Convention on Cybercrime and its 2006 Additional Protocol or the 2009 Shanghai Cooperation Organization's International Information Security Agreement, have had relatively low adherence with many major countries not to adhere to them.

NATO through the CCD OE in Tallinn, Estonia, has established a set of legal standards that constitute the LoAC, The Tallinn Manual Law of Armed Conflicts, known as Tallinn Manual, (NATO CCD OE, 2013). According to the CCDOE website, the Tallinn Manual “is the result of a three year effort to examine how extant international legal norms apply to this “new form of warfare”. As any other LoAC Manual ever published, the Tallinn Manual emphasizes the fact that this does not represent a legally binding document but are the opinion of the group of experts, which participated in the redaction of the Manual, and does not represent the position of the CCDCOE nor NATO or any of its member States, .

To this extent, the Manual is indeed more a statement, a clarification and an articulation tool of the existing law, rather than a proper law-making process. By restating certain applicable norms and particularly Customary International Law, CIL, the Manual work cane slightly related to a legal authority; the States are not bound by the Manual itself but by CIL, (Juan, 2016).

According to (Vihul, 2013), in late 2009, NATO CCD COE invited a group of twenty international law scholars and operational legal advisers (the International Group of Experts), under the directorship of Professor Michael Schmitt of the United States Naval War College, to conduct a three year research project examining the norms applicable during cyber warfare. The product of this effort is the “Tallinn Manual on the International Law Applicable to Cyber Warfare”, published in March by Cambridge University Press. The Tallinn Manual’s primary focus is the jus ad bellum (the law governing the use of force) and *jus in bello* (international humanitarian law). To a lesser extent, it touches upon related areas of international law, such as the concepts of sovereignty and jurisdiction, as well as the law of State responsibility. The Manual is design as a reference tool for State legal advisors, policymakers, and operational planners, although scholars and students will find it useful as well. NATO

CCD COE has launched a three-year follow-on project, “Tallinn 2.0”, which will expand the scope of the Tallinn Manual. The Tallinn Manual is strictly an expression of opinions of the International Group of Experts and as such does not represent the official positions of the Center or NATO. This will also be the status of Tallinn 2.0. ”, (Https://www.ejiltalk.org/the-tallinn-manual-on-the-international-law-applicable-to-cyber-warfare/)

In addition to these initiatives, there have previously been others of a more restricted nature with the aim of regulating this issue, for example, The San Remo Manual on International Law Applicable to Armed Conflict at Sea, (San Remo Manual, 12 June 1994), The Manual on the Law of Non-International Armed Conflict With Commentary, (Schmitt, et al., 2006) or The Harvard Manual on the International Law Applicable to Air and Missile Warfare, (The Program on Humanitarian Policy and Conflict Research at Harvard University, 2013).

However, none of these legal frameworks can consider a regulation of international application with regard to the regulation of cyber actions, in particular of actions related to cyber war. In fact, they only translate concepts and opinions and they have not accepted or adopt by a large number of countries. In particular, they it are not yet recognized by world organizations, such as the United Nations, UN.

Although many countries and organizations recognize the need for such a framework, there remains, however, a legal vacuum at the global level with regard to the regulation of this type of cyber actions.

3. Establishment of a Universal Humanitarian Convention for Cyber War Actions

3.1 Why the Establishment of a Universal Humanitarian Convention for Cyber War Actions

From a humanitarian point of view, the main problem with cyber war actions is that they it can be interpreted as “*Bloodless”* actions, but cyber-attacks do not cause, directly and necessarily, "*blood*". These attacks are, from this point of view, much more subtle and clean! This aspect is very important in raising awareness of the need to establish humanitarian regulations for cyber war, such as the *IHL*, the *International Humanitarian Law*, for conventional kinetic warfare.

In conventional kinetic wars, the visual effects related to the physical destruction of buildings, industrial installations and even a large part of cities, create an image in those who see, for example, television reports and other media that create commiseration and compassion. The effect of this influence on observers assists these affected populations and, as a result, society, governments and international organizations are receptive and even conditioned to have to assist these populations, by offering direct humanitarian aid. Another consequence is the creation of legal mechanisms to contain and moderate this type of conventional war actions.

The impact of these visual effects of conventional wars was at the origin of the creation of the Red Cross. The emergence of this organization is associated to the Swiss Henry Dunant, who directly witnessed the horrors left by the *Battle of Solferino*, in northern Italy, on June 24, 1859. Moved by the situation of the wounded, Henry would write the publication *A Memory of Solferino* and would make a humanitarian appeal to society to organize in times of peace to provide relief in times of war. Dunant's lobby with political leaders has therefore led to the adoption of measures to protect the victims of war. Its two main ideas envisaged a treaty that would force armies to care for all wounded soldiers and the creation of national societies that would assist military health services. His experience in Solferino then gave rise to the International Red Cross and Red Crescent Movement, which today comprises the International Committee of the Red Cross (ICRC), 189 National Societies and the International Federation of Red Cross and Red Crescent Societies, (CIVC, s.d.).

The lack of visibility or not the so high visibility of cyber war actions due to the fact that it does not create these physical destructions with these visually striking and not so physically reproducible effects in television reports, can lead to the disregard of these actions and also the great suffering that they can create in affected civilian populations.

This fact is therefore very worrying a priori. The lack of existence of these not-so-impacting effects that as such may not create the same sentimental effects as conventional wars can then mask the real effects of cyber warfare actions that. Unlike conventional warfare actions, may even have a greater scope and dimension, because actions aren´t so localized, their areas of influence can be quite superior to those of conventional war.

They can have systemic effects on systems other than those directly affected by cyber weapons and deprive civilian populations of food and fuel by altering or destroying, for example, logistical distribution chains or even the inability to acquire such goods due to lack of resources, financial resources for this. This resource, which may be unavailable due to cyber-attacks on banks or other financial institutions, deprived of potable water due to the destruction of the respective treatment plants or deprived of electricity due to the allocation of the respective energy producing systems.

The combined effect of these actions can be catastrophic for the affected civilian populations! In this way, the establishment of a Universal Humanitarian Convention for Cyber War Actions is essential!

3.2 Difficulty in establishing a Universal Humanitarian Convention for Cyber War Actions

Further analysis of this issue leads to complex issues that may not be resolvable or easy to resolve. In fact, the targets of the cyber actions are not the computer system itself, but its data or indirectly the systems that these computer systems manage, in what possibly known as cyber to kinetic effects of the cyber weapons.

The computer systems that manage, for example, industrial systems do not necessarily need to be destroyed, but modified to cause disruptions to the physical systems they control. Examples of this are the cyber actions that became known as *StuxNet*, in which Iranian centrifuges purportedly used for the purification and concentration of materials for the production of nuclear weapons, were physically destroyed, (Zetter, 2014). Also relevant are the experiences carried out in 2003, in the national test program *SCADA*, *Aurora Generator Test* carried out at *Idaho National Lab, INL*, by the *American Department of Energy* (Clark & Knake, 2010), in (Zetter, 2014). In this test, it remains clear that it was easy and fast (Singer & Friedman, 2014) to destroy a generator similar to those existing in American hydroelectric power plants. In the event of a similar real action, orchestrated as an act of cyber war, a country's electrical generation system, or part of it, could easily be stop. These generators do not exist for immediate replacement, since they are *costume made* manufactured. If alternative energy continued to allow its production, it would only take approximately three months to build it beyond the time required for its transport and installation, (Barbosa, 2018). During this period, there could be a deprivation of electricity in a large area with the consequent major adverse effects.

This dual use nature of the cyberspace is a major obstacle to limiting and consequently establishing regulations for cyber war that prevent cyber actions that directly affect civilian populations.

The interconnection between civilian and military computer systems and more than that of civilian systems not necessarily linked to the military machine can be preferred targets for cyber-attacks. This is the case with dams, water treatment plants or power plants for the production and transformation of energy. These types of targets were already preferred targets for conventional kinetic warfare, but the increasing digitization of this type of infrastructure also makes them increasingly desirable, and in some ways easier, targets for cyber warfare.

In addition to these cyber to kinetic actions, we also have cyber to cyber actions, such as actions against banking or financial institutions or computer systems linked to governments. In the case of cyber actions, for example against banking institutions, civilian populations can be seriously affected by the destruction or alteration of bank data that may make it impossible for these populations to have means of payment that make it impossible to purchase food.

The destruction or the impairment of some data, e.g. financial data, can then cause human suffering due to the privations of basic means of subsistence. It can also lead to civil wars or even serious internal changes in public order, which can lead to chaos in countries, already affected by cyber actions and even contribute to the loss of sovereignty in those countries.

These impacting systemic effects on civil society, for example the affection of financial systems or the like, can spread to countries other than the initial cyber attacked, also causing deprivation and consequent suffering of the civilian populations of these other countries and lead to the involvement of these other countries in the war, either with cyber or conventional means.

3.3 Principles of a Universal Humanitarian Convention for Cyber War Actions

Despite this great need to establish rules humanitarian aspects related to the assistance to populations in the event of cyber war, the establishment of mechanisms of assistance to these populations and or the containment that of the harmful effects on people, almost nothing has been done at international level with a view to within reach of that goal.

In 2014, the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) publish an Occasional Policy Paper called "Humanitarianism in the Age of Cyber-warfare: Towards the Principled and Secure Use of Information in Humanitarian Emergencies", where the main issues related to humanitarian policies to be considered in cyber war scenarios, (OCHA, 2014).

The convention to be established for the creation of humanitarian protection and safeguard mechanisms must have a Universal nature, applicable to all nations and all types of population, whether they are military or civilians. It should include a scale so large that it can equally include care for the environment, for example, the non-physical extermination of nuclear power installations whose destruction could seriously contaminate the environment, animals and nature in general!

4. Conclusion

The need for the establishment of a Universal Humanitarian Convention for Cyber War Actions is evident and urgent!

Despite the apparent non-physical destruction caused by cyber warfare, it has great potential and a great capacity to cause human suffering by depriving civilians of goods essential to their livelihood and survival.

This deprivation may be due to the destruction by cyber means of critical infrastructures such as dams, power plants and energy transformation, management systems of banks and financial institutions.

Such a convention should be as comprehensive as possible and more than international, it should be universal in the sense that in addition to being a duty and having to be applicable to all and in all nations. It must be applicable to for both civilians and the military population, it must take into account against the effects of cyber weapons actions on the environment, animals and general nature.

Acknowledgement

I thank the Coimbra Polytechnic Institute – ISEC/IPC for the support and facilities granted.

References

Barbosa, J., 2018. Pequenas potências militares convencionais, Grandes potências militares cibernéticas - Abordagem da utilização de meios informáticos na defesa/ataque militar moderno, Lisboa: IDN.

Bunker, R., 2000. Weapons of Mass Disruption and Terrorism. Terrorism and Political Violence, Vol 12(No.1), pp. 37-46..

Carr, J., 2012. Inside Cyber Warfare. Second Edition ed. Sebastopol, CA, USA: O'Reilly Media, Inc..

Casimiro, S. d. V., 2018. Quadro Legal para a Cibersegurança e a Ciberdefesa. Em: I. d. D. Nacional, ed. Contributos para uma Estratégia Nacional de Ciberdefesa. Lisboa: IDN, p. 47~62.

CIVC, s.d. CICV. [Online]
Available at: https://www.icrc.org/pt/o-cicv/historia
[Acedido em January 2020].

Clark, R. A. & Knake, R. K., 2010. Cyber War: the next threat to national security and what to do about. New York: HarperCollins Publishers Inc..

Clemente, D., 2013. Cyber Security and Global Interdependence: What Is Critical?, London: The Royal Institute of International Affairs.

Delibasis, D., 2002. The Right of States to Use Force in Cyberspace: Defining the Rules of Engagement. Information & Communication Technology Law, 11(3), pp. 255-268.

Evans, M., 2005. Just War Theory. A Reappraisal. Edinburgh: Edinburgh University Press.

FLOWERS, A. & ZEADALLY, S., 2014. Cyberwar: The What, When, Why, and How. IEEE TECHNOLOGY AND SOCIETY MAGAZINE.

Joyner, C. & Lotrionte, C., 2001. Information Warfare as International Coercion: Elements of Legal Framework. European Journal of International Law, Vol 12(No.5), pp. 825-865.

Juan, G., 2016. The Challenges of Cyber Warfare to the Laws of Armed Conflict : Humanitarian Impact and Regulation Attempts, Glasgow: University of Glasgow.

Kostopoulos, G. K., 2013. CYBERSPACE and CYBERSECURITY. New York: CRC Press, Taylor & Francis Group.

Liaropoulos, A., 2010. War and Ethics in Cyberspace: Cyber-Conflict and Just War Theory. Proceedings of the 9th European Conference on Information Warfare and Security, 1-2 July, pp. 177-182.

Mariarosaria Taddeo, L. G., 2017. Ethics and Policies for Cyber Operations. EBook Edition ed. Switzerland: Springer.

MULVENON, J., 2005. Toward a Cyberconflict Studies Research Agenda. IEEE SECURITY & PRIVACY, Vol 3(No.4), pp. 52-55..

NATO CCD OE, 2013. Tallinn Manual on the International Law Applicable to Cyber Warfare. [Online]
Available at: https://ccdcoe.org/tallinnmanual
[Acedido em 10 February 2020].

OCHA, 2014. Humanitarianism in the Age of Cyber-warfare: Towards the Principled and Secure Use of Information in Humanitarian Emergencies, s.l.: OCHA.

Prelas, M. A., 2005. Terrorism and Weapons of Mass Destruction. s.l., s.n.

Ranger, S., 2018. What is cyberwar? Everything you need to know about the frightening future of digital conflict. [Online]
Available at: https://www.zdnet.com/article/cyberwar-a-guide-to-the-frightening-future-of-online-conflict/
[Acedido em 10 Fevereira 2020].

Rauscher, K., 2013. It's Time to Write the Rules of Cyberwar. IEEE Spectrum, 27 Nov.

San Remo Manual, 12 June 1994. San Remo Manual on International Law Applicable to Armed Conflicts at Sea. San Remo: International Institute of Humanitarian Law.

Schmitt, M. N., Garraway, C. H. & Dinstein, Y., 2006. The Manual on the Law of NonInternational Armed Conflict With Commentary, San Remo: International Institute of Humanitarian Law.

Seviş, K. N. & Seker, E., 2016. Cyber Warfare: Terms, Issues, Laws and Controversies. 2016 International Conference On Cyber Security And Protection Of Digital Services (Cyber Security), pp. 1-9.

Singer, P. W. & Friedman, A., 2014. Cybersecurity and Cyberwar - What everyone needs to know. N. Y,: Oxford University Press.

The Program on Humanitarian Policy and Conflict Research at Harvard University, 2013. HPCR MANUAL ON INTERNATIONAL LAW APPLICABLE TO AIR AND MISS ILE WARFARE, 32 Avenue of the Americas, New York, NY 10013-2473, USA: Cambridge University Press.

Thomas, J. T. & Mukta Sharma, 2019. Cyber Poison. International Journal of Computer Applications, April, pp. 50-54.

Tim Stevens, 2018. CYBERWEAPONS: POWER AND THE GOVERNANCE OF THE INVISIBLE. Accepted for publication in special issue of International Politics, on global prohibition regimes.

Vihul, L., 2013. The Tallinn Manual on the International Law applicable to Cyber Warfare, Tallinn, Estonia.: NATO Cooperative Cyber Defence Centre of Excellence.

Zetter, K., 2014. Countdown to Zero Day: Stuxnet and the launch of the world's first digital weapon". N.Y.: Crown Publishers.

Detecting a Rogue Switch using Network Automation

Vijay Bhuse and Vineet James
Grand Valley State University, Allendale, MI, USA
bhusevij@gvsu.edu
jamesvine@mail.gvsu.edu

DOI: 10.34190/EWS.20.068

Abstract: The problem of detecting a rogue switch is still unsolved. The rogue wireless access point problem has been solved. The wired rogue switch detection problem remains unsolved because of the implicit assumption that wired is safer. In this paper, we apply core networking concepts and demonstrate the effectiveness of our solution by combining the latest Automation techniques with highly effective software toolsets available for detecting malicious systems connected to a rogue switch. This solution promises quick detection and requires Zero Downtime which could prove to be an ideal solution for enterprises having managed switch production networks. We achieve this by continuously filtering and analyzing network traffic for any broadcast storms or new Address Resolution protocol packets using Packet Analyzers and then effectively tracing the malicious host connected to the rogue switch by deploying automation techniques. This technique also helps detecting rogue unmanaged switches ("plug and play" devices) having pre-loaded configuration.

Keywords: Rogue switch, LAN, SNMP, ARP, VLAN, Detection

1. Introduction

We begin with looking at two typical types of switches in the Network Infrastructure – Managed and Unmanaged. The key difference between them lies in the fact that a managed switch can be configured and it can prioritize LAN traffic so that the most important information gets through.

An unmanaged switch on the other hand behaves like a "plug and play" device. It comes pre-configured and simply allows the devices to communicate with one another. They do not have an IP address and some also have no MAC address which make them very difficult to trace.

In this article, we initially discuss what are the challenges that network administrators face to detect an unmanaged rogue switch on the network (Prins et al, 2018 and Bhuse et al, 2019). Then we discuss a network-based solution collecting evidences to finally reach a conclusion that a rogue switch is present.

This smart solution is implemented and demonstrated by combining the latest Automation techniques with highly effective software toolsets available for effectively detecting and tracing malicious systems connected to a rogue switch. This paper is an extended version of our internal publication (James, 2019).

2. Challenges to detect rogue switch

Following are some of the reasons why we think it is challenging to detect a rogue switch.

2.1 Unmanaged switches are untraceable

An unmanaged switch comes pre-configured and simply allows the devices to communicate with one another. These are Datalink Layer (referring to the OSI layer model) devices and do not have IP addresses and therefore ICMP tools such as traceroute cannot be used to trace the device. Moreover, some unmanaged switches also do not have MAC address which means there will be no entry of the switch on any MAC Address table of other switches. This makes the job of a network administrator who is trying to trace the rogue switch extremely difficult.

2.2 Discovery Protocols are ineffective

Switches are Layer 2 devices, so we can use discovery protocols to monitor directly connected neighbors and update connectivity status. The scope of this paper is to define a protocol and management elements, suitable for advertising information to stations attached to the same IEEE 802 LAN. However, protocols such as Link Layer Discovery Protocol (LLDP) (802.1b, 2009), Cisco Discovery Protocol (CDP) (CDP, 2016) will not be able to retrieve information about the connected rogue switch as most unmanaged switches do not support

Discovery protocols, which means the managed switches CDP or LLDP updates won't be acknowledged by the directly connected rogue switch and would totally bypass it.

2.3 Even Spanning Tree BPDUs can't help detect rogue switches

Bridge Protocol Data Units (BPDUs) are frames that contain information about the Spanning tree protocol (STP), which helps avoiding loops in the network. Managed Switches send BPDUs for STP algorithms to function, the switches need to share information about themselves and their connections. What they share are bridge protocol data units (BPDUs) (Cisco, 2017).

Typically, a BPDU guard is configured on the ports of managed switches to examine STP and disables the port which receives BPDU packets if it's not coming from a legitimate switch. However, Unmanaged switches do not have support STP and hence do not send BPDUs making this possibility of tracing it ineffective.

2.4 IP sweep tools cannot capture rogue switch location

Another possible way to detect a rogue switch is to somehow get some information about the hosts that are connected to that rogue switch. There is a possibility of running an IP sweep using tools in the entire LAN and examine the output for any anomalies in the network. These results can be utilized by the Network administrator to get clues on where to find the rogue switch and its actual physical connection on the network.

Today, we have such IP sweep tools available to us like NMAP (Network Mapper) which is a security scanner which can be used for this purpose. It is used to discover hosts and services on a computer network, thus building a "map" of the network. To accomplish its goal, these tools sends specially crafted packets to the target host(s) and then analyzes the responses (Namp, 2020).

Performing an IP sweep with such tools seems to be a good option, however, if you don't know what you are looking for it will be almost impossible to find. The problem with unmanaged switches is that there is no way to get information out of them as it has no management IP or supports SNMP. This makes finding the IP of the host connected to that rogue switch among a large pool of IP addresses impossible. The rogue switch remains entirely invisible to the mapping tool and does not show up in the topology map simulated by i**t.**

3. Concepts leading to the proposed solution

The solution that this article proposes is based upon certain networking concepts like IP Address Conflict Detection (ACD) (Cheshire et al, 2008)., Address Resolution Protocol (ARP) (Plummer, 1982). and Broadcast packets in addition to Wireshark tool which is a packet sniffer that we shall use to analyze frames on the LAN. Before we actually work on the solution for detecting a rogue switch on the network, let's understand few Network concepts that we will be using that lead us to the solution.

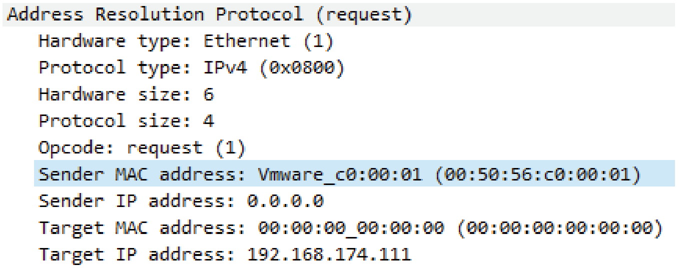
3.1 ACD, ARP-Probe and ARP Announcement

ACD is a utility of IPv4 which is used to avoid IP conflict in a LAN environment. This applies to all IEEE 802 Local Area Networks (LANs) [802], including Ethernet [802.3], Token-Ring [802.5], and IEEE 802.11 wireless LANs.

The ACD utility in IPv4 uses 'ARP Probe' a term used to refer to an ARP Request packet, broadcasted on the local link. Before beginning to use an IPv4 address (whether received from manual configuration, DHCP, or some other means), a Network Interface Card in a host must test to see if the address is already in use, by broadcasting ARP probe packets.

An ARP Probe conveys both a question ("Is anyone using this address?") and an implied statement ("This is the address I hope to use.").

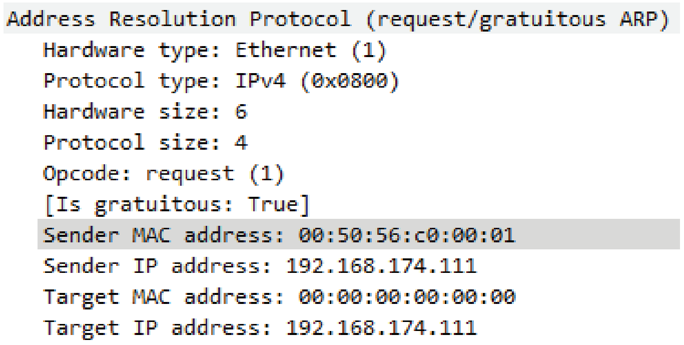
ARP Probe packets are broadcasts with an all-zero 'sender IP address'. The 'sender hardware address' MUST contain the hardware address of the interface sending the packet. The 'sender IP address' field MUST be set to all zeroes. The 'target hardware address' field is ignored and SHOULD be set to all zeroes. The 'target IP address' field MUST be set to the address being probed. ARP Probe header as illustrated in Figure 1 below.



```
Address Resolution Protocol (request)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (1)
    Sender MAC address: Vmware_c0:00:01 (00:50:56:c0:00:01)
    Sender IP address: 0.0.0.0
    Target MAC address: 00:00:00_00:00:00 (00:00:00:00:00:00)
    Target IP address: 192.168.174.111
```

Figure 1: ARP Probe

This process sends 3 ARP Probes, and if no one responds, the host officially claims the IP address with an ARP Announcement. Finally, the IP address is mapped to the physical address of the Source host. Figure 2 below illustrates the ARP Announcement header when the host finally maps the physical address and tries to confirm if it has set a unique IP address on the LAN.



```
Address Resolution Protocol (request/gratuitous ARP)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (1)
    [Is gratuitous: True]
    Sender MAC address: 00:50:56:c0:00:01
    Sender IP address: 192.168.174.111
    Target MAC address: 00:00:00:00:00:00
    Target IP address: 192.168.174.111
```

Figure 2: ARP Announcement

3.2 Switch behaviour when connected to a network

The switch table is initially empty. For each incoming frame received on an interface, the switch stores in its table the MAC address in the frame's *source address field*, the interface from which the frame arrived. In this manner the switch records in its table the LAN segment on which the sender resides. In case, the Destination MAC address is not found in the table, the switch forwards copies of the frame to the output buffers preceding *all* interfaces except for incoming interface. In other words, if there is no entry for the destination address, the switch broadcasts the frame (Kurose et al, 2013). As soon as the host tries to initiate a connection to other hosts, the switch tries to learn MAC addresses from the broadcast messages.

3.3 SNMP for monitoring

Simple Network Management Protocol (SNMP) is an Internet Standard protocol for collecting and organizing information about managed devices on IP networks and for modifying that information to change device behavior. Devices that typically support SNMP include cable modems, routers, switches, servers, workstations, printers, and more (Mauro et al, 2001).

SNMP-managed network consists of two key components:

- Agent – software which runs on devices that we want to monitor. This can be configured in such a manner that it sends Trap signal to the SNMP Manager reporting any unknown behavior on it.
- Network management station (NMS) – software which runs on the manager. A network management station executes applications that monitor and control managed devices.

4. The solution

This section covers how we can collaborate all concepts defined and explained in the previous section and co-relate evidences gathered using those concepts to finally detect a Rogue switch in a network.

Let us assume a network design as shown in Figure 3, where we see 4 managed switches S1, S2, S3, S4 are connected, distributing the network into 3 departments. Here, assuming no Port-security features are configured on the managed switch, meaning that anyone can walk in and plug in an unmanaged switch in one of the ports. Here we consider a Rogue switch is plugged in port 5 of Switch S4, as illustrated in Figure 3 below.

We also observe from Figure 3 (circled and connection marked in 'red') that three hosts are connected to the rogue switch which are trying to consume the entire bandwidth of the LAN by downloading heavy files using Bit-Torrent from the Internet.

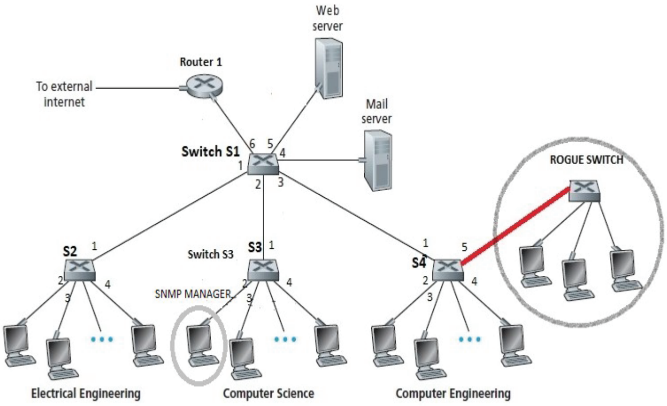


Figure 3: Network Topology

In this design, the internal LAN is connected to the Internet via a Router R1 and the host connection go through the router to establish a connection with the outside world.

The following steps were implemented and demonstrated on a real network emulator under a testing environment setup. The implementation details will be discussed in the next section. GNS3 was used to setup the network topology, configured with real Cisco images. The solution was tested on a simple network as illustrated in figure 3 and also on a multi-VLAN complex network which also displayed the same results.

4.1 Step 1 – Detect Brodcasts using SNMP

As already discussed in the above section, we know that a switch's CAM table initially has zero entries when it is plugged into an existing network unless a host gets connected to any port of that switch and tries to communicate. As soon as a host is plugged into one of rogue switch's port, it will start broadcasting traffic to all ports as it has no MAC entries in its CAM table. This causes a broadcast storm on the network which is received by all switches in the network.

All legitimate switches in the network are configured as SNMP agents and pick up these broadcasts. The agents are configured to report Broadcast storms in the network and send Trap signals to the SNMP Manager (As in Figure 3) on the network.

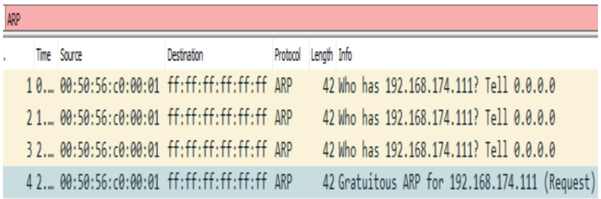
These SNMP Trap messages will alert the Network Administrator. This broadcast storm indicates that a switch has been connected on the network, however, more evidence will be needed to find the physical location of the Rogue switch.

4.2 Step 2 – Using Wireshark to examine ARP Probes

We also discussed about ARP Probes and ARP announcements in the Section above. To trace the Rogue switch, we need to first trace the IP addresses of hosts that are connected to the rogue switch. We need to know exactly what we're looking for on this busy network.

We use Wireshark to sniff all ARP Probes around the time since we received an SNMP Trap for broadcasts in the network. Filtering ARP packets and looking for ARP Probes will give us the most recently configured IP addresses on the network, one of which can trace us back to the Rogue switch.

Figure 4 illustrates the Wireshark trace showing 3 ARP Probe packets followed by an ARP announcement which shows the host attains an IP of 192.168.174.111.



```
ARP
  Time Source            Destination       Protocol Length Info
1 0... 00:50:56:c0:00:01 ff:ff:ff:ff:ff:ff ARP      42 Who has 192.168.174.111? Tell 0.0.0.0
2 1... 00:50:56:c0:00:01 ff:ff:ff:ff:ff:ff ARP      42 Who has 192.168.174.111? Tell 0.0.0.0
3 2... 00:50:56:c0:00:01 ff:ff:ff:ff:ff:ff ARP      42 Who has 192.168.174.111? Tell 0.0.0.0
4 2... 00:50:56:c0:00:01 ff:ff:ff:ff:ff:ff ARP      42 Gratuitous ARP for 192.168.174.111 (Request)
```

Figure 4: Packet capture

Now, that we the host whose ARP probes corresponds to the same timestamp as the broadcast signal. We can use that IP to trace the Rogue switch.

4.3 Step 3 - ARP table at the Router

From the uplink router, it being a managed device the ARP entries can be seen for that specific IP as it has been downloading files from the Internet using Router R1.

Once we get the MAC address of the Host, we can trace the switch going downwards from the Router to the switches.in their forwarding databases. The MAC found on a switch link can be traced by following the link to the next switch until you find the access-port on which the MAC is listed. This is where the malicious rogue switch is connected.

5. Implementation Details

GNS3 network emulator software (GNS3, 2019) and VM used for creating a virtual network topology using actual Cisco images.

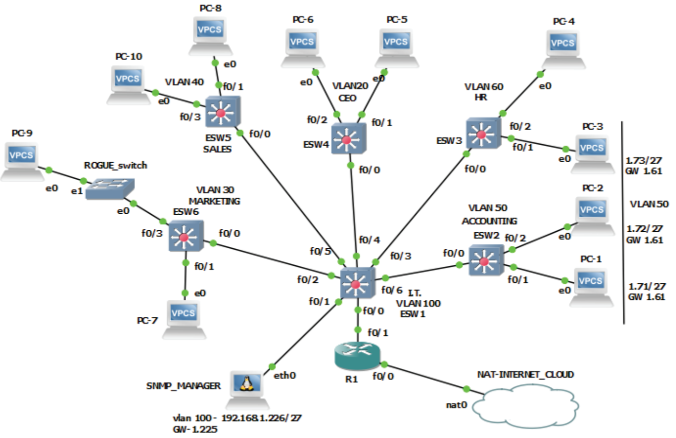
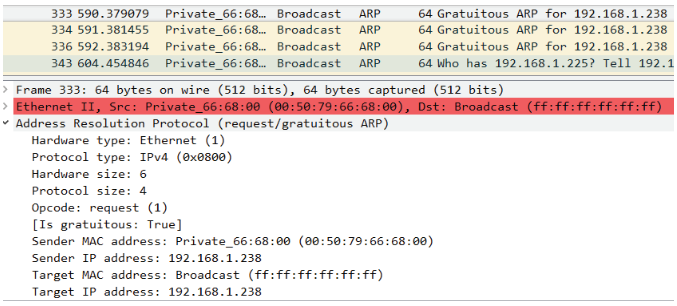


Figure 5: GNS 3 – Multi-VLAN Network topology

Wireshark packet analyzer used for network monitoring and analyzing ARP and broadcasts. Monitoring of special ACD (Address Conflict Detection) packets proved vital in finding out the host IP connected to the rogue switch.



```
333 590.379079  Private_66:68… Broadcast  ARP   64 Gratuitous ARP for 192.168.1.238
334 591.381455  Private_66:68… Broadcast  ARP   64 Gratuitous ARP for 192.168.1.238
336 592.383194  Private_66:68… Broadcast  ARP   64 Gratuitous ARP for 192.168.1.238
343 604.454846  Private_66:68… Broadcast  ARP   64 Who has 192.168.1.225? Tell 192.1

> Frame 333: 64 bytes on wire (512 bits), 64 bytes captured (512 bits)
> Ethernet II, Src: Private_66:68:00 (00:50:79:66:68:00), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
v Address Resolution Protocol (request/gratuitous ARP)
    Hardware type: Ethernet (1)
    Protocol type: IPv4 (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (1)
    [Is gratuitous: True]
    Sender MAC address: Private_66:68:00 (00:50:79:66:68:00)
    Sender IP address: 192.168.1.238
    Target MAC address: Broadcast (ff:ff:ff:ff:ff:ff)
    Target IP address: 192.168.1.238
```

Figure 6: ACD packet capture

Python Scripts were implemented using the Netmiko library (Byers, K. et al 2016), for configuring Cisco devices. Scripts were pushed on the Docker cloud container acting as the management system to automatically Trace the malicious system downwards from the uplink L3 device having its own ARP table using Python automation scripts.

6. Conclusions

We successfully demonstrate how to detect a rogue switch on the wired network, however, there is a lot of future scope for enhancing this solution into a fully functional centralized management tool for effective network management and with advanced fault detection features.

Plans are in place to continue building a fully functional centralized management tool with advanced GUI features. GUI features would include smart network configuration, backups, management and fault detection capabilities of network equipment. Full scale Integration with hardware networking equipment. A Wireshark plug-in could be developed providing full integration with the platform.

References

802.1ab (2009). IEEE LAN/MAN Standards Committee, “Station and Media Access Control Connectivity Discovery”. IEEE, Std. 802.1ab, 2009.

Bhuse, V., Kalafut, A., & Dohn L. (2019). “Detection of a Rogue Switch in a Local Area Network”. International Conference on Internet Monitoring and Protection, Nice, France.

Byers, K. (2016). “Multi-vendor library to simplify Paramiko SSH connections to network devices”, https://github.com/ktbyers/netmiko, Accessed January 2020.

CDP (2016). “Cisco Discovery Protocol”, https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/cdp/configuration/15-mt/cdp-15-mt-book/nm-cdp-discover.html, Accesses Feb 2020.

Cheshire, S. (July 2008). "RFC 5227 - IPv4 Address Conflict Detection". Internet Engineering Task Force.

Cisco (May 2007). "Configuring Spanning Tree Protocol". Retrieved 2014-06-10

GNS3 (2019). GNS3 – official documentation, February 2, 2019

James, V. (2019). “Network Automation Methodology for Detecting Rogue Switch”, https://scholarworks.gvsu.edu/cistechlib/337/, Accessed Februray 2020.

Kurose, J. F. and Ross, K. W. (2013). “Computer Networking: A Top-Down Approach”, 6th Edition, Pearson, ISBN13: 9780132856201.

Mauro, D. R. & Schmidt, K. J. (2001). “Essential SNMP (1st ed.)”. Sebastopol, CA: O’Reilly & Associates.

Nmap (January 2020). “The History and Future of Nmap”. Nmap.org. Accessed January 2020.

Plummer, D. (1982). “An Ethernet Address Resolution Protocol”, http://www.hjp.at/doc/rfc/rfc826.html, November 1982.

Prins K., Bhuse V. (2018). “Forced Vacation: A Rogue Switch Detection Technique”, 17th European Conference on Cyber Warfare and Security, Oslo, Norway.

An Investigation into Evidence (Digital) Artefacts Resulting from the use of Cryptocurrency Applications

Stephen Blenkin, Gareth Davies and Peter Eden
University of South Wales, Treforest, Pontypridd, UK
stephen.blenkin@btinternet.com
gareth.davies@southwales.ac.uk
peter.eden@southwales.ac.uk

DOI: 10.34190/EWS.20.096

Abstract: Is cash still king? With card payments now overtaking cash for the first time ever as reported in media 2017 (morningadvertiser, Accessed 2019), coupled with the rise of 1000s of cryptocurrencies, does this mean that cash now takes a back seat? Electronic transference of currency from one bank account to another anywhere in the world is not new, and with the strict guidelines that the banking institutions adhere to these transactions can be tracked from source to destination. However, some national banks; and indeed, some of the Governments that regulate them are not trustworthy. Chaum (1982) proposed 'blind' cryptography which protects payees' anonymity from invasion. Criminals commit their crimes for profit, therefore the digital currency is a perfect place for criminals to hide their financial affairs. Payments being made as a result of money laundering, blackmail, extortion or tax evasion, can be moved seamlessly anywhere in the world with few or no records tying an individual to 'monies'. A digital wallet where you store and commit transactions is used as an interface to your cryptocurrency. Creating a digital wallet can be done anonymously and does not need 'linking' to any financial institutions. A digital wallet allows purchasing of goods and services anywhere in the world anonymously by using your digital wallet's address, which is a unique digital key. Currently Microsoft Windows could be classed as the de-facto operating system worldwide, with Windows 10 being the latest version. Most owners are all too familiar with installing and uninstalling programs. Once the application has been uninstalled can analysis still identify any digital artefacts left on the host computer?This paper proposes a virtual step-by-step guideline for a forensic analysis to identify quickly the usage of digital a wallet, even if the application has been removed from the computer under investigation.

Keywords: Digital Wallet, digital wallet address, digital forensic, bitcoin, Windows 10 taxonomy

1. Introduction

Early in 2019 a threatening email was sent to an employee of a construction company demanding money with menaces. An online Bitcoin Wallet Address was included in the email for payment. At the time, Microsoft Windows was classed as the de-facto operating system worldwide with Windows 10 being the latest version. Most owners are all too familiar with installing and uninstalling programs. By forensically analysing a Windows 10 computer hard drive investigators can find evidence (digital) artefacts that can identify usage of a digital wallet application and emails, tying the user of the computer with activities. These artefacts can be found at numerous locations throughout the hard drive and can be advantageous in forensic investigations.

This study is an investigation into where evidence (digital) artefacts from digital wallets reside, as it must be noted that digital wallets are not all secure and written to the same standards, processes and file locations on a Windows 10 Operating System. However, most typically follow a similar process; and a wallet.dat file contains encrypted data which is the private key and unique digital wallet address (Heid, 2013).

Files reside within the Windows operating system,in order to speed up program execution (Singh and Singh 2018). They describe the taxonomy of program execution artefact structure as:

- User artefacts
  - IconCache.db
  - Jump Lists
  - Shortcut

- Systems artefacts
  - SRUDB.dat
  - Prefetch
  - Amcach.hve

- Registry artefacts

- Sam
- Security
- Software
- System
- Default
  - NTUSER.dat
  - UsrClass.dat

This research will focus on the existence of a digital wallet's address in plain text, before switching into an investigation of the installation/uninstallation of the digital wallet applications themselves. This will be achieved with standard forensic tools (AccessData 2019, Sleuthkit 2019, Thumbcache 2019)

This paper is organised as follows: Section 2 is a brief introduction into the forensic tools used in the investigation. Section 3 is an introduction of digital wallets, Bitcoin and how data was gathered for experiments and analysis. Section 4 contains analysis of the digital images obtained. Conclusions are then drawn in Section 5.

2. Forensic Tools

Forensics is all about finding out how and what happened. Throughout this text the word 'image' is being used frequently. An 'image' is an exact bit for bit duplication of any digital media at a given time, in this case a Windows 10 computer hard drive.

For this investigation FTK Imager will be used to capture a forensic digital image required for analysis. Once you have a bit-for-bit digital replica of the media in question, analysis can then take place. For this investigation Autopsy will be used as the principal open source application for investigation and analysis work.

Some Windows 10 registry files are stored in non-human readable formats, therefore a Registry Explorer and/or 'RegRipper' application is required to form this data into a human readable format. There are also several other Windows 10 files to be discussed further in this text that need to be translated into a human viewable structure, including ThumbCache and Prefetch files, discussed in sections 2.4 and 2.5.

When dealing with a significant amount of data in text format; albeit human readable; human error will always be a factor. Therefore, when carrying out analysis a simple text, a comparison tool is used to eliminate any possible chance of errors occurring.

2.1 FTK Imager

In order to preserve data integrity at any given moment a forensic image of digital media must be taken. This is to ensure that evidence cannot be tampered with before analysis. AccessData's Forensic ToolKit or FTK Imager captures the digital state of many forms of media (AccessData, 2019). In this research the same computer hard drive is imaged at various stages throughout the investigation. FTK can image a disk and, for the purpose of this investigation, store it in an E01 format. E01 is an abbreviation for "Encase Image File Format" which is also commonly referred to EWF or Expert Witness Format (Forensicsware, 2019).

2.2 Autopsy

The main analysis in this investigation will employ use of The Sleuth Kit's Autopsy application. Autopsy has an array of features, including the ability to import E01 images, captured using FTK. Autopsy constructs its own database out of the E01 files imported and presents them in a 'Windows' like file structure for analysis (sleuthkit, 2019)

2.3 Registry Explorer/Ripper

Windows 10 has a collection of registry files which record system information. These are not stored in human readable format. Some of the more common ways to decipher this information is to use an open source Registry Explorer and/or a RegRipper. For this investigation Registry Explorer v1.3.0.0 and RegRipper 2.8 are utilised. (Kristensen, 2020)

2.4 Thumbcache Viewer

Thumbcache, or Iconcache, is a built-in database within Windows 10 containing a collection of .bmp, .png or .jpg files. These files are basically images used to represent the icons a user typically clicks on to open an application. There are two major issues with investigating Thumbcache and Iconcache; extracting the files from a database; and interpreting what the individual files actually are. A Thumbcache viewer is used within this investigation simply to extract the files from the database and show what the image actually is (Thumbcache, 2019)

2.5 Prefetch Reader

Since the launch of Windows XP prefetch files have been used to speed up program execution. However, when viewed in their native form they are not in a human readable format. For this investigation an open source prefetch viewer was obtained enabling prefetch file format to be read in a standard text editor (NirSoft, 2019)

3. Digital Wallets, Bitcoin and obtaining the data

A digital wallet allows a user to store and commit transactions and is used as an interface to your cryptocurrency. Creating a digital wallet can be done anonymously and does not need 'linking' to any financial institutions. A digital wallet allows purchasing of goods and services anywhere in the world anonymously by using your digital wallets address. (Digital wallet, 2020)

After downloading a digital wallet's executable program, it needs be installed on the device being used. Part of this process involves producing a unique digital key, or more commonly known as a wallet's address. To construct this address most wallet providers; certainly, the ones used in this research; use what is known as a seed which is then encrypted to produce a unique identifier. A seed file is a random generated collection of words or a phrase that is used to create a private key.

3.1 Bitcoin

'A new type of cryptography based on blind signatures'. The initial introduction of this cryptography was to deny third parties' access to the payees 'lifestyle' (Chaum 1982). With access to payment card information Chaum deemed that you could build quite a complex picture of an individual, for example, where they visit on holiday, what films they like to watch, what they like to eat and even any illness they have.

Pseudonym Nakamoto (Nakamoto, 2008) took this a stage further with the introduction of a 'Peer-to-Peer Electronic Cash System', mainly for use for electronic commerce over the Internet and in 2009 Bitcoin was born.

3.2 Data extraction

For this investigation a systemic approach was required. As previously discussed not all digital wallets are coded to the same specification. A high-end out of the box vanilla laptop was procured as a baseline for the experiments. A sequence of downloads, transactions and removal of various Bitcoin digital wallets were then carried out. The steps are as follows:

1. Download first Digital Wallet
2. Receive and remove Bitcoin
3. Image the disk
4. Uninstall first Digital Wallet
5. Image the disk
6. Download two Digital Wallets, three Windows 10 'App Store' Wallets.
7. Buy Bitcoin on open market
8. Send Bitcoin to Digital Wallets
9. Image the disk
10. Uninstall/remove all Digital Wallets
11. Image the disk
12. Reset laptop to factory setting and Image the disk

4. Analysis

This section presents a step-by-step analysis methodology to investigating possible criminal activity mentioned in Section 1, by analysing data obtained in section 3.2, therefore allowing investigators the ability to link user activities on a Windows 10 laptop to such criminal behaviour.

The forensic analysis of the captured images discussed in section 3 are dealt with here, together with keyword searches for any unique identifier such as a digital wallet's address. This analysis is conducted with a variety of forensic tools and techniques using specialist open source readers and tools discussed in section 2.

4.1 Unique Identifier, Wallet Address

The unique identifier to all digital wallets is its digital address, and this is normally formed by a 'seed' of random words which are used to create a private key. This private key is then used with a one-way Elliptic Curve Multiplication to produce a public key. This public key then uses a one-way hashing algorithm to produce a unique digital address, always starting with a '1' (Antonopoulos, 2015).

For the next part of this analysis three of the digital wallets were selected and their corresponding digital addresses will be used. They are:

- Armory 1BKqcYKihbMvuw2RTTmka3tS1rvCPqWdkV
- Bitpay 1Gf8dtaiGXQs5ScCRS7aeGSkJLjD2W8Hb3
- Electrum 1M4yT8Ai6f6bn6ZAUPsTn4WHbx58nLtCyN

Loading the image file after uninstalling removal of all Digital Wallets into Autopsy application for analysis. A keyword search was performed on each of the digital wallet addresses. This should identify if the wallets address is held in plain text. This keyword search returned several results. Firstly, the installation of the digital wallet application 'Bitpay' which is located in Figure 1 below:

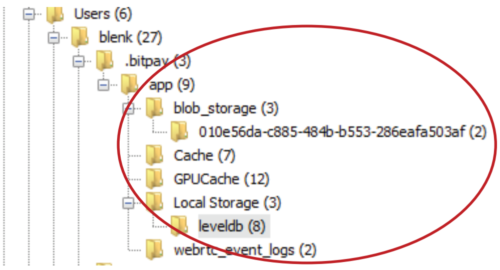


Figure 1: A Digital Wallet Application even after uninstallation of the program.

The digital wallet address is also held unencrypted in a log file after the program has been uninstalled. An extract from the log file is shown below:

[{"id":"5cd4173c12025b0a3946733f","txid":"146cdc577f3110bcf499f3f8eaa0216762da06819abea3c6f487452566158eb9","confirmations":0,"blockheight":null,"fees":12026,"time":1557403452.277,"size":116,"amount":10000,"action":"received","outputs":[{"address":"1Gf8dtaiGXQs5ScCRS7aeGSkJLjD2W8Hb3","amount":10000,"message":null}],"dust":false,"message":null,"creatorName":"","hasUnconfirmedInputs":

The second result was found from an experimental email that was sent replicating the threatening email mentioned earlier. Email data can be found in either one of the Unistore folders located here:

user/<user_name>/AppData/Local/Comms

and the data is held within a .dat file. An extract of the data is shown below:

--></style></head><body lang=EN-GB link=blue vlink="#954F72"><div class=WordSection1><p class=MsoNormal>My wallet address is 1Gf8dtaiGXQs5ScCRS7aeGSkJLjD2W8Hb3</p><p class=MsoNormal><o:p>&nbsp;</o:p></p><p class=MsoNormal>Sent from <a

4.2 Digital artefacts and Windows taxonomy

There was no evidence of digital artefacts on either the Vanilla baseline laptop before experiments commenced or once the laptop had been reformatted. As per the bullet point list set out in section 1 this research will now investigate Windows taxonomy at the various different stages of the image captured, where digital wallets have been downloaded and installed, transactions performed and then uninstalled.

4.3 User artefacts

4.3.1 Thumbcache.db / Iconcache.db

Before getting into lower level details of log files and the like, one of the quickest ways to identify digital wallets is by their logo, as the saying goes '*a picture speaks a thousand words'.* Cryptocurrency logos are quite unique as with most company logos and can be found with a standard internet search 'Bitcoin wallet logo'.

Internet searches leave a digital footprint including image artefacts. However, these artefacts are always stored in their own location as part of the browser history, the locations of three popular browser history files are:

InternetExplorer: \Users\<*Username*>\AppData\Local\Microsoft\Windows\History

GoogleChrome:\Users\<*username*>\AppData\Local\Google\Chrome\User Data\Default

MicrosoftEdge: Users\<*username*>\AppData\Local\Packages\Microsoft.MicrosoftEdge_8wekyb3d8bbwe\AC

IconCache is a database of stored icons linked to executable programs, therefore web search data could not be stored in this database by viewing internet pages. As discussed in section 2.4 Thumbcache.db or Iconcache.db is a visual representation of icon associated with starting applications on Windows 10 and are stored here:

Users\<username>\AppData\local\Microsoft\Windows\Explorer

After the first wallet 'Electrum' in the experiments was installed and transactions performed an Icon was identified using a Thumbcache viewer as seen in Figure 2 below:

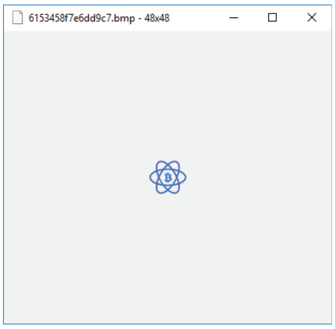
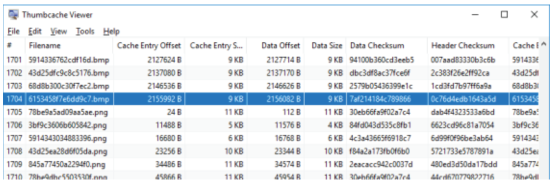

6153458f7e6dd9c7.bmp - 48x48

Figure 2: Electrum's logo

Once Electrum had been uninstalled and a new image taken the same icon was found and conatined the same filename name, as seen in Figure 3:


Thumbcache Viewer

#	Filename	Cache Entry Offset	Cache Entry S...	Data Offset	Data Size	Data Checksum	Header Checksum	Cache E
1701	5914336762cdf16d.bmp	2127624 B	9 KB	2127714 B	9 KB	94100b360cd3eeb5	007aad83330b3c6b	591433
1702	43d25dfc9c8c5176.bmp	2137080 B	9 KB	2137170 B	9 KB	dbc3df8ac37fce6f	2c383f26e2ff92ca	43d25d
1703	68d8b300c30f7ec2.bmp	2146536 B	9 KB	2146626 B	9 KB	2579b05436399e1c	1cd3fd7b97ff6a9a	68d8b3
1704	6153458f7e6dd9c7.bmp	2155992 B	9 KB	2156082 B	9 KB	7af214184c789866	0c76d4edb1643a5d	615345
1705	78be9a5ad09aa5ae.png	24 B	11 KB	112 B	11 KB	30eb66fa9f02a7c4	dab4f4323533a6bd	78be9a
1706	3bf9c3606b605842.png	11488 B	5 KB	11576 B	4 KB	84fd043d535c8fb1	6623cd96c81a7054	3bf9c3
1707	5914343034883396.png	16680 B	6 KB	16768 B	6 KB	4c3a43665f6918c7	6d99f0f96be3ab64	591434
1708	43d25ea28d6f05da.png	23256 B	10 KB	23344 B	10 KB	f84a2a173fb0f6b0	5721733e5787891a	43d25e
1709	845a77450a2294f0.png	34486 B	11 KB	34574 B	11 KB	2eacacc942c0037d	480ed3d50da17bdd	845a77
1710	78be9dbc5503530f.png	45866 B	11 KB	45954 B	11 KB	30eb66fa9f02a7c4	44cd670779822716	78be9d

Figure 3: Same Electrum filename and icon

Two further online digital wallets, Armory and Coinomi, were download and installed in addition to three wallets that were installed from the Windows App store directly, BitPay, Bither and CoinTradr. Icons were identified in the Iconcache.db. After purchasing Bitcoin and carrying out transactions between the wallets and then uninstalling all digital wallets, icons of all digital wallet were still identified, seen in Figure(s) 4.

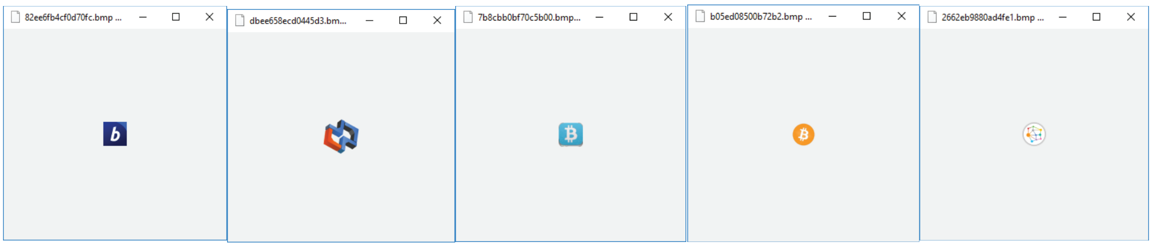


Figure 4: Five Digital wallet icons after uninstalling the applications

4.3.2 Jump List

Having identified digital wallet logos in the previous section, the next step is to take a deeper dive into the directory structure, looking at other areas in the Windows 10 taxonomy. Jump List files are a record of recently opened files and the directory locations can be found at:

```
<username>\AppData\Roaming\Microsoft\Windows\Recent\AutomaticDestinations
<username>\AppData\Roaming\Microsoft\Windows\Recent\CustomDestinations
```

Artefacts were found where digital wallets had been downloaded and installed. However, not every digital wallet was found in the Jump List. It is thought that this is because not all wallets had transactions applied and the application only installed and shutdown.

Shown in the below log extract of the Jump List file the same data resides for the digital wallet Bitpay after all the digital wallets had been uninstalled.

```
ms-windows-store://pdp/?productId=9NBR15SK4ZJV&ocid=&cid=&referrer=bitpay.com&scenario=click&webig=7886484d-be69-43de-
```

Also found in a jump list log file was a verification email from the Bitcoin trading house where Bitcoin was purchased in order to carry out these experiments. Extract shown below.

```
mailto:verification@coinfloor.co.uk,AppXydk58wgm44se4b399557yyyj1w7mbmvd1SPSU(L
```

4.3.3 Shortcut

Shortcut files contain useful forensic artefacts in the form of data objects, which are used to link to other data objects when launching applications. This process is well documented by Microsoft and is commonly known as Object Linking and Embedding (OLE). The directory location for Shortcut can be found here:

```
<username>\AppData\Roaming\Microsoft\Windows\Recent
```

The same data resides in memory when digital wallets have been downloaded, installed and then have been uninstalled, and can be seen in the following log file extract. **NB:** not all digital wallet artefacts were found.

```
Blade Pro SSDC:\Users\blenk\AppData\Roaming\Armory
```

4.4 System Artefacts

4.4.1 SRUDB.dat

System Resource Usage Monitor (SRUM) utilises backend databases, storing information about system resources. These entries can be investigated and analysed to identify which programs have been run and for what time period. One of these storage files is called SRUDB.dat. SRUDB.dat is a database based on Microsoft's ESE storage structure and the location path can be found at:

```
Windows\System32\sru\SRUDB.dat
```

Analysis of SRUDB.dat utilising FTK yields several references to 'Electrum' after uninstallation. An extract shown below shows where it can be seen that electrum appears to have been uninstalled.

```
.P.r.o.g.r.a.m. .F.i.l.e.s. .(.x.8.6.).\.E.l.e.c.t.r.u.m.\.U.n.i.n.s.t.a.l.l...e.x.e.
```

Certain log files in the same location yielded results (shown below) identifying the execution of some, but not all, digital wallets.

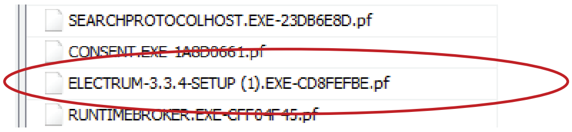
\device\harddiskvolume5\program files\coinomi\wallet\coinomi.exe

4.4.2 Prefetch Files

Prefetch files are used to speed up the execution of Windows programs. They have a file extension of .pf and contain metadata. Windows 10 Prefetch file location can be found here

C:\Windows\Prefetch

As discussed in section 2.5 a Prefetch viewer is used to interpret the data into a human readable format. Figure 5 below shows the first digital wallet Electrum setup that was run. This screenshot was taken after the digital wallet was installed.



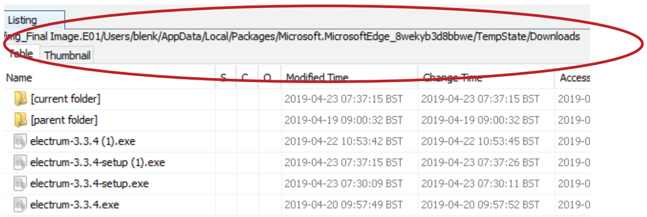
SEARCHPROTOCOLHOST.EXE-23DB6E8D.pf
CONSENT.EXE-1A8D0661.pf
ELECTRUM-3.3.4-SETUP (1).EXE-CD8FEFBE.pf
RUNTIMEBROKER.EXE-CFF04F45.pf

Figure 5: Electrum setup file in Windows 10 Prefetch directory location

Once the wallet was uninstalled the same entry in the Prefetch file location was found. This was exported from Autopsy and parsed though the Prefetch viewer. The analysis from the Prefetch viewer gave the directory location of the .exe file itself which is identified below.

<username>\AppData\LOCAL\PACKAGES\MICROSOFT.MICROSOFTEDGE__8wekyb3d8bbwe\TEMPSTATE\DOWNLOADS\ELECTRUM-3.3.4-SETUP(1).EXE

When this location was traced back in the image the .exe programs for Electrum were identified, shown in Figure 6 below:



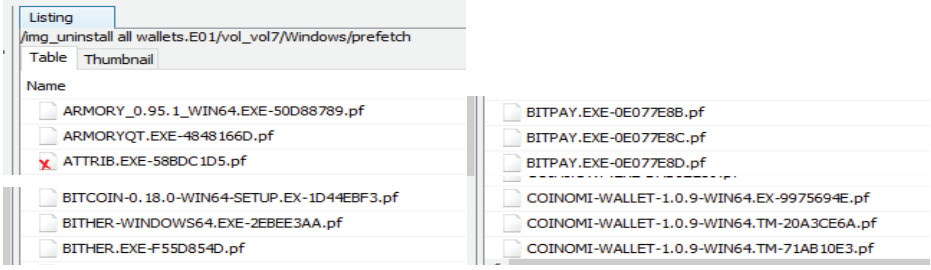
Listing
/img_Final Image.E01/Users/blenk/AppData/Local/Packages/Microsoft.MicrosoftEdge_8wekyb3d8bbwe/TempState/Downloads

Name	S	C	O	Modified Time	Change Time	Access
[current folder]				2019-04-23 07:37:15 BST	2019-04-23 07:37:15 BST	2019-0
[parent folder]				2019-04-19 09:00:32 BST	2019-04-19 09:00:32 BST	2019-0
electrum-3.3.4 (1).exe				2019-04-22 10:53:42 BST	2019-04-22 10:53:45 BST	2019-0
electrum-3.3.4-setup (1).exe				2019-04-23 07:37:15 BST	2019-04-23 07:37:26 BST	2019-0
electrum-3.3.4-setup.exe				2019-04-23 07:30:09 BST	2019-04-23 07:30:11 BST	2019-0
electrum-3.3.4.exe				2019-04-20 09:57:49 BST	2019-04-20 09:57:52 BST	2019-0

Figure 6: file location path showing the .exe files

The same experiment was carried out using a selection of wallets previously mentioned. Three wallets were found to be contained in the Prefetch directory - Armory, Bitpay and Coinomi.

The Prefetch file contained more information after uninstalling all the digital wallets than when they were installed, it is believed that the application started itself in order for it to be uninstalled. Figure 7 shows four digital wallets located in the Prefetch file - Armory, Bitpay, Bither and Coinomi.



Listing
/img_uninstall all wallets.E01/vol_vol7/Windows/prefetch

Name
ARMORY_0.95.1_WIN64.EXE-50D88789.pf
ARMORYQT.EXE-4848166D.pf
ATTRIB.EXE-58BDC1D5.pf
BITCOIN-0.18.0-WIN64-SETUP.EX-1D44EBF3.pf
BITHER-WINDOWS64.EXE-2EBEE3AA.pf
BITHER.EXE-F55D854D.pf
BITPAY.EXE-0E077E8B.pf
BITPAY.EXE-0E077E8C.pf
BITPAY.EXE-0E077E8D.pf
COINOMI-WALLET-1.0.9-WIN64.EX-9975694E.pf
COINOMI-WALLET-1.0.9-WIN64.TM-20A3CE6A.pf
COINOMI-WALLET-1.0.9-WIN64.TM-71AB10E3.pf

Figure 7: Prefetch file location with more digital wallets after uninstallation

4.4.3 Amcache.hiv

Amcache.hiv stores information on recently executed programs and their associated history. In Windows 10 the Amcache.hiv files can be found here:
Windows\appcompat\Programs

Analysis of Amcache.hve yields several references to 'Electrum' after uninstallation extract below, where it can be seen that electrum.exe appears to have been run.

```
\.u.s.e.r.s.\.b.l.e.n.k.\.a.p.p.d.a.t.a.\.l.o.c.a.l.\.p.a.c.k.a.g.e.s.\.m.i.c.r.o.s.o.f.t...m.i.c.r.o.s.o.f.t.e.d.g.e._.8.w.e.k.y.b.3.d.8.b.b.w.e.\.t.e.m.p.s.t
.a.t.e.\.d.o.w.n.l.o.a.d.s.\.e.l.e.c.t.r.u.m.-.3...3...4.-.s.e.t.u.p...e.x.e
```

Certain log files in the same location yielded results identifying the uninstallation of 'Electrum'.

```
c:\users\blenk\appdata\local\packages\microsoft.microsoftedge_8wekyb3d8bbwe\tempstate\downloads\electrum-3.3.4-setup.exe
HKEY_USERS\S-1-5-21-170591366-2771221896-939485899-1001\Software\Microsoft\Windows\CurrentVersion\Uninstall\Electrum
```

After the installation of a selection of digital wallets artefacts relating to all but one of the digital wallets was found either in Amache.hve or associated log files, as shown below.

```
FileCreate=C:\Program Files (x86)\Armory\ArmoryDB.exe
FileCreate=C:\Program Files (x86)\Armory\MSVCP90.dll
```

Once all the wallets have been removed and uninstalled the same log file as discussed above was identified along with a reference to Coinomi wallet, shown below.

```
FileCreate=C:\Program Files\Coinomi\Wallet\unins000.exe
FileCreate=C:\Program Files\Coinomi\Wallet\jre\bin\attach.dll
```

4.5 Registry Artefacts

Registry artefacts are contained in a hierarchical central database in the Windows Registry. This Registry contains a substantial amount of information which may be of significant importance in a forensic examination of a computer system.

Windows 10 registry files are located in the paths below. Exported from a forensic analysis tool they can be read with a RegRipper application.

1. Sam – C:\Windows\System32\Config\SAM
2. Security C:\Windows\System32\Config\SECURITY
3. Software C:\Windows\System32\Config\SOFTWARE
4. System C:\Windows\System32\Config\SYSTEM
5. Default C:\Windows\System32\Config\DEFAULT
6. NTUSER.dat C:\Users\<username>\NTUSER.DAT
7. UsrClass.dat C:\Users\<username>\AppData\local\Microsoft\Windows\UsrClass.dat

Nothing of interest is found in SAM, SECURITY or DEFAULT registry files. However, the rest of the registry files could contain digital evidence artefacts relating to digital wallets.

4.5.1 Software

The SOFTWARE file contains over fifty-six thousand lines of text, far too much for a general search, as this would be prone to human error. A comparison tool mentioned in section 2.6 is employed. These experiments will compare one image with the previous one in order to distinguish changes to the file.

Nothing of significance was found between the first image containing a single wallet and when the wallet had been uninstalled. However, when a comparison was carried out between the third and fourth image, one containing a selection of wallets and the last uninstallation of all wallets, the comparison tool highlighted a notable difference. An extract of the difference is shown below.

```
1557414229|REG|||[Uninstall] - Coinomi Wallet version 1.0.9 v.1.0.9
1557223731|REG|||[Uninstall] - Bitcoin Armory v.0.95.1.0
```

When the image after ‘Electrum’ was uninstalled it was then analysed using FTK. Artefacts were identified and extracted below to make it easier to understand:

```
S.o.f.t.w.a.r.e.\.M.i.c.r.o.s.o.f.t.\.W.i.n.d.o.w.s.\.C.u.r.r.e.n.t.V.e.r.s.i.o.n.\.U.n.i.n.s.t.a.l.l...E.l.e.c.t.r.u.m...C.:.\.P.r.o.g.r.a.m.
```

Similar artefacts were identified after the uninstallation of all the wallets and analysed using FTK, as described in the extract below:

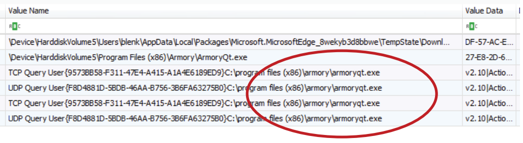
```
\.P.r.o.g.r.a.m. .F.i.l.e.s.\.W.i.n.d.o.w.s.A.p.p.s.\.T.i.m.o.P.a.r.t.l...C.o.i.n.i._.1...3...4...1.0.0.0._.x.6.4
```

4.5.2 System

The SYSTEM file contains over ten thousand lines of text. As previously mentioned both the first wallet installed image and first wallet uninstalled image where loaded into the comparison tool. Although there were differences between the two files most appeared to be system related. However, a word search was applied looking for references to the first digital wallet Electrum. Both installed image and uninstalled images contained the same references to Electrum extract below.

```
1555750669|REG|||M...                    AppCompatCache                    -
C:\Users\blenk\AppData\Local\Packages\Microsoft.MicrosoftEdge_8wekyb3d8bbwe\TempState\Downloads\electrum-3.3.4.exe
```

Although the RegRipper identified nothing of interest between the selection of wallets imaged and the uninstallation image, both Registry Explorer and FTK yielded the same results. Armory, Coini, bitpay and BitcoinTradr were all found. Figure 9 below shows the result of a search on Armory in Registry Explorer.



Value Name	Value Data
\Device\HarddiskVolume5\Users\blenk\AppData\Local\Packages\Microsoft.MicrosoftEdge_8wekyb3d8bbwe\TempState\Downl...	DF-57-AC-E...
\Device\HarddiskVolume5\Program Files (x86)\Armory\ArmoryQt.exe	27-E8-2D-6...
TCP Query User{9573BB58-F311-47E4-A415-A1A4E6189ED9}C:\program files (x86)\armory\armoryqt.exe	v2.10|Actio...
UDP Query User{F8D4881D-5BDB-46AA-B756-3B6FA63275B0}C:\program files (x86)\armory\armoryqt.exe	v2.10|Actio...
TCP Query User{9573BB58-F311-47E4-A415-A1A4E6189ED9}C:\program files (x86)\armory\armoryqt.exe	v2.10|Actio...
UDP Query User{F8D4881D-5BDB-46AA-B756-3B6FA63275B0}C:\program files (x86)\armory\armoryqt.exe	v2.10|Actio...

Figure 9: Result of a search on Armory in Registry Explorer

4.5.3 NTUSER.dat

NTUSER.dat file in the first two images where the first wallet was installed and uninstalled, contains the same entries shown in the extract below.

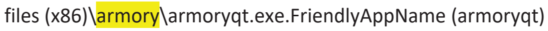
```
Tue Apr 23 06:29:11 2019                    8 = electrum-master.zip

1556000951|REG|||RecentDocs - Software\Microsoft\Windows\CurrentVersion\Explorer\RecentDocs\.zip - electrum-master.zip
```

The same file has been extracted for the image that had all the digital wallets which were uninstalled, and interestingly more entries were identified in this file than the image where the selection of wallets where installed, as shown below.

```
C:\Users\blenk\AppData\Roaming\JWrapper-Bither\JWrapper-Windows64JRE-00054022688-complete\bin\Bither.exe

Thu May  9 16:26:53 2019 Z
{A77F5D77-2E2B-44C3-A6A2-ABA601054A51}\Bither\Uninstall Bither.lnk (3)
Thu May  9 16:16:17 2019 Z
C:\Users\Public\Desktop\Coinomi Wallet.lnk (1)
Thu May  9 14:48:56 2019 Z
C:\Users\Public\Desktop\Bitcoin Armory.lnk (7)

UserAssist - 51239JoeLevy.BitcoinTradr_2rba8c4v97x22!App (1)
1557223633|REG|||[Program                    Execution]                    UserAssist                    -
C:\Users\blenk\AppData\Local\Packages\Microsoft.MicrosoftEdge_8wekyb3d8bbwe\TempState\Downloads\armory_0.95.1_w
in64.exe (1)
1557419213|REG|||[Program Execution] UserAssist - {A77F5D77-2E2B-44C3-A6A2-ABA601054A51}\Bither\Uninstall Bither.lnk
(3)
1557418577|REG|||[Program Execution] UserAssist - C:\Users\Public\Desktop\Coinomi Wallet.lnk (1)
```

4.5.4 UsrClass.dat

UsrClass.dat file location which was exported from Autopsy so it could be passed though the RegRipper application and analysed using a standard notepad text editor. Windows 10 UsrClass.dat file also contained references to several digital wallets even after they have all been uninstalled on Windows 10, as descibed below:



```
files (x86)\armory\armoryqt.exe.FriendlyAppName (armoryqt)
```

5. Conclusions

This paper presented a forensic analysis of the Windows 10 operating system and the evidence (digital) artefacts remaining once digital wallet applications had been uninstalled. Together with identifying emails containing a digital wallet address, it was established that when the digital wallet applications Electrum, Armory, Coinomi, Bitpay, Bither, BitTradr and Coini have been downloaded installed/uninstalled; and experimental emails containing a digital wallet address; do leave digital artefacts. In fact, an abundance of evidence (digital) artefacts was traced through the analysis of Windows 10 forensically imaged hard drive(s), taken at various different stages through a progression of simulated experiments. This research has established that artefacts of significant forensic interest can be located throughout various locations within Windows 10. Digital wallet addresses were identified in plain text stored within log files in various applications themselves.

References

Alex Heid, 2013. Analysis of the Cryptocurrency Marketplace [online] http://www.hackmiami.org/whitepapers/HackMiami-Analysis_of_the_Cryptocurrency_Marketplace.pdf [accessed 8/4/2019].

Andreas M Antonopoulos, (2015), Mastering Bitcoin, O'Reilly Media Inc.

Bhupendra Singh, Upasna Singh, (2018), Program execution analysis in Windows: A study of data sources, their format and comparison of forensic capability. Elsevier.

Chaum D, (1982), White Paper: Blind Signatures for Untraceable Payments.

https://accessdata.com/products-services/forensic-toolkit-ftk (on-line accessed 27/7/2019).

https://www.sleuthkit.org/autopsy/features.php (on-line accessed 27/7/2019).

https://thumbcacheviewer.github.io/ (on-line accessed 25/7/2019).

http://www.forensicsware.com/blog/e01-file-format.html(on-line accessed 27/7/2019).

https://www.forensicswiki.org/wiki/Regripper (on-line accessed 26/7/2019).

https://www.nirsoft.net/utils/win_prefetch_view.html (on-line accessed 25/7/2019).

Kristensen, R., 2020. A Guide To Regripper And The Art Of Timeline Building. [online] Forensic Focus - Articles. Available at: <https://articles.forensicfocus.com/2014/09/25/a-guide-to-regripper-and-the-art-of-timeline-building/> (on-line accessed 27/7/2019).

PCMAG. 2020. Definition Of Digital Wallet. [online] Available at: <https://www.pcmag.com/encyclopedia/term/digital-wallet> (on-line accessed 27/7/2019).

Nakamoto S, (2008), Bitcoin: A Peer-to-Peer Electronic Cash System.

Information Sharing in Cyber Defence Exercises

Agnė Brilingaitė[1], Linas Bukauskas[1], Aušrius Juozapavičius[2] and Eduardas Kutka[1]
[1]Vilnius University, Lithuania
[2]General Jonas Žemaitis Military Academy of Lithuania, Vilnius, Lithuania
agne.brilingaite@mif.vu.lt
linas.bukauskas@mif.vu.lt
ausrius.juozapavicius@lka.lt
eduardas.kutka@mif.vu.lt

DOI: 10.34190/EWS.20.034

Abstract: Availability and easy access to sophisticated cyber penetration testing tools enable exploitation of vulnerabilities in different systems globally. Repetitive nature and recognisable signatures of attacks raise demand for effective information sharing. Timely warnings about cyber incidents in other systems make it possible to identify related attacks locally. International cyber community supports several commercial and open-source threat information sharing platforms. Efficient use of these systems depends both on the quality of submitted information and the ability of the security specialist to receive, interpret, and integrate indicators of compromise into local defence systems. Business stakeholders tend to emphasise the importance of threat hunters, while the information-sharing aspect is overlooked. Therefore, there is a need for professionals who can assess risk levels of cyber incidents in a broad context and share concise information with team members, superiors, relevant institutions, and community. The complex nature of cyber attacks raised the popularity of live cyber defence exercises (CDX), where cybersecurity specialists are trained using simulated real-life scenarios. However, the exercises are mostly oriented towards the development of technical competences. This paper addresses the problem of proper development of information sharing competence during the CDX. We performed a case study of two annual international CDX. Research data were collected using several techniques. First, the participants filled in pre-event and post-event questionnaires. Additionally, each defending team was continuously observed by a dedicated evaluation team member. Finally, incident reports in short and long forms were gathered. We distinguished challenges related to internal team collaboration, information sharing among the teams, and reporting to relevant authorities. Based on the findings, we present a methodology to integrate information sharing into the planning and execution of CDX. The methodology encompasses activities, scoring strategies, scenario recommendations, tools, and communication-encouragement components. The presented enhancement creates an observable added value to the CDX training event.

Keywords: cyber defence exercises, incident information sharing, indicators of compromise, collaborative defence

1. Introduction

Criminal cyber activities tend to be global and do not respect the borders of actual jurisdiction. The global reach of the internet and a variety of cyber threats from outside of organisations make intelligence information on possible attacks critical. The solution is to share commonly known indicators of compromise (IoCs) among organisations so that defensive infrastructures learn from each other.

Information sharing is beneficial to any organisation — from governmental institutions (Scott, 2009) to businesses (Cui et al., 2015). Usually, sharing is considered as a self-explanatory action. The lack of a clear definition of what exactly is the information and the sharing process leads to the breakdown of the process (Beynon-Davies and Wang, 2019). Both the sender and receiver of information need to have the same definitions. On the one hand, sharing of cyber threat intelligence information is a standardised matter — there is a finite number of clear IoCs one may use to describe a cyber attack. On the other hand, many organisations are creating different classification schemes of the incidents (such as ATT&CK by MITRE Corporation (2019b) or the famous Kill Chain by Lockheed Martin). There are also many different platforms with their unique and loosely defined parameters a user has to understand to describe or read the data correctly. As a consequence, specialised training is needed for cybersecurity specialists and threat hunters, in particular, to successfully use and produce threat intelligence information.

The main contribution of the paper is the analysis and discussion of possible adjustments to the execution of cybersecurity exercises. The purpose of the adjustments is to improve information sharing competence of the exercise participants — cybersecurity specialists who should be responsible for active collaboration with their peers in other organisations. The findings are based on the analysis of live exercises. We present a methodology to integrate information sharing into the planning and execution of CDX. The methodology enriches Identify, Plan, and Execute phases of the event with activities, scenario recommendations, and examples of tools.

The paper is structured as follows. Section 2 describes the importance and challenges of information sharing in the area of threat intelligence. A case study is presented in Section 3. Section 4 covers discussion and suggested improvements for future CDX. The paper ends up with conclusions and future work.

2. Related work

Threat intelligence includes valuable information from various sources and can benefit security staff using the right tools (Pokorny, 2019). Intelligence is a product of tedious processing, analysis, and interpretation of data that should be understood and gathered correctly. Feeds of threat data coming from reliable external sources provide an advantageous input for the threat intelligence.

Vulnerability databases provide one source of data commonly used by cybersecurity specialists worldwide, e.g., Common Weaknesses Enumeration (MITRE Corporation, 2019a) and National Vulnerability Database (NIST, 2019). Such sources focus on the technical properties of exploits (Pokorny, 2019) and provide threat risk assessment. The vulnerabilities may be abused in many different ways, and MITRE Corporation developed its ATT&CK Framework (MITRE Corporation, 2019b) to classify attack methods. ATT&CK is a knowledge base categorising and describing known adversary tactics and techniques based on observed cyber attacks. Additionally, several standards were created to facilitate the exchange of intelligence information between different organisations, e.g., the STIX format for presenting the information (Barnum, 2012) and the TAXII protocol for sharing it (Connolly et al., 2014).

The cybersecurity community maintains several proprietary systems to record cyber incidents. To address the standardisation issues, the European Union co-financed an open-source Malware Information Sharing Platform (MISP) (MISP Project, 2019). Wagner et al. (2016) present a possible case of MISP usage and implementation. They describe important features like sharing levels, multi-organisational MISP support and integration, and customisation of taxonomies for different organisations. The system enables usage of pull, push, and cherry-picking methods for event distribution among organisations. Also, to break down the resistance to information sharing among the organisations, the Arizona Cyber Threat Response Alliance was established (Haass et al., 2015).

Despite the attempts at standardisation, there exist several formats for threat information sharing. The variety of formats prevents an efficient spread and use of information as there is no standard definition for the components of the different formats. Menges et al. (2019) try to overcome the problem by creating a unified meta-model that describes the essential properties and conventional notation for entities of various threat information sharing formats. Apart from the format problem, a threat intelligence specialist may encounter other challenges when working with multiple intelligence sources, combining and enriching data for fine-grained intelligence, or determining intelligence relevance based on technical constructs (Brown et al., 2015).

Brilingaitė et al. (2020) emphasise the lack of cybersecurity specialists in the world and encourage educating them through various methods. Based on their experience with planning, executing and monitoring cyber exercises, the authors propose an extended competence development and assessment framework, which could be used to enhance participant experience and improve learning outcomes of CDX.

Cyber-attacks have global nature, and therefore information sharing must cross country boundaries as well (ENISA, 2013, 2016). However, it has many obstacles related to legal issues, technical and procedural problems, lack of trust, and insufficient interest of partners. Johnson et al. (2016) emphasise that one of the benefits of information sharing is shared situational awareness. At the same time, organisations have to deal with various challenges, e.g., establishing trust, enabling information consumption, safeguarding sensitive information, and evaluating the quality of received information.

Ring (2014) considers the importance of information sharing within the area of the threat intelligence community and addresses its problems. Interestingly, information sharing among teams of the same organisation is even a bigger problem than sharing among several organisations, because it requires feedback from another team. On a higher level, sharing between organisations is also challenging. Although there is much information available both in paid and free intelligence systems, the systems themselves have typical problems like a limited number of queries, expiring passwords, or outdated information. Also, cyber-attacks become personalised, and organisations tend to focus on specific threat analysis instead of spreading information. As a consequence, the number of sharers is small, and many organisations only consume threat indicators, even though an organisation may benefit substantially by producing threat information (Johnson et al. 2016).

Sander and Hailpern (2015) discuss the importance of user experience (UX) in the Threat Information Sharing Platform (TISP). According to them, the sharing of rich contextual data is more valuable than just automatically acquired and aggregated data. Profiling of typical TISP users indicates that most of them are 23–34 years old. Therefore, TISP interface and functions have to exhibit a modern look and feel to be easily understood by people of that age. As a side effect, this view allows TISP designers to identify UX problems and present future requirements to improve the system.

The requirement of anonymity as an encouragement for information sharing is raised by Murdoch and Leaver (2015). On the other hand, security specialists may want to be acknowledged and recognised for their contributions to the community. Therefore, so-called social-network-like profiles are suggested in TISP systems as a gamification feature. Every reporter would have an anonymous profile while retaining the ability to gather credibility and reputation points.

Privacy issues should be addressed when sharing information to avoid negative consequences, e.g., legal problems, reputation damage, and disclosure of infrastructure details. Thus, van de Kamp et al. (2016) present a proof of concept of a secure sharing system using private and encrypted reporting of sightings. In that case, only trusted subscribers can use shared data. Almost two years ago, the European Commission issued the General Data Protection Regulation, GDPR, (Council of European Union, 2016). The regulation also influences the information sharing process, as cyber threat intelligence data-sets might contain personal data. Albakri et al. (2019) suggest using the Traffic Light Protocol (TLP) to indicate different levels of sensibility for distributed data. They present a cyber threat intelligence policy space and a decision graph to identify the access level of the particular shared data using TLP. The strategy of different sharing levels is implemented in MISP (MISP Project, 2019) as described by Wagner et al. (2016). Similarly, Murdoch and Leaver (2015) noticed that many organisations restrict information sharing by defining what information can be shared only internally within a team, only with partners, or with everybody externally.

Trust is an essential feature in the threat information sharing platforms, as data-sets contain sensitive information. Wagner et al. (2018) present a trust taxonomy to enable a trusted sharing community. Burger et al. (2014) proposed a layered taxonomy to assess and analyse different threat intelligence ontologies. The taxonomy follows the ISO OSI protocol model.

The related work stresses the importance of cyber threat information sharing and some possible technological and organisational solutions to overcome its challenges. As another piece of the puzzle, we address the educational part of information sharers in this article.

3. Case study

We performed a case study of two international cybersecurity exercises in autumn of 2018 and 2019. The primary training audience were cybersecurity specialists of critical infrastructure enterprises. They were grouped into the Blue teams as defenders of a simulated infrastructure. There were four and eight Blue teams in 2018 and 2019, respectively, with up to 8 persons per team. Each team had an identical cyber range to defend against an ever-growing amount of various attacks performed by the Red team. We assigned one person to each team as a dedicated Observer. Also, we had full access to all team reports, service availability data, and the attack matrix.

The exercises were oriented towards cybersecurity specialists with little experience. Indeed, 83% of participants in 2018 and 60% in 2019 attended this type of event for the first or second time only. When asked to self-assess their cybersecurity competence level, 92% of the Blue team members stated they were on an intermediate level or lower in 2018. Therefore, in 2019 organisers emphasised training over the competition, and special care was taken to create a stress-free environment. E.g., teams could see their score based on service availability, but there was no aggregated scoreboard to compare against other teams.

Apart from the direct observation of the Blue teams during the exercises, we also gathered their self-evaluation data using pre-event and post-event questionnaires to understand their attitude towards information sharing better. In 2018, 24 Blue teams members submitted their answers, and 42% identified soft skills as one of their strengths. During the exercises, 48% of the respondents used soft skills, and 32% said they would like to improve them further. Problems related to poor communication were often mentioned in the questionnaire. Observers also confirmed poor inter-team and almost non-existing cross-team information sharing in 2018. Most of the participants were satisfied with their overall experience giving 4.25 points on average out of 5, but they wanted more opportunities to exchange knowledge about the attacks with one another. In the feedback, the participants requested for some extra time for team-building before the exercises and additional social activities (including more coffee breaks) before and during the event — both to get to know team members better and to foster cross-team communication. The organisers accommodated the suggestions and allotted a whole full week of training before the live part of the exercises in 2019. Additionally, many tools were introduced to facilitate information exchange.

During the exercises in 2019, a particular emphasis was placed on training threat hunters and enhancing collaboration in all aspects. The participants were given a set of lectures about the use of the Security Onion threat hunting platform, the proper methodology of cyber incident triage, and the tools and forms needed to report the incidents they encounter during the live stage of the exercises. After detecting an incident, a Blue team had to submit a short initial report to an emulated management of a fictitious company the team was defending. The management could ask the team to submit a comprehensive report about the attack after the incident was closed. The report form was similar to the template provided by the US-CERT Incident Reporting System by following the NIST Computer Security Incident Handling Guide (Cichonski e al., 2012). There were no points given for incident reports, although the teams were encouraged to share incident information. Every day the Evaluation team gave feedback on the report quality, and teams with the best reports had to present them during hot wash-up meetings. The training objectives included learning to identify threats and share threat intelligence with other teams without application of fully-automated sharing platforms such as TAXII.

All forms of information exchange and cross-team communication were promoted in 2019. It was allowed to visit other teams, there were multiple coffee breaks, and each team had to present their main challenges, tools, defence methods, and lessons learned at the end of each day. Most importantly, they exchanged IoCs. Filters of the Security Onion systems had been pre-configured to integrate the MISP service. As cyber ranges and attacks were almost identical, each team could benefit from others if correct IoCs were submitted to MISP.

The teams had to dedicate at least one member to reporting and information exchange. In 2018, teams had to submit free-form incident reports using an email. In 2019, the information-sharing specialists had to use a substantial number of different platforms, presented in Figure 1. Specialists had to respond to the tickets created by simulated users using the Request Tracker system. Also, they had to submit incident reports to company managers using two forms on a Collaboration platform and enter IoCs in the MISP. Additional information was communicated by email. Interestingly, the teams were slightly wary of one another during the initial days. However, on the third day of live exercises in 2019, they had spontaneously decided to use an instant messaging platform. As one of the teams explained during the daily hot wash-up, the MISP platform was deemed too complicated, slow, and non-intuitive to facilitate fast information exchange.

The pre-event questionnaire of the 2019 CDX was filled in by 43 Blue team members, and 67% of them claimed their cybersecurity skills are on an intermediate level or below. Thus, the percentage of experienced participants increased substantially compared to the previous year. When asked about their strengths, the participants gave the largest score to their Windows OS skills (2.98 on a scale of 1 to 5), with soft skills (2.61) and reporting skills (2.33) not far behind. All other skills received a lower average value with forensics skills (1.63) at the bottom. Finally, 44 Blue team members submitted the post-event questionnaire. It revealed that 64% of them did use Windows OS skills during the exercises while only 20% of used forensics skills.

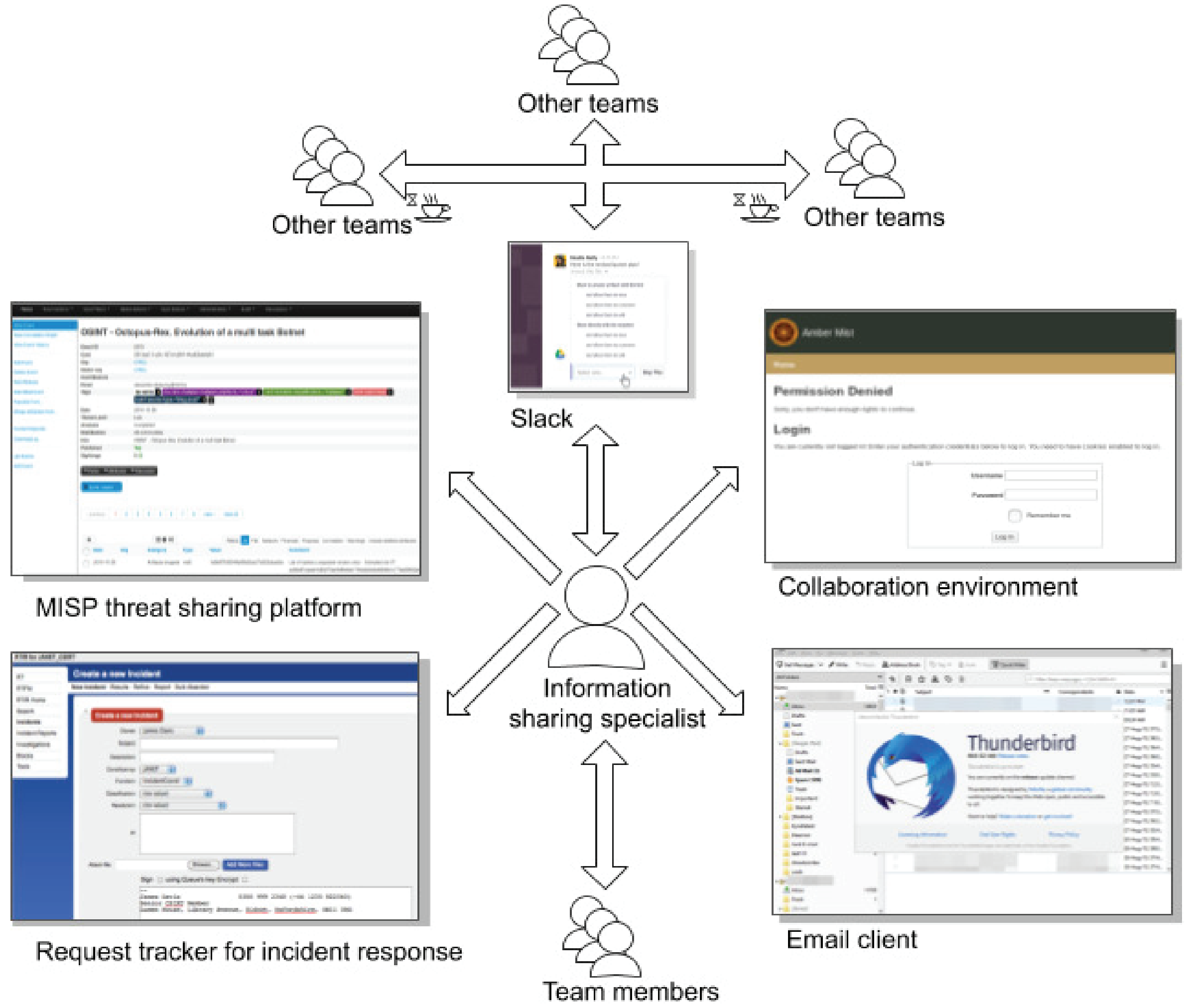

Other teams
Other teams
Other teams
Slack
MISP threat sharing platform
Collaboration environment
Information sharing specialist
Request tracker for incident response
Email client
Thunderbird
Team members

Figure 1: Communication channels of an information sharing specialist

During the training week in 2019, the participants were given lectures about a threat hunting system (Security Onion) and a threat intelligence sharing system (MISP). In the pre-event questionnaire, only 3 participants (7%) mentioned they had previous experience with Security Onion, and nobody mentioned MISP or any other threat intelligence sharing platform. In the post-event questionnaire, 61% of the respondents claimed they used Security Onion during the exercises. On the other end, only 25% of the participants said they used MISP, although 41% claimed they were involved in incident reporting activities. Additionally, 70% of the respondents specified they would like to learn more about threat hunting, 55% about Security Onion, 36% about MISP, and 34% about reporting. Consequently, despite a similar introduction of the two systems and a constant encouragement of the organisers to use MISP, the users clearly chose to focus their attention on the threat hunting platform, while reporting activities were given the lowest priority. This, in turn, led to poor quality of most of the incident reports as evaluated by the organisers.

4. Discussion

The results of the case study show that information sharing does not get enough attention during the CDX. Several reasons affect the attitude towards reporting and information sharing (RIS):

1. RIS tasks require a lot of technical skills to assess the risk of the attack, define IoCs, map attack in the knowledge base of adversarial techniques, and use threat intelligence information. For example, IoCs include not only technical data but also a likely impact, handling methods, phases of the Kill Chain, and other related features.
2. RIS tasks are carried out using information systems, e.g., some sharing platform developed or applied for the particular CDX event. Thus, participants should have experience in using the platform, or the platform should be very user-friendly. For example, publishing the attack pattern with meta-data in

MISP has several steps or states (stored or published). Therefore, knowledge about the workflow and special buttons of its graphical user interface is required.

3. RIS activities are treated as low-priority because technical cybersecurity specialists are the primary target audience of CDX. Thus, teams prioritise technical tasks (firewall settings, network monitoring, updates of user policy, restarting services) to deal with an ongoing attack instead of reporting it.
4. Participants do not associate RIS activities with a global context outside the CDX event. The activities seem to have only local importance and no use outside the game. Thus, participants do not feel motivated to develop RIS-related competences.

The listed factors create a situation when the assignment of an experienced technical person to RIS tasks is treated as a sacrifice from the perspective of the team. Therefore, during the CDX event, it is essential to demonstrate the strategic value of RIS. The gained RIS experience and competences would benefit the global cybersecurity community afterwards.

We present a methodology to enhance the CDX event with components encouraging communication and collaboration. A usual CDX event follows four phases: Identify (P1), Plan (P2), Conduct (P3), and Evaluate (P4). These standard phases overlap with competence development phases (Brilingaitė et al., 2020): Pre-exercise assessment (C1), Pre-exercise training (C2), LiveEx (C3), and Post-exercise assessment (C4). We propose several techniques to develop RIS related competences. Phase C1 starts in P1 and ends in P2, phase C2 continues through P2 and P3, while C3 is a part of P3. The key points in the methodology are:

- identify competences and objectives related to RIS,
- choose a representative attack as an example,
- train all participants using the example attack before the LiveEx, and
- emphasise the role of an individual in global security.

The following tasks should be carried out during the corresponding stages of CDX to integrate the development of RIS competences:

Stage P1-C1: Information sharing should be defined as one of the goals of the CDX, and learning objectives related to RIS should be specified. The objectives should address the ability to follow the procedure of sharing attack patterns using a sharing platform, to assign correct IoCs, to attribute the attack based on intelligence reports.

Stage P2-C1: Choose tools and standards to integrate into the CDX. For example, the organisers could (a) choose an open-source platform MISP, and (b) decide to apply MITRE ATT&CK framework. Choose a few representative cases/examples of attacks so that they can be reported using the selected tools. Introduce at least one case closely related to the CDX scenario. Preparation of the case includes: intelligence information, live attack demonstration, demonstration of the sharing of the attack pattern, and recommended learning material. For example, organisers choose the Bypass User Account Control and Access Token Manipulation in the Privilege Escalation category from the ATT&CK Matrix. Then, an attack demonstration can be prepared for these specific techniques. A scoring system should also include RIS activities. For example, each report could be given points from 0 to N. Each published pattern of an attack is treated as a bonus of value M. If a team fails to fill in a report or does not share an attack pattern during a time interval T, then penalties are calculated using a function P(T, M, N). A criteria matrix CM should be prepared for the evaluation of reports and a specific individual or group RIS-related competences.

Stage P2-C2: Distribute learning material for participants with relevant practical tasks and self-control challenges. Assess the level of the participants (ideally, use some remote learning platform for this) and update the scoring system based on the analysis to credit more points in the areas where more development is needed.

Stage P3-C2: Introduce participants on-site to the prepared cyber range and run the test cases of the attacks. Tasks should be given to participants in practical sessions to get hands-on experience. Then challenges could ensure the entry-level of competences before the LiveEx: the experience with sharing platforms, needed technical skills, and a common understanding of data flow among systems and participants.

Stage P3-C3: During the LiveEx of the CDX, use observers to assess the RIS-related competences of the participants. Some Evaluation team members should be dedicated to the evaluation of reports based on the criteria matrix CM.

Three aspects should be considered to stimulate information sharing behaviour of the participants: competence, opportunity, and motivation. Correctly defined CDX objectives and training plan with a competence matrix would ensure the development of RIS-related competences. The right toolset (e.g., MISP), scenario, and training material would create opportunities for immersive learning and demonstration of competences. The right scoring system and clear added value would foster the motivation of the participants to share the threat intelligence information. As an encouragement for a team to value work done by RIS members, one can add team intrinsic motivation factor by assigning significant score points for every successfully shared IoC and multiplied scores for implementing recommended solutions that were applied by other teams.

5. Conclusions and future work

Information sharing among organisations has always been a challenge, although the successful defence of infrastructures largely depends on external information. One of the factors inhibiting successful dissemination of relevant information is a limited competence of the information sharing specialist. Based on our findings during a CDX case study, we propose a methodology to stimulate the development of the information sharing competences of the CDX participants. The suggested minor changes in the CDX execution would, in turn, would be beneficial to the broader cybersecurity community.

Our work can be extended in several directions. Firstly, a detailed model of the CDX scoring system could be developed and tested. This model should be customisable depending on the chosen importance of RIS in the exercises. Another direction of the research might be an investigation of homogeneous environments that provide a multitude of combined services in a dashboard to simplify communication and ease the work of information sharing specialists.

Acknowledgements

The authors of the paper would like to express their gratitude to the organisers, participants, and the evaluation team of the international cybersecurity exercises "Amber Mist 2018–2019" for the opportunity to gather the data. The authors are also thankful for the computing resources provided by the IT Research Center of Vilnius University. This work was partially supported by EU Horizon 2020 project No. 830892 "SPARTA".

References

Albakri, A., Boiten, E.A., de Lemos, R. (2019) Sharing cyber threat intelligence under the general data protection regulation, in: Naldi, M., Italiano, G.F., Rannenberg, K., Medina, M., Bourka, A. (Eds.), Privacy Technologies and Policy - 7th Annual Privacy Forum (APF), Rome, Italy, Proceedings, Springer. pp. 28–41. doi:10.1007/978-3-030-21752-5_3.

Barnum, S. (2012) Standardizing cyber threat intelligence information with the structured threat information expression (stix). Mitre Corporation 11, 1–22.

Beynon-Davies, P., Wang, Y. (2019) Deconstructing information sharing. Journal of the Association for Information Systems 20, 476–498. doi:10.17705/1.jais.00541.

Brilingaitė, A., Bukauskas, L., Juozapavičius, A. (2020) A framework for competence development and assessment in hybrid cybersecurity exercises. Computers & Security 88, 101607. doi:10.1016/j.cose.2019.101607.

Brown, S., Gommers, J., Serrano, O.S. (2015) From cyber security information sharing to threat management, in: Ray et al. (2015). pp. 43–49. doi:10.1145/2808128.2808133.

Burger, E.W., Goodman, M.D., Kampanakis, P., Zhu, K.A. (2014) Taxonomy model for cyber threat intelligence information exchange technologies, in: Ahn, G., Sander, T. (Eds.), Proceedings of the ACM Workshop on Information Sharing & Collaborative Security (WISCS), Scottsdale, Arizona, USA, ACM. pp. 51–60. doi:10. 1145/2663876.2663883.

Cichonski, P., Millar , T., Grance, K., Scarfone, K., (2012). NIST special publication 800-61 Rev 2: computer security incident handling guide. NIST.

Connolly, J., Davidson, M., Schmidt, C. (2014) The trusted automated exchange of indicator information (taxii). The MITRE Corporation , 1–20.

Council of European Union (2016) Regulation (ec) no 2016/679 of the european parliament and the council of the european union of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/. Official Journal European Communities 59, 1–88.

Cui, R., Allon, G., Bassamboo, A., Mieghem, J.A.V. (2015) Information Sharing in Supply Chains: An Empirical and Theoretical Valuation. Management Science 61, 2803–2824. doi:10.1287/mnsc.2014.2132.

ENISA (2013) Detect, SHARE, Protect: Solutions for Improving Threat Data Exchange among CERTs. URL: https://www.enisa.europa.eu/publications/detect-share-protect-solutions-for-improving-threat-data-exchange-among-certs.

ENISA (2016) Report on Cyber Security Information Sharing in the Energy Sector. doi:10.2824/960067.

Haass, J.C., Ahn, G., Grimmelmann, F. (2015) ACTRA: A case study for threat information sharing, in: Ray et al. (2015). pp. 23–26. doi:10.1145/2808128.2808135.

Johnson, C., Badger, L., Waltermire, D., Snyder, J., Skorupka, C., (2016). NIST special publication 800-150: guide to cyber threat information sharing. NIST, Technical Report.

van de Kamp, T., Peter, A., Everts, M.H., Jonker, W. (2016) Private sharing of iocs and sightings, in: Katzenbeisser et al. (2016). pp. 35–38. doi:10.1145/2994539.

Katzenbeisser, S., Weippl, E.R., Blass, E., Kerschbaum, F. (Eds.) (2016) Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security (WISCS), Vienna, Austria, ACM. doi:10.1145/2994539.

Menges, F., Sperl, C., Pernul, G. (2019) Unifying cyber threat intelligence, in: Gritzalis, S., Weippl, E.R., Katsikas, S.K., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (Eds.), Trust, Privacy and Security in Digital Business - 16th International Conference (TrustBus) Linz, Austria, Proceedings, Springer. pp. 161–175. doi:10.1007/978-3-030-27813-7_11.

MISP Project (2019) MISP - Malware Information Sharing Platform and Threat Sharing - The Open Source Threat Intelligence Platform. URL: https://www.misp-project.org/.

MITRE Corporation (2019a). Common Weaknesses Enumeration. URL: https://cwe.mitre.org/.

MITRE Corporation (2019b). MITRE ATT&CK^{TM}. URL: https://attack.mitre.org/.

Murdoch, S., Leaver, N. (2015) Anonymity vs. trust in cyber-security collaboration, in: Ray et al. (2015). pp. 27–29. doi:10.1145/2808128. 2808134.

NIST (2019) National Vulnerability Database. URL: https://nvd.nist.org/.

Pokorny, Z. (Ed.) (2019) The Threat Intelligence Handbook: Moving Toward a Security Intelligence Program. Second ed., CyberEdge Group.

Ray, I., Sander, T., Yung, M. (Eds.) (2015) Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security (WISCS), Denver, Colorado, USA. ACM. doi:10.1145/2808128.

Ring, T. (2014) Threat intelligence: why people don't share. Computer Fraud & Security 2014, 5–9.

Sander, T., Hailpern, J.M. (2015) UX aspects of threat information sharing platforms: An examination & lessons learned using personas, in: Ray et al. (2015). pp. 51–59. doi:10.1145/2808128.2808136.

Scott, E.D. (2009) Police Information Sharing: All-Crimes Approach to Homeland Security. LFB Scholarly Pub.

Wagner, C., Dulaunoy, A., Wagener, G., Iklody, A. (2016) MISP: the design and implementation of a collaborative threat intelligence sharing platform, in: Katzenbeisser et al. (2016). pp. 49–56. doi:10.1145/2994539.

Wagner, T.D., Palomar, E., Mahbub, K., Abdallah, A.E. (2018) A novel trust taxonomy for shared cyber threat intelligence. Security and Communication Networks 2018, 9634507:1–9634507:11. doi:10.1155/2018/9634507.

Analysis of Reports on Cyber Threats and Attacks Using Text-Analytical Software

Ladislav Burita
University of Defence; Brno, Czech Republic
ladislav.burita@unob.cz

DOI: 10.34190/EWS.20.050

Abstract: The aim of this research is to analyze reports on cyber warfare, threats and attacks, using a text-analytical (TA) software (SW). The first part is a literature review of similar solutions and a description of used methodology and tools. The TA SW used is Tovek Tools (TT), the professional product for information analysis produced by Tovek Company, Czech Republic. The used methodology includes steps of data preparation, a simple analysis of documents and their segmentation, and the detailed analysis. The author obtained reports from the Computer Incident Response Center of the Army of the Czech Republic. The data preparation contains the transformation procedure of PDF documents to plain text, the integration of these documents, selection of relevant documents, and their numbering for easy identification; the result is 302 files. Document segmentation categorizes documents, depending on the goal of the research, under the following subsets: "Cyber warfare or cyber conflict" (CWC), "Cyber threats and attacks" (CTA), "Country mentioned in reports" (CMR), and "Aggressive groups" (AG). In each segment, the focus of the analysis is specified as well as the correlation with other segments. In the CWC segment, examples of CWC acts and cyber espionage can be found. The CTA segment contains particular tactics and techniques, and details about used malware and other tools and procedures, obtained from the analyzed documents. The CMR is a supporting segment, with documents grouped under this referring to various countries; it can be used in association with other segments. The AG segment, in collaboration with other segments, reveals most known groups and can be used to analyze their activities and cooperation. The detailed analysis, provided by investigation of the reports and the typical methods and procedures used, utilized all of the TT modules. The final part of the paper is the discussion and conclusions.

Keywords: Cyber warfare, cyber threats/attacks, reports analysis, text-analytical SW, Tovek Tools

1. Introduction

The aim of the article is to analyze the "Cyber Warfare" and "Cyber Threats / Attacks" reports obtained from the Computer Incident Response Center (CIRC) of the Army of the Czech Republic (ACR). These reports are preprocessed documents converted to plain text. They are first divided into groups (segmentation), and then each group is analyzed separately. The text analytical (TA) SW, Tovek Tools (TT), is used for the analysis. In terms of scope, there are 302 reports, of which the most important ones (17) are in the reference list.

The documents come from public sources and have been issued by:

- Governmental organizations, example (Beehner, Collins, Ferenzi, Person, and Brantly, 2018).
- Cyber security workplaces, example (APT28, 2014), (RED LINE, 2016).
- ICT companies, example (HP Security, 2014), (EQUATION, 2015).
- Professional journals, example (JR03, 2010).
- Universities, example (Ellis, Jensen, Johansen, Lee, and Liles, 2013).

2. The literature review

The source of the literature was the Scopus portal (https://www.scopus.com). A search for the keywords "Cyber Warfare", "Cyber Threats", "Cyber Attacks", and "Cyber Conflict" returned 9495 documents (2019-12-04). The search was then limited to open-access documents published between 2018 and 2019; this narrowed the results to 367 papers. Finally, the papers with the maximum number of citations (10 or more citations) were selected (11 papers in all). The second part of the literature review focused on information analysis, with a search for the keywords, "information analysis" and "text document" yielding 117 documents (2019-12-04). When the search was limited to the years 2017-2019, 15 papers were eligible; these papers were subsequently ordered by the number of citations and only those with at least one citation were selected (9 papers in all).

In today's hyper-connected world, the critical infrastructure is more than ever vulnerable to cyber threats from criminal groups or individuals. As the number of interconnected devices increases, the number of potential access points to critical infrastructure for hackers grows. The paper (Tariq, et al., 2019) aims to improve understanding of the challenges encountered in securing digital infrastructure that generates big data; the functionality-based fog architecture is also defined. In addition, a comprehensive review of security requirements in fog-enabled internet of things (IoT) systems is presented. The paper (Tuptuk and Hailes, 2018) discusses the security of industrial and manufacturing systems, existing vulnerabilities of the systems, potential cyberattacks, the inefficiencies in existing measures, the levels of awareness and preparedness for future security challenges, and the need for security to play a key role in underpinning the development of future smart manufacturing systems. The paper (Davarikia and Barati, 2018) proposes a novel approach that could be taken by power system planners to improve power grid resilience by suggesting appropriate impenetrable strategies against man-made attacks or natural hazards.

The paper (Moustafa, Adi, Turnbull and Hu, 2018) addresses challenges by proposing a new threat intelligence scheme that models the dynamic interactions of industry 4.0 components, including physical and network systems. The scheme consists of a smart management module and a threat intelligence module. The proposed threat intelligence technique is designed based on beta mixture-hidden Markov models for discovering anomalous activities against both physical and network systems. For the purpose of identifying cyber threats and possible attacks, intrusion detection systems (IDS) are widely deployed in computer networks. In order to enhance the detection capability of single IDS, collaborative intrusion detection networks have been developed, which allow IDS nodes to exchange data with each other. As blockchain can protect the integrity of data storage and ensure process transparency, it has a potential to be applied to intrusion detection domain; the paper (Meng, et al., 2018) provides a review of the intersection of IDS and blockchains.

The paper (Taleby Ahvanooey, Li, Shim, and Huang, 2018) presents a comparative analysis of information hiding techniques, especially those techniques which are focused on modifying the structure and content of digital texts. Herein, the characteristics of various text watermarking and text steganography techniques are highlighted, along with their applications. In addition, various types of attacks are described and their effects analyzed, in order to highlight the advantages and inefficiencies of current techniques. The report (Ruppert, et al., 2017) presents an interactive system for the creation and validation of text clustering workflows, with the goal of exploring document collections. The system allows users to control every step of the text clustering workflow. A similar theme is elaborated in the paper (Sari, Fauzi, and Adikara, 2017). The clustering method starts with text preprocessing and indexing using inverted index. Then the index is trimmed by thresholding. After that, Term Document Matrix is built and the next step uses Latent Semantic Indexing to extract important features. The following process is seed selection via Pillar algorithm.

Résumé of the literature review: The papers are oriented to one or two topics in cyber security and document analysis. The source of the papers used by the author for document analysis in chosen topics is almost large and contains complex reports, and the analysis of the papers contains more steps in methodology that are quite original.

3. Methodology of research and tools used

The methodology of research includes some steps and used tools:

1. Data preparation.
2. Text-analytical SW Tovek Tools.
3. Simple document analysis (segmentation).
4. Detailed document analysis of respective segments.
5. Evaluation of results and discussion.

3.1 Data preparation

The data preparation step involves the integration, selection and enumeration of documents with respect to the goal of the paper. Obtained reports for information analysis are at first retrieved using simple queries, with the aim of extracting only reports (302 of them) relevant to the goals of the paper. The next step is indexing and preparation of reports for segmentation.

3.2 Text-analytical SW Tovek Tools

Tovek Tools (TT) are used to easily find information in large text data from various sources (disk files, e-mails, databases, archives, news, articles, and more). They allow the indexing of data in the Index Manager module, so that searches can be performed quickly in the Tovek Agent module using other options (date restrictions, author, origin, etc.), or using automatic search for entities (name, date, number, URL address, geographic information) in texts. Source information can be in different formats. With TT, projects can store multiple queries, including their settings and other information, and create different sets of documents that can be updated with new data.

In addition to simple queries containing only a few search terms, complex queries can also be created using the Query Editor module. The Tovek language can formulate queries that contain the precise search terms being looked for in documents. The InfoRating module, for context analysis, allows further analysis of found documents with respect to other queries and the discovery of how the individual issues represented by context queries are related. Documents can be exported to the Harvester module, can be used for automatic detection of document content (content analysis), and, on the basis of statistical calculations, for finding out what is in the documents (SW Tovek, 2019). All five modules are used for analysis in the paper.

3.3 Simple document analysis (segmentation)

Segmentation of analyzed documents is based on searching the documents for given keywords. Results are depicted in tables and graphs.

3.4 Detail document analysis of respective segments

The detailed document analysis is in segments 1, 2, and 4, oriented to the important documents ranked as significant (see Tab. 2). Segment 1 includes eight documents (450 pages), segment 2 includes seven documents (almost 330 pages), and segment 4 includes six documents (280 pages). The strategy of the process is complex query, context and content analyses. The goal is to present important and interesting information in relation to the themes mentioned in section 4. The detailed analysis in segment 3 is limited by the scope of the article and will not be addressed.

3.5 Evaluation of results and discussion

Results of the analysis are evaluated with relevant information in the cyber security domain and in the context of methodology of research. The results show the important decisions in the large volume of reports (302), which altogether are nearly 10 thousand pages.

4. Document segmentation

The simple document analysis includes segmentation into the four themes (Tab. 1). Segmentation according to the countries mentioned in the documents is in Fig. 1. The most relevant papers (see Tab. 2) belong to more segments and cover more topics than the others do. All documents from Tab. 2 (report 004 to 296) are included in the reference list.

Table 1: Segmentation of the 302 analyzed documents - source own

Segment	Num-doc	%
1. Cyber warfare or cyber conflict	14	4,6
2. Cyber threats / attacks	27	8,9
3. By country, mentioned in reports	223	73,8
4. Aggressive groups	68	22,5

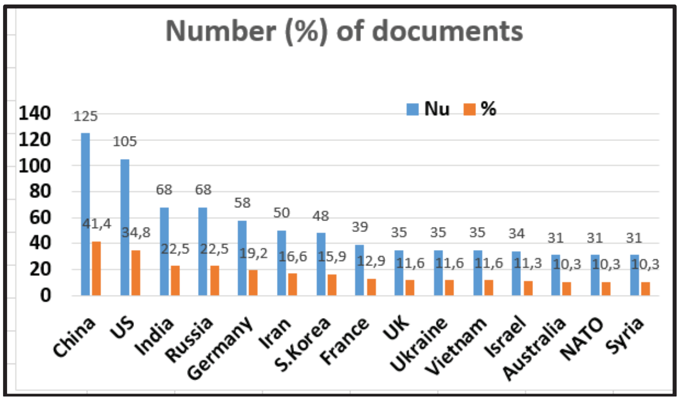


Figure 1: Countries and organizations mentioned in the documents - source own

Analysis of Segment 1 shows examples of cyber warfare in hybrid war (Russia - Georgia), cases of cyber-attacks in industries and the private sector, and samples of cyber espionage, where victims were diplomatic missions, government entities, organizations of national security and defense, and academics or journalists. Analysis of Segment 2 reveals examples of a certain cyber-attack methodology and includes description of its tactics, techniques, and procedures. Content analysis depicted the “cyber” situation. Results of segment 3 can be seen in Fig. 1. Analysis of Segment 4 shows examples of APT28, APT1, and Lazarus group.

Table 2: Set of important documents - source own

Document number:		4	5	16	24	42	44	51	72	95	110	111	121	145	158	206	220	296
1	**Seg1**	1	1					1		1	1		1	1			1	
2	**Seg2**		1	1				1		1			1	1	1			
3	Seg3 **China**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	Seg3 **US**	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
5	Seg3 **India**				1		1	1	1	1	1	1		1	1	1	1	1
6	Seg3 **Russia**	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1
7	Seg3 **Germany**	1	1		1	1	1	1	1		1	1	1			1	1	1
8	Seg3 **Iran**	1	1	1	1	1	1	1	1	1	1	1			1			
9	Seg3 **S.Korea**			1		1		1		1	1	1	1	1	1	1		
10	Seg3 **France**		1		1	1	1	1	1		1	1		1		1		1
11	Seg3 **UK**		1	1		1	1	1	1		1			1		1	1	1
12	Seg3 **Ukraine**	1	1	1	1	1						1						1
13	Seg3 **Vietnam**		1		1					1	1	1				1		1
14	Seg3 **Israel**	1	1			1		1		1	1		1			1		
15	Seg3 **Australia**				1		1			1	1	1		1		1	1	
16	Seg3 **NATO**	1		1		1				1	1	1			1		1	
17	Seg3 **Syria**			1		1			1	1		1	1	1				
18	**Seg4**			1					1			1		1	1	1		
SUMMARY:		9	12	11	10	12	9	12	10	13	14	14	9	10	9	12	9	9

5. Detailed document analysis

The detailed analysis focuses on finding the most important and interesting themes with respect to methodology of the research and results in section 4. The SW TT is used in context and content analysis.

5.1 Segment 1: Cyber warfare or cyber conflict

The 2008 Russia-Georgia War (Beehner, Collins, Ferenzi, Person, and Brantly, 2018) highlighted not necessarily a new form of conflict but rather the incorporation of a new dimension to that conflict, namely cyberspace. Where states once tried to control the radio waves, broadcast television channels, newspapers, and other forms of communications, they now add to these sources of information the control of cyberspace and its component aspects, like websites and social media. The conflict was a limited war with limited objectives; it marked the first invasion by Russian ground forces of a sovereign nation since the Cold War. It also marked a breakthrough in the integration of cyberwarfare and other non-kinetic tools into conventional strategies, what some authors in the West have termed “hybrid warfare”. The digital attacks, which coincided with Russia’s advances into Georgia and indicated a change in warfighting tactics, constituted a new form of warfare. The timing of the digital attacks was problematic to the Russian narrative for several reasons. First, it indicated coordination, if not direction, by the Russian state in the perpetration of cyberattacks against Georgian entities. Second, it highlighted the near fait accompli and inevitability of the conflict. What made the cyberattacks against Georgia more significant was not how they affected the human layer of cyberspace domestically, but rather how it constrained the tools available to the government in Tbilisi for conveying their narrative in the early stages of the conflict to the international community. Attacks that degraded the communication capabilities of the state also included the influx of propaganda and disinformation. The degraded communications were achieved through three primary types of attacks against the logical layers of the Georgian Internet, viz.: cross-site scripting, SQL injections, and distributed denial of service (DDoS) attacks. The attacks reinforced the Russian view of cyberspace as a tool for psychological manipulation and information warfare and highlighted the role of third forces, Russian “patriot’s hackers”, and other non-state actors in the modern battlefield.

The lessons learned from the conflict (Beehner, Collins, Ferenzi, Person, and Brantly, 2018) include:

- There is a need to establish and develop greater territorial and stronger national defense.
- It is important that civilian leaders are prevented from micromanaging and expanding urban-operations training.
- There is an urgent need to improve their combined arms capability and restructure their air force.

- Nations must strive to strengthen their cyber defenses and ties with Western institutions.

Saga espionage (Ellis, Jensen, Johansen, Lee, and Liles, 2013) of long-term campaigns has been garnering a lot of attention, and for good reasons. Some have asserted that certain campaigns have existed since the early 2000's and their existence has only recently come to light in the private sector. The damage caused by these types of breaches is difficult to estimate because it occurred over such a long time span, but in some cases terabytes of data were stolen over a period of a few months. When taken in relation to the oil industry, where proprietary information like bid exploration data is the lifeblood of an organization, this can be a disastrous blow. The list of campaigns/attacks follows.

- NightDragon (2007-2009) - Exfiltration of competitive proprietary operations and project-financing information with regards to oil and gas field bids and operations and collection of data from SCADA systems. Entry methods: Social engineering, spear phishing, SQL injection. Affected countries: US, Taiwan, Kazakhstan, and Greece.
- Elderwood (2009) - The wholesale gathering of intelligence and intellectual property. Entry methods: Watering-hole, spear phishing.
- ShadyRAT - Exfiltration of a historically unprecedented transfer of wealth, closely guarded national secrets, source code, bug databases, email archives, negotiation plans and exploration details for new oil and gas field auctions, document stores, legal contracts, supervisory control and data acquisition, design schematics, and much more. Entry method: Spear phishing. Affected country: US.
- Mirage (2012) - Theft of intellectual property and company secrets. Entry Method: Social engineering, spear phishing, SQL injection of web servers. Affected countries: Philippines, Canada.
- Red October (2007) - Intelligence gathering from the compromised organizations. Entry methods: Social engineering, spear phishing, SQL-injection of web servers. Countries with companies affected: Azerbaijan, Belarus, Turkmenistan, and UAE.

The report (JR03, 2010) contains an analysis of data stolen from politically sensitive targets (Offices of the Dalai Lama and agencies of the Indian security establishment) and recovered during investigation. Databases containing sensitive information on citizens of numerous third-party countries, as well as personal, financial, and business information, were also infiltrated and recovered during the course of the investigation. Victims were diplomatic missions, government entities, security and defense organizations, academics, and journalists. China has a hacker community that has been associated with targeted attacks in the past, and has been linked through informal channels to elements of the Chinese state, although the nature and extent of the connections remains unclear. One common theme regarding attribution of related attacks emerges from variations of a privateering model, in which the state authorizes private persons to perform attacks against enemies of the state. This "patriotic hacking" model emerged because, as studies have shown, there is no direct government control over the loosely connected groups of hackers. However, this ambiguous relationship does not mean that there is no connection between the activities of Chinese hackers and the state. China relies on a broad informal network of students, tourists, teachers, and foreign workers in host nations to collect small bits of information to form a composite picture of the environment.

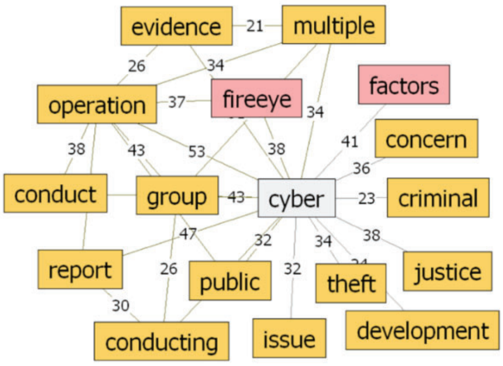


Figure 2: Word connections to the word "cyber" in analyzed reports- source own

Content analysis using TT SW can highlight word connections in various sets of the document sources. Fig. 2 depicts the connections with the word "cyber" in important papers (upon the Tab. 2).

5.2 Segment 2: Cyber threats/attacks

Since at least 2012 (Siboni, Fallon, Weatherford, Rogers, and Moss, 2014), Iranian actors have directly attacked, established persistence, and extracted highly sensitive materials from the networks of government agencies and major critical infrastructure companies in the following countries: Canada, China, England, France, Germany, India, Israel, Kuwait, Mexico, Pakistan, Qatar, Saudi Arabia, South Korea, Turkey, United Arab Emirates, and the US. Iran is the new China.

5.3 Tactics, techniques and procedures used by Iran (Siboni, Fallon, Weatherford, Rogers, and Moss, 2014)

5.3.1 Initial compromise

This phase gets the attackers their first foothold into the target network. Once the ability to execute arbitrary code has been established, an attacker's job becomes quite a bit easier. The vector of initial compromise is usually determined by what the target is vulnerable to. Used techniques: SQL injection, spear phishing.

5.3.2 Privilege escalation & pivoting

Escalation is a category of techniques that describes the process of going from a less privileged user on a compromised computer to a more privileged user. This increase in privileges allows the attacker to gain access to privileged areas of the operating system as well as to infect other computers on the target network. The team did not utilize any novel methods of privilege escalation, but only used a variety of publicly known exploits for escalation of privileges on unpatched Windows operating systems, from an unprivileged user to kernel-level privilege. List of used techniques and tools is as follows:

- Cached Credential Dumping – A method of pivoting on a Windows system network is to extract domain credentials which have been used on the compromised computer from a credential cache.
- zhMimikatz and MimikatzWrapper - Two similar applications that determine which version of Mimikatz to use, depending on whether the computer's version of Windows is 32bit or 64bit.
- PsExec Spreading - Once an attacker has credentials extracted from the cache, whether in hash form or in plaintext form, PsExec can be used to run commands on any other computer, which accepts those domain credentials. If this technique is combined with cached credential dumping, it can be used to jump from computer to computer on a compromised network.
- MS08-067 Exploit – This is a vulnerability in Microsoft Windows, made popular by the Conficker worm which can be exploited by a specially crafted packet to the operating system's RPC network interface.
- Cain & Abel - This is a publicly available toolkit, which covers a wide range of functionality that assists attackers once they have compromised a node on a network.

5.3.3 Exfiltration

Exfiltration is the process of moving information to an external site; it is the process of stealing information without being detected. List of used techniques and tools is as follows:

- Anonymous FTP Servers – These are enabled in order to pilfer large quantities of information. The servers allow attackers to use existing command line available on their targets to upload information.
- SMTP – Here, the sending is performed by internally developed malware samples in order to exfiltrate information about the infected computer.
- SOAP – A sub-protocol communicated via HTTP, SOAP is used as the command and control protocol for TinyZBot, which is the preferred backdoor; SOAP has undergone long-term development.

5.3.4 Persistence

This is the phase where access to a compromised network is sustained. TinyZBot is a backdoor developed in C#. The goal of TinyZBot is to gather information from an infected computer as well as to gain and maintain access into a compromised network.

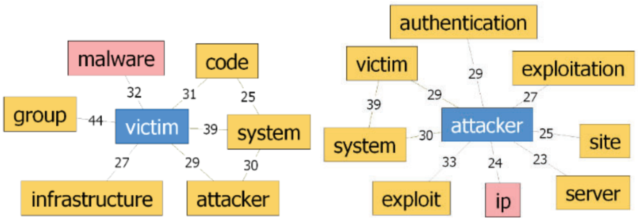


Figure 3: Word connections to the words "attacker" and "victim" - source own

Fig. 3 depicts the result of content analysis in the connections with the words “attacker” and “victim”.

The APT1 (APT1, 2014) has a well-defined attack methodology honed over the years and designed to steal massive quantities of intellectual property. The APT1 files begin with aggressive spear phishing, proceed to deploy custom digital weapons, and end by exporting compressed bundles of files to China, before beginning the cycle again. They employ good English in socially engineered emails; they have evolved their digital weapons for more than seven years, resulting in continual upgrades, as part of their own software release cycle.

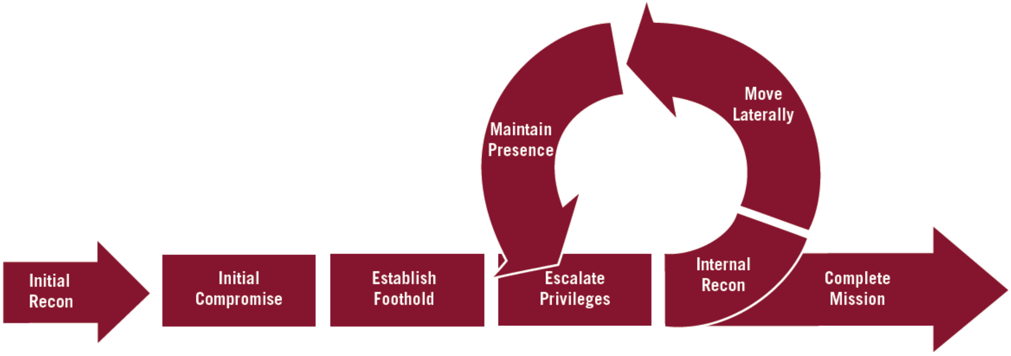


Figure 4: APT1 attack life cycle - source (APT1, 2014)

5.4 The APT1 attack life cycle, see Fig. 4, (APT1, 2014)

5.4.1 The Initial Reconnaissance and Compromise

This phase represents the methods intruders use to first penetrate a target organization’s network. Spear phishing is the most commonly used technique; emails contain either a malicious attachment or a hyperlink to a malicious file. The subject line and the text in the email body are usually relevant to the recipient. The webmail accounts are created using real people’s names that are familiar to the recipient, such as a colleague, a company executive, or an IT department employee, and these accounts are used to send the emails. Establishing a foothold involves actions that ensure the control of the target network’s systems from outside the network.

5.4.2 Establish a foothold

APT1 establishes a foothold once email recipients open a malicious file and a backdoor is subsequently installed. In almost every case, backdoors initiate outbound connections to the intruder’s command and control server. Intruders employ this tactic because while network firewalls are generally adept at keeping malware outside the network from initiating communication with systems inside the network, they are less reliable at keeping malware that is already inside the network from communicating to systems outside.

5.4.3 Escalate privileges

This phase involves acquiring items (most often usernames and passwords) that will allow access to more resources within the network. In this and the next two stages, publicly available tools are predominantly used to dump password hashes from victim systems, in order to obtain legitimate user credentials.

5.4.4 Internal Reconnaissance

In this stage, the intruder collects information about the victim environment. APT1 primarily uses built-in operating system commands to explore a compromised system and its networked environment.

5.4.5 Lateral Movement

Once an intruder gains a foothold inside the network and has access to a set of legitimate credentials, it is simple to move around the network undetected. They can connect to shared resources on other systems or execute commands on other systems using the publicly available “psexec” tool from Microsoft Sysinternals or the built-in Windows Task Scheduler (“at.exe”). These actions are hard to detect because legitimate system administrators also use these techniques to perform actions around the network.

5.4.6 Maintain Presence

The intruders take actions to ensure continued, long-term control over key systems in the network environment. They achieve this by installing new backdoors on systems, by using legitimate VPN credentials, or by logging in to web portals.

5.4.7 Completing the Mission

When APT1 find files of interest, they pack them into archived files before stealing them. Intruders most commonly use the RAR archiving utility for this task and ensure that the archives are password-protected.

5.5 Segment 4: Aggressive groups

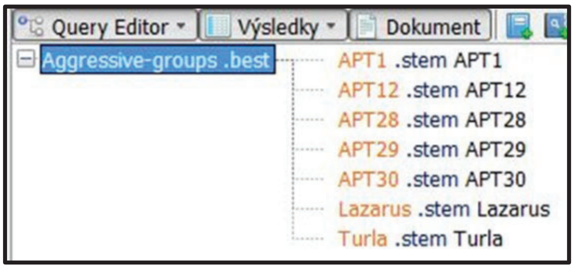
APT28 targets insider information related to governments, the military, and security organizations that would likely benefit the Russian government. Since 2007, the group has systematically evolved its malware, using flexible and lasting platforms indicative of plans for long-term use. The coding practices evident in the group's malware suggest both a high level of skill and an interest in complicating reverse engineering efforts. Malware compile times suggest that the team of developers have consistently updated their tools over the last seven years. APT28 uses spear phishing emails to target its victims, a common tactic in which the threat group crafts emails that mention specific topics relevant to recipients. This increases the likelihood of recipients opening the message, opening any attached files, or clicking on a link in the body of the email, since their interest has been piqued and they believe that the email is legitimate (APT28, 2014).

APT1 is a single organization of operators that has conducted a cyber espionage campaign against a broad range of victims since at least 2006 (APT1, 2014). It is one of the most prolific cyber espionage groups in terms of the sheer quantity of information stolen. The activity we have directly observed likely represents only a small fraction of the cyber espionage that the group has conducted – the group's intrusions against nearly 150 victims over seven years was analyzed and traced to four large networks in Shanghai, two of which are allocated directly to the Pudong New Area. APT1 has systematically stolen hundreds of terabytes of data from at least 141 organizations, and has demonstrated the capability and intent to steal from dozens of organizations simultaneously. Since 2006, 141 companies, spanning 20 major industries, have been compromised. The group has a well-defined attack methodology (see section 5.2), honed over the years and designed to steal large volumes of valuable intellectual property. Once access has been established, the APT1 files periodically revisit the victim's network over several months or years and steal broad categories of intellectual property, including technology blueprints, proprietary manufacturing processes, test results, business plans, pricing documents, partnership agreements, and emails and contact lists from the victim organization's leadership. APT1 focuses on compromising organizations across a broad range of industries in English-speaking countries and maintains an extensive infrastructure of computer systems around the world. In the years 2012 and 2013, APT1 established a minimum of 937 command and control servers, hosted on 849 distinct IP addresses in 13 countries. The majority of these addresses were registered to organizations in China (709), followed by the US (109) (APT1, 2014).

The Lazarus group (Operation Blockbuster, 2016) has been active since at least 2009, and was responsible for the November 2014 destructive wiper attack against Sony Pictures Entertainment (SPE). The attack broke new ground not only as a destructive malware attack on a US commercial entity, but also due to the fact that the US government attributed the attack to North Korea and enacted small reciprocal measures. The characteristics of the Lazarus group include:

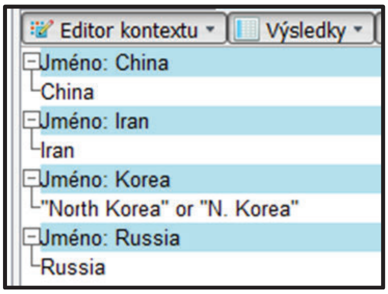
1. It is a well-established group that produces various sets of custom malware.
2. The group demonstrates varying levels of technical proficiency in computer network operations.
3. It shows a heavy reliance on shared code, techniques, and ideas from other previously developed tool components, as well as from outside sources.
4. The malware analyzed in the SPE operation has been used to target government, media, military, aerospace, financial, and critical infrastructures, primarily in South Korea and the US.
5. The set of malware uncovered and analyzed during this operation, more than 45 unique families to date, consists of a wide variety of attack tools, viz.: Rats, General Tools, Uninstallers, installers, spreaders, proxy, Keylogger, DDoS Bot, Loaders, and Hard Drive Wipers.
6. The frequency and type of code sharing across malware families suggests the same group of author(s) across families or extensive sharing of resources between closely linked groups.
7. The group has also been observed to share cryptographic keys across malware families as well as general techniques in other unrelated malware families.

The procedure and results of context analysis of the Lazarus group are shown in Figures 5, 6, 7, and 8. First, a complex query for APT groups was created using Query editor (see Fig. 5). The result of the query was then selected only for "Lazarus", and a contextual query (see Fig. 6) was elaborated on the result to find out whether the Lazarus group has connections with Russia, China, and other countries. The result is in the context matrix (see Fig. 7). Relations between Lazarus, Russia and China are analyzed in four relevant documents (see Fig. 8) and are commented after the detailed investigation.



Aggressive-groups .best
- APT1 .stem APT1
- APT12 .stem APT12
- APT28 .stem APT28
- APT29 .stem APT29
- APT30 .stem APT30
- Lazarus .stem Lazarus
- Turla .stem Turla

Figure 5: The complex query for APT groups - source own



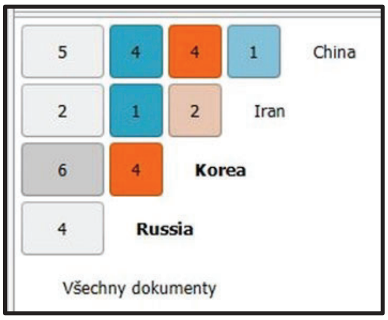
Jméno: China
- China

Jméno: Iran
- Iran

Jméno: Korea
- "North Korea" or "N. Korea"

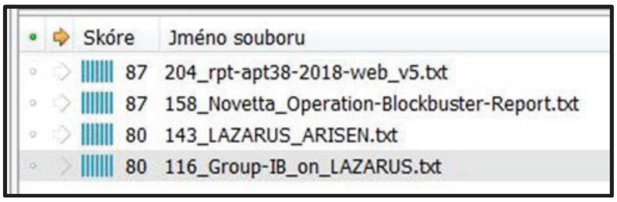
Jméno: Russia
- Russia

Figure 6: The context query for Russia, China, and other countries - source own



5	4	4	1	China
2	1	2	Iran	
6	4	Korea		
4	Russia			
Všechny dokumenty				

Figure 7: The context matrix to the context query - source own



Skóre	Jméno souboru
87	204_rpt-apt38-2018-web_v5.txt
87	158_Novetta_Operation-Blockbuster-Report.txt
80	143_LAZARUS_ARISEN.txt
80	116_Group-IB_on_LAZARUS.txt

Figure 8: Documents for analysis of relations between Lazarus, Russia and China - source own

5.6 Explanation of the relations between Lazarus, Russia and China from the documents in Fig. 8

Lazarus-Russia:

- The hackers have tried to mask their activity by pretending to be Russian hackers.
- The Client_TrafficForwarder module includes debugging strings containing Russian words.
- "Russian commands" received from the server are not typical of a Russian native speaker.
- Hackers used Enigma Protector, a commercial product created by a Russian SW developer.
- The sets of exploits were borrowed from Russian-speaking hackers.
- Evidence in the Sony hack attack suggests possible involvement by Iran, China or Russia.
- The NACHOCHEESE malware contained poorly translated Russian-language strings.

Lazarus-China:

- Pharmaceutical companies in Japan and China were attacked.

- The used network belongs to an autonomous system China Netcom; other C2 locations include Taiwan, Indonesia, India, and China. IP range is registered to a company in China.
- The group malware was found in other Asian countries like Taiwan, China, Japan, and India.

6. Results evaluation and discussion

The main result of the reports analysis in the paper is a methodology how makes order in a huge volume of information and get interesting and important knowledge. Segmentation upon the strategic goals of analysis and selection about the criterions is a way for choosing the most relevant papers that are object to detail investigation and getting the desired parts. The used approach is possible to apply in all areas of interest for the analytical reason. One of the interesting information from analyzed sources in cyber security is so called "patriotic hacking", in our case for Russia and China. In case of Russia it was "selfless" support of aggression against Georgia and Ukraine. In case of China, an expression of a deep ground or good state organization leads to the continuous sending of information to hacker's country of origin.

7. Conclusions

This paper analyzed documents in cyber security area gotten from CIRC ACR. Results of the analysis have been divided into segments and, in each segment, interesting and important information is published from the document sources. The own research methodology was applied, supported by SW TT. The results are clear and useful, although more recent reports are available. These examples clearly illustrate how vulnerable the cyberspace is. To reduce and prevent cyberattacks, it is necessary for governments and organizations to be relentless in the fight against cyber threats.

To keep track of cybersecurity, it is advisable continuously monitor reports and journals such as FireEye Reports, Information warfare monitor. NOVETTA Reports, etc.

Acknowledgements

This article presents the results of the research in cyber security as part of the project (DZRO-209, 2019).

References

Adversary. (2019). "Adversary tradecraft and the importance of speed", Global threat report, 75 pp. [report 111], https://www.crowdstrike.com/resources/reports/2019-crowdstrike-global-threat-report/

APT1. (2014). "APT1 Exposing One of China's Cyber Espionage Units", Mandian, 76 pp. [report 145], https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf

APT28. (2014). "A window into Russia's cyber espionage operations?", FireEye Inc., 45 pp. [report 016], https://www.fireeye.com/content/dam/fireeye-www/global/en/current-threats/pdfs/rpt-apt28.pdf

Beehner Lionel, Collins Liam, Ferenzi Steve, Person Robert, and Brantly Aaron. (2018). "Analyzing the Russian Way of War, Evidence from the 2008 Conflict with Georgia", Modern War Institute at West Point, 98 pp. [report 004], https://mwi.usma.edu/wp-content/uploads/2018/03/Analyzing-the-Russian-Way-of-War.pdf

Crouching. (2012). "Whitepaper: Crouching Tiger, Hidden Dragon, Stolen Data", Context Information Security, 33 pp. [report 044], https://paper.seebug.org/papers/APT/APT_CyberCriminal_Campagin/2012/Crouching_tiger_hidden_dragon.pdf

Davarikia, H., and Barati, M. (2018). "A tri-level programming model for attack-resilient control of power grids". Journal of Modern Power Systems and Clean Energy, volume 6, issue 5, pp. 918-929.

DZRO-209. (2019). "Development of systems C4I and cyber security," Brno, University of Defence, 2016-2020.

Ellis, K., Jensen, W., Johansen, T., Lee, K., and Liles, S. B. (2013). "CERIAS Tech Report 2013-9 Crude Faux: An analysis of cyber conflict within the oil & gas industries", Center for Education and Research Information Assurance and Security Purdue University, West Lafayette, 36 pp. [report 005], https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2013-9.pdf

EQUATION. (2015). "Equation group: Questions and answers, Version 1.5", Kaspersky Lab HQ, Moscow, Russian Federation, 44 pp. [report 072], https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/08064459/Equation_group_questions_and_answers.pdf

Fagerland Snorre and Grange Waylon. (2014). "Whitepaper: The Inception Framework: Cloud - hosted APT", Blue Coat Systems, Inc., 62 pp. [report 024], https://paper.seebug.org/papers/APT/APT_CyberCriminal_Campagin/2014/bcs_wp_InceptionReport_EN_v12914.pdf

Geers Kenneth, Kindlund Darien, Moran Ned, and Rachwald Rob. (2014). "WORLD WAR C: Understanding Nation-State Motives Behind Today's Advanced Cyber Attacks", FireEye, 22 pp. [report 095], https://www.fireeye.com/content/dam/fireeye-www/global/en/current-threats/pdfs/fireeye-wwc-report.pdf

HP Security. (2014). “Profiling an enigma: The mystery of North Korea’s cyber threat landscape”, HP Security Briefing Episode 16, 75 pp. [report 121], https://www.slideshare.net/crash1980/profiling-an-enigma-the-mystery-of-north-koreas-cyber-threat-landscape

JR02. (2009). “Tracking GhostNet: Investigating a Cyber Espionage Network”, Information Warfare Monitor, 53 pp. [report 110], http://www.nartv.org/mirror/ghostnet.pdf

JR03. (2010). “Shadows in the Cloud: Investigating Cyber Espionage 2.0”, Information Warfare Monitor, 52 pp. [report 220], https://itlaw.wikia.org/wiki/Shadows_in_the_Cloud:_Investigating_Cyber_Espionage_2.0

Meng, W., Tischhauser, E.W., Wang, Q., Wang, Y., and Han, J. (2018). “When intrusion detection meets blockchain technology: A review”. IEEE Access, volume 6, pp. 10179-10188.

Moustafa, N., Adi, E., Turnbull, B., and Hu, J. (2018). “A New Threat Intelligence Scheme for Safeguarding Industry 4.0 Systems”. IEEE Access, volume 6, pp. 32910-32924.

Operation Blockbuster. (2016). “Unraveling the long Thread of the Sony Attack”, NOVETTA, 58 pp. [report 158], https://www.novetta.com/2016/02/operation-blockbuster-unraveling-the-long-thread-of-the-sony-attack/

RED LINE. (2016). “Red Line Drawn: China Recalculates Its Use of Cyber Espionage”, FireEye, 16 pp. [report 206], https://www.fireeye.com/blog/threat-research/2016/06/red-line-drawn-china-espionage.html

Ruppert, T., Staab, M., and all. (2017). “Visual interactive creation and validation of text clustering workflows to explore document collections”. IS and T International Symposium on Electronic Imaging Science and Technology 2017, pp. 46-57.

Sari, Y.A., Fauzi, M.A., and Adikara, P.P. (2017). “Optimizing K-means text document clustering using latent semantic indexing and pillar algorithm”. The 5th International Symposium on Computational and Business Intelligence, article number 8053549, pp. 81-85.

Siboni Gabi, Fallon William J., Weatherford Mark, Rogers Michael, and Moss Jeff. (2014). “Cylance Operation”, Cleaver Report, 68 pp. [report 051], https://www.cylance.com/content/dam/cylance/pages/operation-cleaver/Cylance_Operation_Cleaver_Report.pdf

SW Tovek. (2019). “The text-analytical software TOVEK”, [online], TOVEK, Prague, https://www.tovek.cz

Taleby Ahvanooey, M., Li, Q., Shim, H.J., and Huang, Y. (2018). “A Comparative Analysis of Information Hiding Techniques for Copyright Protection of Text Documents”. Security and Communication Networks, volume 2018, article number 5325040.

Tariq, N., Asim, M., and all. (2019). “The security of big data in fog-enabled IoT applications including blockchain: A survey”. Sensors (Switzerland), issue 8, article number 1788.

Tuptuk, N., and Hailes, S. (2018). “Security of smart manufacturing systems”. Journal of Manufacturing Systems, volume 47, pp. 93-106.

Villeneuve Nart and Bennett James. (2012) “Detecting APT Activity with Network Traffic Analysis”, Trend Micro Incorporated Research Paper, 13 pp. [report 296], https://www.bankinfosecurity.com/whitepapers/detecting-apt-activity-network-traffic-analysis-w-663

Tracking Botnets on Nation Research and Education Network

Ivan Daniel Burke[1] and Alan Herbert[2]
[1]Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Rhodes University, Grahamstown, South Africa
iburke@csir.co.za
a.herbert@ru.ac.za

DOI: 10.34190/EWS.20.107

Abstract: The South African National Research and Education Network (SA NREN) proves network connectivity and services to all tertiary education networks and research councils within South Africa. The NREN forms part of South Africa's national integrated cyber infrastructure, as such, it is a potential target for cyber-attacks. Due to the large volume of traffic and decentralised nature of the SA NREN, monitoring, reporting and mitigating cyber-attacks is a complex problem. The NREN Cyber Incident Response Team (CSIRT) uses network flow data to identify early indicators of cyber-attacks. In this paper the focus will be on the mechanisms used to identify malicious botnet traffic using network flow analysis.

Keywords: Network flow analysis, NREN, Network Traffic Analysis, botnet detection, cyber threat detection.

1. Introduction

The South African National Research and Education Network (SA NREN) provides the backbone infrastructure for most of South African research and education institutions. This network operates similar to the American, ESnet, or the European SURFnet and CESNET organisations. These NRENs provide services for these research institutions such as network interconnectivity, large data transfer capability, data warehousing and data processing services. In the SA NREN, the backbone is maintained and secured through the collaboration of two institutions: the Tertiary Education and Research Network of South Africa (TENET) which is a service organisation which maintains the services running within the NREN and the South African National Research Network Competency Area (SANReN CA) both these organisations form part of the National Integrated Cyber Infrastructure System (NICIS) organisation.

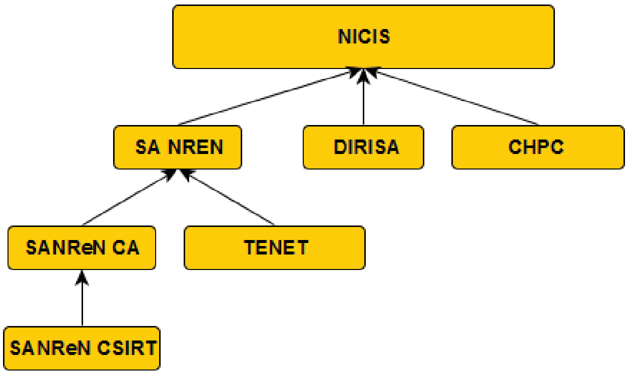


Figure 1: NICIS organogram

Within the SANReN CA a Cyber Security Incident Team (CSIRT) was established in 2016 to address the growing concern of cyber threats against the NREN. Within the CSIRT various tools and mechanisms are used to detect network anomalies and investigate cyber-attacks. The CSIRT aims to provide a proactive cyber incident capability to the SA NREN user base. Thus the CSIRT monitors the NREN for signs of anomalies or abuse before it is reported to the SANReN CSIRT team. As part of this capability the CSIRT seeks to detect: Port scans, Brute force attacks, Denial of Service and Botnet detection. However, this paper focuses the work done within the CSIRT, regarding botnet detection using network flow analytics.

The reason for using network flow analytics is primarily due to the reduced storage size requirement as appose to storing raw packet data. This has the additional benefit of protecting user privacy. Each academic entity within the NREN is entitled to their privacy. Network flow data allows the CSIRT to monitor the traffic flows without violating the user's privacy. The negative trade off of not capturing data payloads is that it makes certain attacks, such as phishing email and drive-by downloads, nearly impossible to detect.

Not all institutions within South Africa have similar budgets to address ICT or cyber-crime related issues. The SANReN CSIRT aims to supplement existing institutional ICT capabilities where needed.

This paper describes the research effort by the SANReN CSIRT to detect Botnets by addressing the following key topics. Firstly the current research methodologies to detect malicious network activity is discussed in section 2. Next, a brief overview of active botnet families, during the observation period is presented in section 3. In section 4 the network flow collection environment is explained. Section 5 showcases the results achieved by applying the research to our data set. In section 6 the paper concludes with our findings and recommendation for future work.

2. Related Work

For the purpose of this paper a botnet is defined as a collection of infected hosts (computers, routers, IoT devices, etc.), known as bots, which are controlled by a bot master via a collection of C2 servers to perform coordinated cyber-attacks. Based on the targets and C2 structures used these botnets can be classified into various botnet families or strains.

Amini, et al. (2014) created a categorisation of techniques used to identify botnets using network flow data. These categories were as follows:

- Correlation: These techniques uses statistical analysis to identify correlations between user access patterns and the number of unique flows. The DISCLOSURE system proposed by Bilge, et al. (2012) is an example of using correlation to detect botnet activity.
- Clustering: This technique aims to cluster together similar traffic patterns into groupings of similar behavioural patterns. This is the technique used by Amini, et al. (2014) to identify botnets.
- IRC channel monitoring: IRC is a well-known communication channel used by botnets to perform Command and Control (C2) activities (Feily, et al., 2009, Amini, et al., 2014). These techniques focus on behavioural fatteners of IRC communications (Mazzariello, 2008). Network flow analysis does not provide new insight into the detection of IRC botnet detection and previous work by other research have covered these techniques in great detail (Mazzariello, 2008).
- Machine learning algorithms: Due to the fact that Machine Learning (ML) algorithms are widely used to identify patterns in large data sets and to perform clustering on seemingly disjoint data sets (Wagner, et al., 2011). ML algorithms was used by Bilge, et al. (2012), Hofstede (2013) and Amini, et al. (2014) to assist with clustering and correlation detection.
- DNS request monitoring: Botnets seek to hide their C2 server infrastructure by using Domain Generating Algorithms (DGA) and Fast-Flux techniques (Grill, et al., 2015, Erquiaga, et al., 2016). This results in an abnormally high number of DNS look-up requests from hosts which are infected by bots.

DISCLOSURE is a botnet detection platform which uses network flow analysis to identify C2 servers within a network (Bilge, et al., 2012). The system uses the random forest machine learning algorithm to classify servers on a network as either benign or malicious. The DISCLOSURE system achieves this classification by first computing several features of the network flow data. The feature extraction algorithms focuses on flow-size, client access patterns and temporal behaviour as input for the classifier.

The flow size features used by the DISCLOSURE system are the unique flow sizes (number of bytes) observed during a period, statistical features such as mean flow size and standard deviation from the mean flow size. The researchers also used autocorrelation to derive additional statistical features based on the flow size (Bilge, et al., 2012).

Bilge, et al. (2012) theorized that benign user interaction with a server results in greater variation in flow sizes than that of a botnet C2 communication. Thus it would be expected that the standard deviation observed by benign servers would be greater that the deviation observed for C2. Furthermore benign servers should have a higher number of unique flow sizes compared to C2 server. These features assisted the classifier in classifying flows as either benign or malicious.

DISCLOSURE used client access patterns to further classify potential C2 servers. Bilge, et al. (2012) monitored regular access patterns by creating a data series wherein the time delay between consecutive server access attempts from each client is captured. Then statistical analysis was performed based on the inter-arrival times of each client. Bilge, et al. (2012) theorized that bots communicate with C2 server in short regular burst in order to avoid detection.

The final feature taken into consideration for the DISCLOSURE classifier is the temporal feature related to the time of the day it is expected for users to be more active. Thus it is uncommon for users to be equally active during the early hours of the day or late at night compared to the activity observed during the middle of the day.

Amini, et al. (2014) used GNS3 network simulator to create a test environment for their cluster testing. Amini, et al. (2014) infected nodes within the network with the Zeus bot. Amini, et al. (2014) created their clusters by first running the simulated environment for a time period whilst collecting the network flow data. After concluding the experimental run benign, traffic was filtered based on whitelist rules and black list rules, resulting in two disjoint data sets. These data sets were then used to identify C2 flow events and create behaviour rules for identifying C2 communication activity. A critique of the methodology used within Amini, et al. (2014) work is that only one family of botnets was tested (namely Zeus) and the test was conducted on a small test network, less than ten nodes, with a known infection. During our own testing, the results obtained by Amini, et al. (2014) could not be replicated in a larger network were the existence of a botnet was unknown.

Hofstede (2013) used the Exponentially Weighted Moving Average (EWMA), a ML algorithm, to detect network intrusions in real time. The EWMA algorithm reduces the significance of previous measurements based on the amount of time which has passed since the measurement was taken. Thus the weight of a measurement on the average measurement is reduced over time. Hofstede (2013) introduced a supplementary algorithm which they refer to as a seasonal EWMA. The seasonal algorithm takes into account that network traffic usage is influenced by the user’s utility of the network at a given time of day. Thus Hofstede (2013) selected a seasonal period of one day and averaged the EWMA results with the measurement taken at the same time the previous day. The seasonal algorithm extension is an alternative to the temporal feature extraction proposed Bilge, et al. (2012), both approaches consider the influence the time of day may have on network utility.

Hofstede (2013) collected data from the CESNET backbone network over a 14-day period in 2012.

3. Active Botnets During the Observation Period

The network attack detection research within the SANReN CSIRT started in November 2017. Based on our current findings a large volume of the attacks are directed at Internet of Things (IoT) devices and specifically Huawei networked devices. This is likely due to Huawei devices being specifically targeted by several botnets (Wijesinghe, et al., 2015, Antonakakis, et al., 2017). According to GlobalStats (2019), Huawei control 25.82% of South Africa’s mobile device market share, see Figure 2.

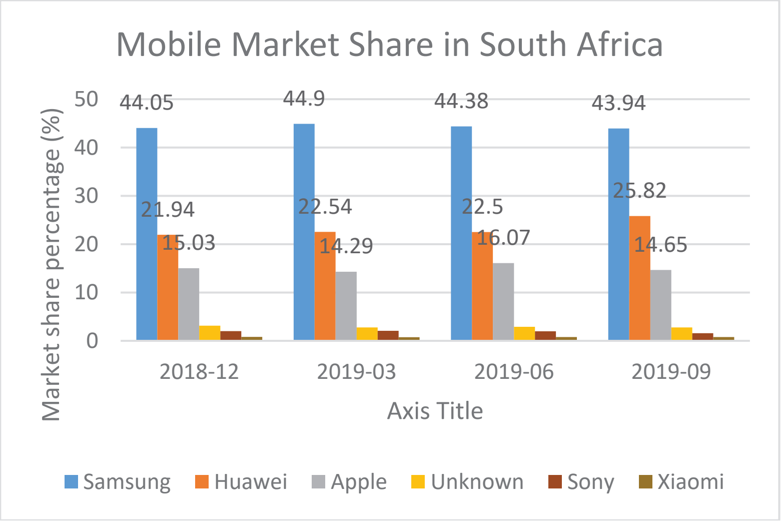


Figure 2: South African mobile market share

Since the start of this research several notable botnet families were observed in the wild. Since the SANREN CSIRT currently only captures IPFIX data we were unable to directly map the cellular operating system to the botnets observed during these attacks. However due to the large market share of Huawei devices and a number of botnets specifically targeting these devices the research conducted within the SANREN CSIRT focused on botnets which are known to target these devices. In this section a brief overview of the main botnet families will be provided.

3.1 Mirai Botnet

The earliest traces of the Mirai botnet was observed in August 2016. The Mirai botnet spreads like a network worm, it primarily targeted Internet of Things (IoT) devices. The most notable attacks performed by the Mirai botnet is the Distributed Denial of Service (DDoS) attack against OVH and Dyn.

The botnet, originally infected new hosts by performing TCP SYN probes on TCP port 23 and TCP 2323 (Telnet). If the target responded to the TCP SYN request, the botnet would enter phase 2 of the infection step by performing a brute force login attack on the service. The Mirai botnet had a list of 62-preconfigired credentials which were known default credentials for vulnerable IoT devices. Once the bot successfully logged into the device a control message would be sent to a Mirai C2 server and a target specific malware payload would be installed on the newly infected device. The newly infected device would then start then start seeking out new devices to infect. Antonakakis, et al., (2017), identified 112 C2 domains and 92 IP addresses related to the Mirai botnet by using active and passive DNS.

3.2 Satori Botnet

Over time Mirai expanded its infection pattern by targeting TCP port 22 (Secure Shell, SSH), 80/8080/443 (web interface, HTTP/S), 7547 (CPE WAN Management Protocol, CWMP), 5555 (Android Debugging Bridge, ADB). The Satori strain, of the Mirai botnet family, exploited CVE-2014-8361 and CVE-2017-17215 via ports 37215 and 52869 to exploit Huawei HG532e home routers without relying on the brute-force attack phase (Anubhav, 2018). In July 2018, it was reported Satori targeted Android devices by sending a malicious payload to the ADB using port 5555(Pour, et al., 2019).

3.3 GoldBrute Botnet

The GoldBrute botnet performs a brute-force attack on servers which have TCP port 3389 (Remote Desktop Protocol, RDP) exposed to the internet (Marinho, 2019). In June 2019, it was reported that the GoldBrute botnet is performing an active attack on over 1.5 million RDP devices. According to Shodan, there are 3.8 million devices (27 589 of which are in South Africa) currently connected on the internet which expose TCP port 3389.

This botnet was controlled via a single C2 server, 104.156.249.231. Bots connect to the server over TCP port 8333. Once the botnet successfully brute-forces a RDP server, the server is instructed to download the botnet payload from the following IP 104.248.167.144. The payload is an 80 megabyte Java application (Marinho, 2019). After the application is downloaded the newly infected server will form part of the botnet and seek out other RDP servers to infect. The GoldBrute C2 server was discovered and taken down in July of 2019.

4. Network flow data collection within SA NREN

The network flow data is captured within the IPFIX data format. IPFIX is a vendor neutral format for storing network flow data, formerly known as NetFlow. The network flow data contains only the meta-data related to a network connection stream, with none of the packet content. At the bare minimum a network flow can be captured as a 6-tuple of information elements: source IP address, destination IP address, source port, destination port, protocol used and time-stamp. IPFIX is based on of NetFlow version 9 which uses 13-tuple to represent a network flow. Each network connection is usually represented by its own flow. These flows are exported to the IPFIX format by an IPFIX export device (usually routers or switches). In order to centralise the analysis process an IPFIX collector is used to fetch all IPFIX logs to a centralised server. Within the SA NREN there are ten primary IPFIX exporters which are used for our analysis. Each exporter is external to the research institutions and each collector collects the data from various institutions. Most of the IPFIX exporters export the flows at a 1:1 ration, however some of the exporters only export a sample of the observed flows. Table 1 shows a list of network flow exporters within the SA NREN as well as their average flows. Each exporter exports well over a billion flows per day (These were the averages observed during the period 1 May 2019 till 31December 2019). Processing such a large volume of data does present a significant challenge to the SANREN CSIRT.

Table 1: Network flow exporters within SA NREN

Device ID	Ratio	Average flows per day
AMS1-IR1	1:1	2.44 x 10^{10}
CPT1_IR1	1:1	2.66 x 10^{10}
CPT1_PE2	1:50	6.37 x 10^{9}
CPT3_PE1	1:50	1.64 x 10^{9}
CPT3_PTA1	1:50	2.14 x 10^{9}
ISD1_PE1	1:50	1.89 x 10^{9}
JNB1_PE1	1:1	8.07 x 10^{9}
LDN1-IR1	1:1	1.83 x 10^{10}
MTZ1_PE1	1:1	1.62 x 10^{10}
PTA1_PE2	1:1	4.94 x 10^{9}

On average 1 terabyte of network flow data is captured each month. Due to the large volume of data to be processed a detection mechanism which is capable of scaling to the needs of the SA NREN is required.

The SANReN CSRIT has an additional constraint in that the network flow data is captured on the SA NREN backbone infrastructure rather than within the specific research institutions themselves. This introduces constrains such an aggregated flows whereby network traffic flow through network aggregation points such as institutional firewall or Network Address Translation (NAT) devices. If flows are aggregated the true source or destination of a network flow may be obfuscated from the network flow logs.

During our initial investigations it was observed that some institutions inject additional packet overhead into their network flow data. This data tends to be related to parity checks, GRE or NAT information. These alterations to the true packet size may influence the detection of malicious traffic if traffic size is used as an indicator.

Another constraint placed on the SANReN CSIRT is that the NREN infrastructure is owned by the constituents and that the CSIRT can only provide advisories to the constituents. The CSIRT has no authority to take down any misbehaving nodes or block connectivity to potentially malicious nodes. In extreme cases TENET can be tasked with limiting network access but this is only as a last resort.

5. Applying research to practise

From the SANReN CSIRT's perspective there are four main activities which are monitored using IPFIX data:

- Port scans detection: Port scans are very prevalent on large scale networks and are not of critical interest to TENET, from a threat perspective. The greater concern is trends or sudden spikes in specific ports being scanned.
- Brute-force login attack detection: Brute-force login attempts fall into a similar category of concern as port scans, in that they are prevalent on the SA NREN. However, brute-force attacks pose a significantly higher risk than port scans if a system is compromised. From TENET's perspective detecting and preventing brute-force attacks aimed at key backbone infrastructure nodes are the highest priority for detection.
- Distributed Denial of Service Attack detection: The IPFIX is used to identify potentially malicious nodes inside the SA NREN which participate in DDoS activities. Due to the meta data capture using network flows volumetric attacks are easily detectable. The SANReN CSIRT is more interested in protocol specific attacks. Specifically Distributed Reflection Denial of Service (DRDoS) attacks which uses SA NREN infrastructure to launch attacks against external services or service providers.
- Botnet detection: From the SANReN CSIRT perspective the team is interested in detecting and monitoring botnet activity at a national level, including the spread of infection within tertiary institutions. TENET is concerned about their infrastructure being targeted by a large scale botnet attack or inadvertently hosting bots on the tertiary infrastructure.

This paper focus specifically on the botnet detection component of the SANREN CSIRT efforts, however the detection efforts rely substantially on the prior work done to detect other threats to the SA NREN.

TENET is responsible for AS2018 (https://ipinfo.io/AS2018), as The SANReN CSIRT needs to monitor several large blocks of IP addresses. In this section several incidences which occurred on these network ranges will be discussed. In order to mask the various institutions identities the IP blocks have been mapped to the private IP address space 10.0.0.0/24.

The SANReN CSIRT does not rely on a single algorithm to detect abusive behaviour or cyber-attacks. Instead, the CSIRT rely on a combination of several algorithms to identify potentially malicious IP addresses. The CSIRT maintains a list of IPs of Interest (IPoI) which are tagged with labels, describing their potential malicious behaviour. The idea of tagging of IPs based on their behaviour was first proposed by Sweeney & Irwin (2017). For example if an IP is observed scanning port 22 (SSH) that IP will be tagged in the database with the tag “Port scan - 22”. IPs within the database can be assigned multiple labels since it is likely that a misbehaving node could potentially perform multiple suspicious activities.

The SANReN CSIRT uses the IPFIX port scan signatures proposed by Grégr (2010) to classify port scan activity based on TCP flag combinations

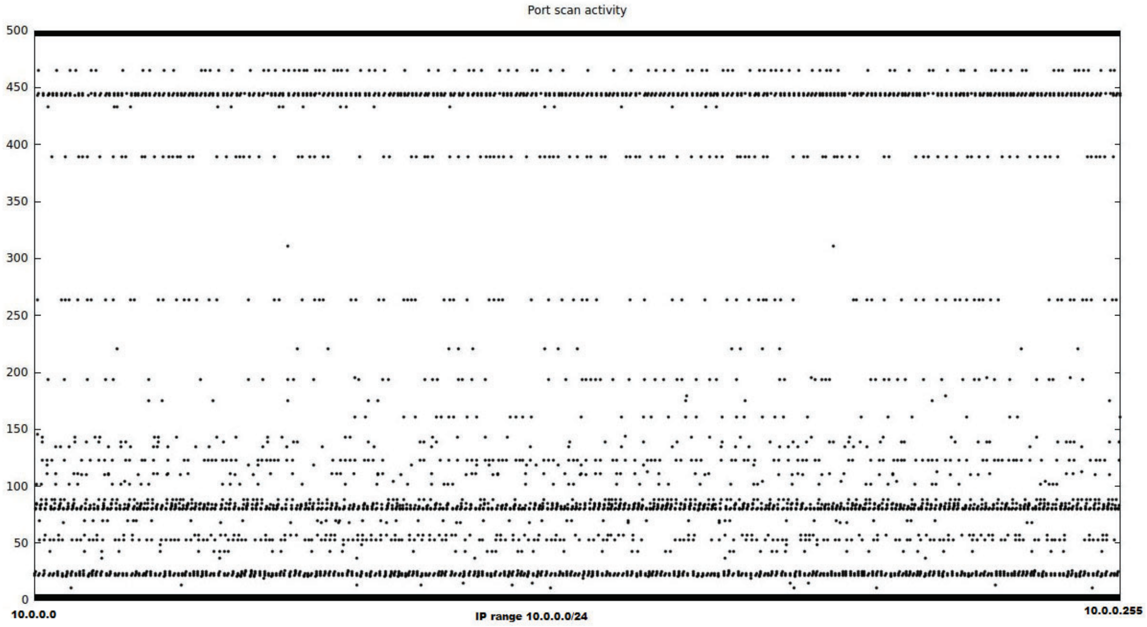


Figure 3: Stealthy port scan activity observed over /24 network range

Figure 3 depicts a scatter diagram of an IPoI performing a SYN scan on a /24 network range within the SA NREN. The y-axis represents the ports being scanned and the x-axis represents the IP address range (which has been mapped to the range 10.0.0.0/24 for privacy concerns). The data depicted in **Figure 3** is from a single IPoI over a period of one week. The IPoI was assigned labels for scanning ports 21-23, 51, 80-84, 143. 189 and 443. Based on the pattern observed this particular port scanner was performing a horizontal scan targeting specific ports over the network range. The port scan activity depicted in **Figure 3** was observed over a 48 day period. During the period the scanner would scan on average three ports on 16 hosts spread across the /24 IP address space. These scans were so slow and distributed that the organisational firewalls and Intrusion Detection Systems (IDS) did not detect any abnormal activity.

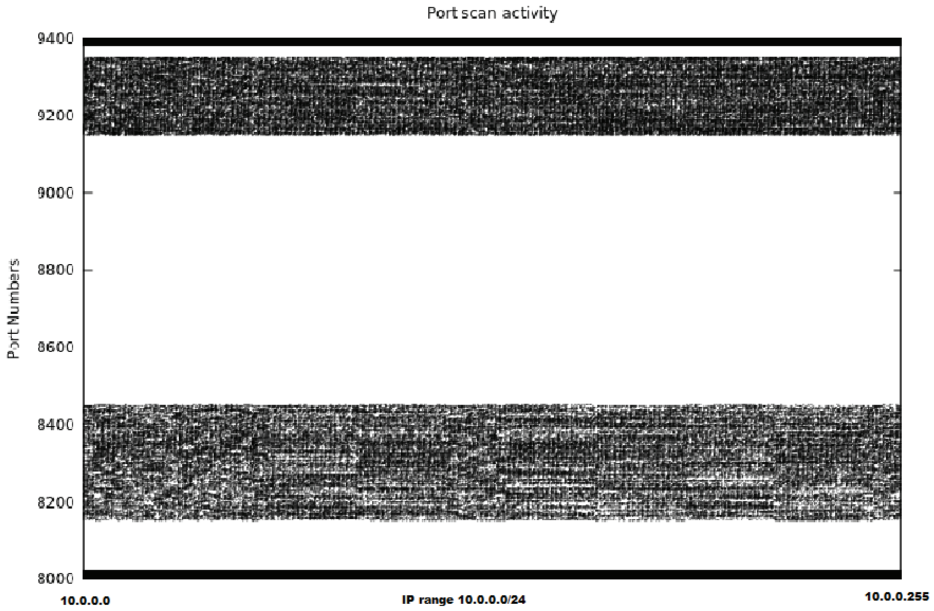


Figure 4: Coordinated block scan activity

In contrast Figure 4 depicts an observation of an aggressive block scan activity. Based on the IPFIX data, the scanner only probed ports which were either between 8151 and 8450 or 9151 and 9350. This specific scanning activity, depicted in Figure 4, formed part of a coordinated scanning effort against one of the tertiary institutions within South Africa. During a 36 hour period, 9 source IPs originating from an internet service provider located in the Netherlands, were performing block scans against the tertiary network. The 9 source IPs never overlapped scanned port blocks, each had its own dedicated block it was scanning. On average the scanning IPs performed 173 probes per host every 18900 seconds (exactly 5 hours and 15 minutes). The activities observed in Figure 3 and Figure 4 resulted in the source IPs being added to the IPoI database.

The brute-force attack detection algorithms within the SANReN CSIRT is based off of work done by Sperotto et al. (2009) and Hellemons et al. (2012).

Sperotto et al. (2009) suggested that the number of packets observed per network flow (ppf) could be an indicator for the attack phases of a SSH brute-force login attack. By observing a time series of ppf observed during an attack Sperotto et al. (2009) was able to delineate between the pre-attack phase (scanning phase), attack phase and the post-attack phase (residue phase).

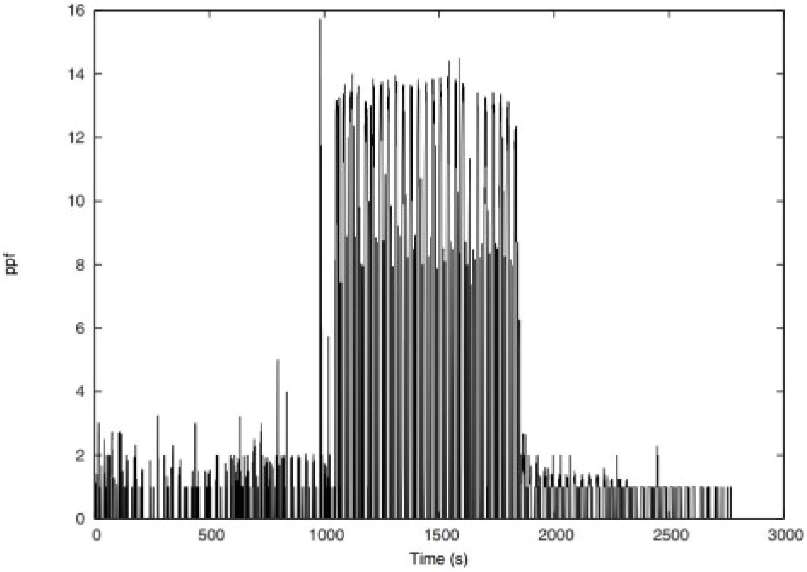


Figure 5: Packets per flow observed during brute force SSH attack Sperotto et al. (2009)

The experimental data used by Sperotto et al. is shown in Figure 5. The scanning phase the majority of ppf measurements are at most 2 and during the brute force attack phase the ppf value fluctuates between 8 and 14 ppf. Sperotto et al. (2009) also noted that between each active brute force attempt a small period of inactivity was observed. By using the data collected during the attack Sperotto et al. (2009) derived the hidden state sequence of the Markov Model. Then by using the hidden sequence the researchers were able to calculate the transition probability between Markov states.

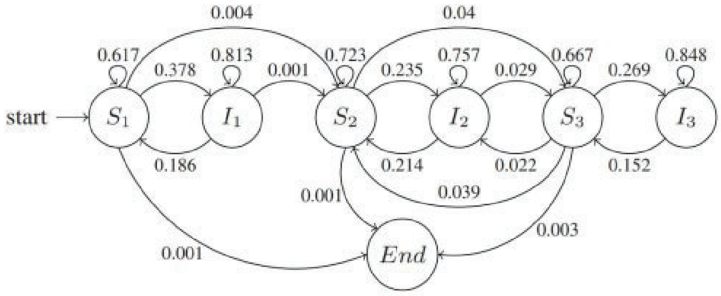


Figure 6: Hidden Markov Model proposed by Sperotto et al. (2009)

The Hidden Markov Model (HMM) to detect SSH attacks, proposed by Sperotto et al. (2009), is shown in Figure 6. S_1, S_2, S_3 represents the three phases of the brute force attack as defined by Hellemons et al. (2012). I_1, I_2, I_3 represents the state of inactivity as observed by Sperotto et al. (2009). It is important to note that based on the proposed model S_1, S_2, S_3 can all proceed to the End state at any point. This implies that an attacker may stop during the attack.

The probabilities shown within the HMM is specific to the attack observed by Sperotto et al. but a similar experimental design was used to define a more generic HMM used by the SANReN CSIRT. The work done by Sperotto et al. (2009) and Hellemons et al. (2012), was also further expanded upon by the SANReN CSIRT to include brute-force attack models for FTP and RDP brute-force attacks.

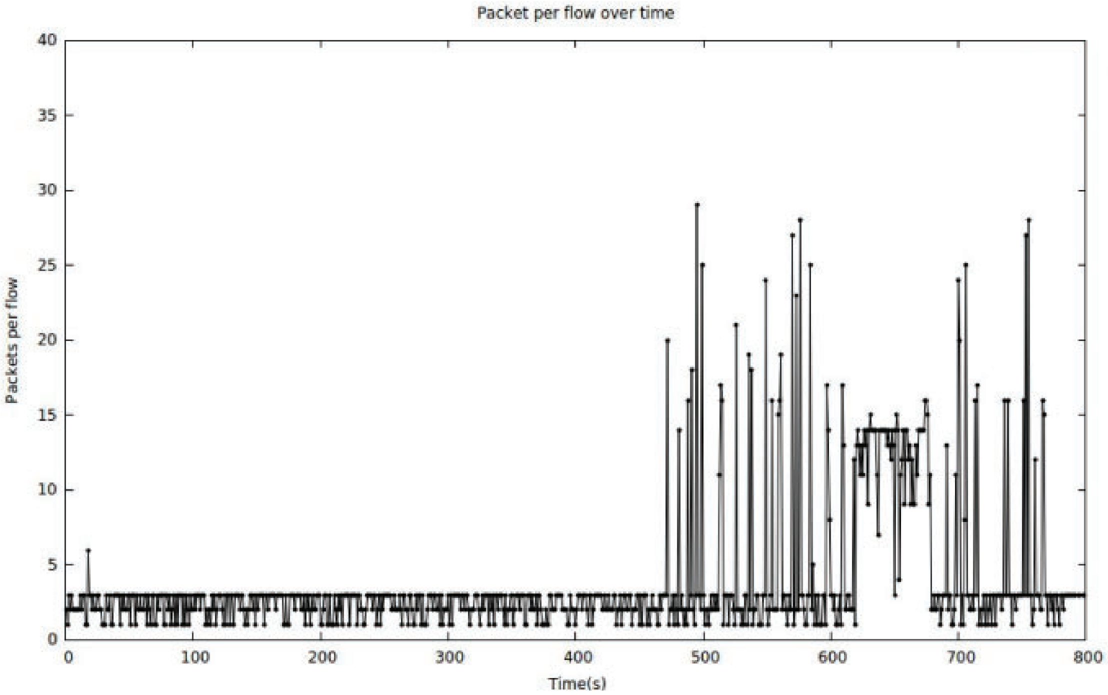
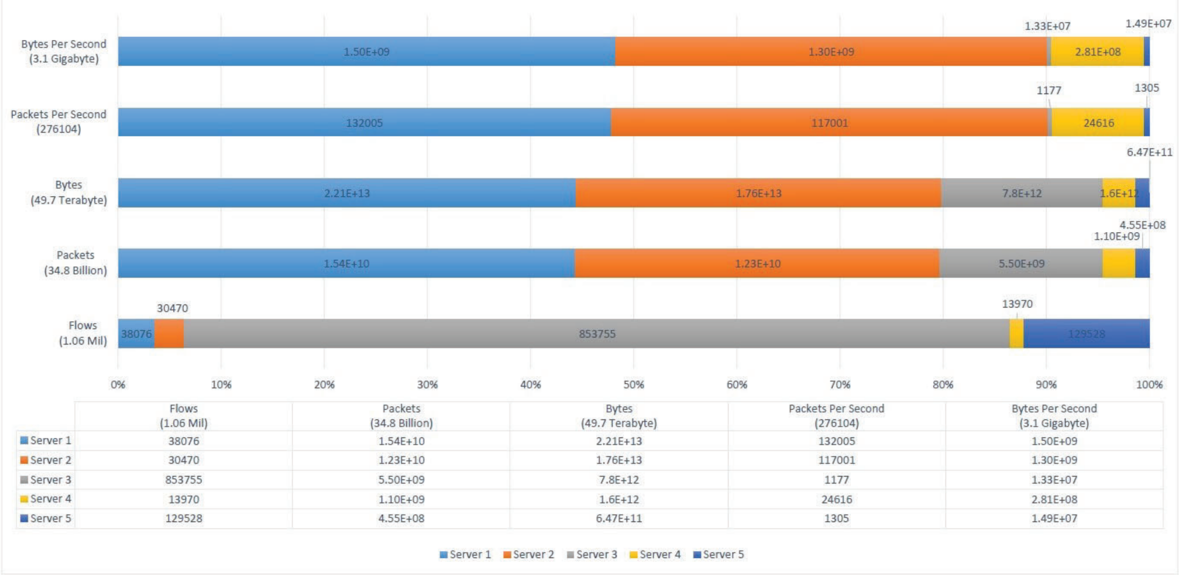


Figure 7: FTP brute-force attack detection within SA NREN

Figure 7 depicts a time series observation of an attempted FTP brute-force login attack against TENET infrastructure components hosted within the SA NREN. The attack pattern is similar to the one observed in Figure 5 by Sperotto et al. During the sensor testing conducted on the NREN network it was determined that fixed using ppf counts as was suggested by Sperotto et al. is not sufficient. Because the SANReN IPFIX collectors are on the TENET backbone the ppf values tended to higher than what was suggested by Sperotto et al. Upon investigation it was determined that security equipment on the SA NREN added additional overhead to the attacker's network flows and as a result the SANReN CSIRT observed higher ppf values. Thus it is recommended that thorough testing and verification of sensor labelling be conducted rather than using the values suggested by Sperotto et al. directly. In order to detect coordinated attacks port scan activity was also considered in our HMM. The HMM thus also needed to be updated to account for attack models which start in S_2 as depicted in Figure 6.

On 28 February 2018, the world's largest Distributed Reflective Denial of Service attack was launched using Memcached servers (Burke et al., 2018). During the attack five SA NREN servers were used to inadvertently attack remote servers. The key findings of the attack using SA NREN is discussed within the SANReN CSIRT's previous publication (Burke et al., 2018), in this paper only the key findings will be highlighted.



	Flows (1.06 Mil)	Packets (34.8 Billion)	Bytes (49.7 Terabyte)	Packets Per Second (276104)	Bytes Per Second (3.1 Gigabyte)
Server 1	38076	1.54E+10	2.21E+13	132005	1.50E+09
Server 2	30470	1.23E+10	1.76E+13	117001	1.30E+09
Server 3	853755	5.50E+09	7.8E+12	1177	1.33E+07
Server 4	13970	1.10E+09	1.6E+12	24616	2.81E+08
Server 5	129528	4.55E+08	6.47E+11	1305	1.49E+07

Figure 8: Summary of SA NREN DRDoS attack, February 2018

During the attack five SA NREN servers were affected by the reflective attack. A summary of the traffic generated between 23 February 2018 and 8 March 2018 is presented in Figure 8. The attack was the first and largest attack detected to date using the IPFIX sensors within the SA NREN. After the attack the following rules were applied to the IPs detected during the attack to label the IPs with the IPoI database:

- If the IP address belongs to a known legitimate user of Memcached service, label the IP as benign. If these benign services form part of an attack, the operators of the services was notified.
- If the number of Memcached requests issued per day increased by more than 50% after 21 February, the IP was labelled as suspicious.
- If the ratio of request traffic to response traffic generated by the IP was greater than 1:100, the IP was labelled as a victim.
- If any of the other network flow sensors, deployed by the SANReN CSIRT, detected the IP as a port scanner, the IP was labelled as a port scanner.
- If any of the port scanner IP addresses were found to only be scanning for port 11211 and using the Memcached protocol, it was assigned an additional label, memcached scanner.

The rule set described above formed the bases for future DDoS attack patterns within the SA NREN.

By combining the observations obtained from the port scan detection algorithms, the brute-force detection filters and the DDoS detection rule sets the SA NREN was able to collect a corpus of label IPoI data. This data was used to train Machine Learning (ML) algorithms to detect potential botnets within the SANReN.

To further enhance the accuracy of the ML models it was supplemented with behavioural models based on patterns observed by common large botnets. For example, as was mentioned in Section 3.2, the Satori botnet Android IoT devices by delivering a malicious payload of ADB (port 5555) in July 2018 (Pour, et al., 2019). Based on the results obtained from the IPoI database it was determined that similar attack behaviour was observed within the SA NREN as early as 5 April 2018.

The SANReN CSIRT built a botnet detection model based on the GoldBrute botnet attack pattern discussed in Section 3.3 and was able to detect similar attack traffic on 11, 15 and 22 September 2019. This despite the GoldBrute C2 servers being taken down in July 2019. The newly observed C2 servers observed to be hosted in Denmark, France, Brazil and Russia.

The SANReN CSIR is currently tracking several variant of Mirai botnets which target IoT devices within the SA NREN. Based on the ports targeted by the botnets and the characteristics of the brute-force attacks used by the bots the ML algorithms are able to cluster the bots into distinct families.

6. Conclusion and future work

In this paper the current efforts of the SANReN CSIRT to detect botnets are discussed. The CSIRT uses a combination of attack sensors to label IPs based on their attack behaviour. This allows the CSIRT to track and label suspicious activity within the SA NREN. At present the majority of the research and detection is inward focused to protect the South African infrastructure. However in future the SANReN CSIR plans to make the IPoI database and label publically accessible for other NRENs to view and contribute.

The botnet components of the SANReN IPFIX attack sensors are still reliant on attack models being built based on known botnet attack models and behavioural analysis. Current research is being conducted within the SANReN CSIRT to develop more generic sensors to detect unknown attack models.

References

Amini, P., Azmi, R. and Araghizadeh, M., 2014. Botnet detection using NetFlow and clustering. Advances in Computer Science: an International Journal, 3(2), pp.139-149.

Antonakakis, M. et al., 2017. Understanding the mirai botnet. In: *26th USENIX Security Symposium USENIX Security 17.*, pp. 1093-1110.

Anubhav, A., 2018. *Masuta: Satori Creators' Second Botnet Weaponizes A New Router Exploit.* [Online] Available at: https://blog.newskysecurity.com/masuta-satori-creators-second-botnet-weaponizes-a-new-router-exploit-2ddc51cc52a7
[Accessed 5 September 2019].

Bilge, L. et al., 2012. Disclosure: detecting botnet command and control servers through large-scale netflow analysis. *Proceedings of the 28th Annual Computer Security Applications Conference, ACM,* pp. 129-138.

Burke, I.D., Herbert, A. and Mooi, R., 2018, September. Using network flow data to analyse DRDoS attacks, as observed on the SANReN. In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT), pp. 164-170.

Erquiaga, M. J., Catania, C. & García, S., 2016. Detecting DGA malware traffic through behavioral models. *IEEE Biennial Congress of Argentina (ARGENCON),* pp. 1-6.

Feily, M., Shahrestani, A. & Ramadass, S., 2009. A survey of botnet and botnet detection. *Third International Conference on Emerging Security Information, Systems and Technologies,* pp. 268-273.

GlobalStats, 2019. *Mobile Vendor Market Share South Africa - Oct 2018 - Oct 2019.* [Online] Available at: https://gs.statcounter.com/vendor-market-share/mobile/south-africa
[Accessed 23 October 2019].

Grégr, M. Portscan detection using netflow data. Proceedings of EEICT10, pp. 229– 233, 2010.

Grill, M., Nikolaev, I., Valeros, V. & Rehak, M., 2015. Detecting DGA malware using NetFlow. *IFIP/IEEE International Symposium on Integrated Network Management,* pp. 1304-1309.

Hellemons, L., Hendriks, L., Hofstede, R., Sperotto, A., Sadre, R., and Pras, A. Sshcure: a flow-based ssh intrusion detection system. In IFIP International Conference on Autonomous Infrastructure, Management and Security, pp. 86–97. Springer, 2012.

Hofstede, R., Bartos, V., Sperotto, A. & Pras, A., 2013. Towards real-time intrusion detection for NetFlow and IPFIX. *Proceedings of the 9th International Conference on Network and Service Management,* pp. 227-234.

Hutchins, M., 2017. *Investigating Command and Control Infrastructure (Emotet).* [Online] Available at: https://www.malwaretech.com/2017/11/investigating-command-and-control-infrastructure-emotet.html
[Accessed 5 October 2019].

Marinho, R., 2019. *GoldBrute Botnet Brute Forcing 1.5 Million RDP Servers.* [Online] Available at: https://isc.sans.edu/forums/diary/GoldBrute+Botnet+Brute+Forcing+15+Million+RDP+Servers/25002/
[Accessed 30 October 2019].

Mazzariello, C., 2008. IRC traffic analysis for botnet detection. *The Fourth International Conference on Information Assurance and Security,* pp. 318-323.

Pour, M. S. et al., 2019. Data-driven Curation, Learning and Analysis for Inferring Evolving IoT Botnets in the Wild. *Proceedings of the 14th International Conference on Availability, Reliability and Security,* pp. 16-26.

Sweeney, M. & Irwin, B., 2017. A NetFlow Scoring Framework For Incident Detection. *Southern Africa Telecommunication Networks and Applications Conference (SATNAC) ,* pp. 300-305.

Wagner, C., François, J., State, R. & Engel, T., 2011. Machine learning approach for IP-flow record anomaly detection. *International Conference on Research in Networking,* pp. 28-39.

Wijesinghe, U., Tupakula, U. & Varadharajan, V., 2015. An enhanced model for network flow based botnet detection. *Proceedings of the 38th australasian computer science conference ,* Volume 27, pp. 101-110.

Fostering Cybersecurity Awareness Among Computing Science Undergraduate Students: Motivating by Example

Tiago Cruz and Paulo Simões
University of Coimbra, Portugal, CISUC, DEI
tjcruz@dei.uc.pt
psimoes@dei.uc.pt

DOI: 10.34190/EWS.20.040

Abstract: To a certain extent, it can be considered that cybersecurity has a PR problem. Despite all the campaigns and publicity surrounding the matter, there are certain age groups for whom the message doesn't go through, either because it sounds too condescending or too technical. This is more of a problem for undergraduate students in computer science courses, especially if we take into account the potential academic and career development paths they may pursue, with many becoming specialized practitioners for areas such as software engineering or information systems design, despite lacking a solid background on cybersecurity and good development practices. Prior experience from the authors has shown that a purely theoretical approach to teaching the fundamental concepts of cybersecurity has several shortcomings, contributing to distance the students from the subject and, in some extreme cases, to foster rejection towards the topic. This is partly due to the fact that the involved formal and theoretical aspects can be quite demanding, especially in terms of subject diversity, involving concepts about operating system design, networking or computer architecture, among others. Presenting this information in an effective way requires taking into account the current level of knowledge expected from the students at a certain point of their academic path, as well as their level of maturity. Engaging the students is a key aspect that has to be central to any strategy. As such, this paper presents an approach to the subject of cybersecurity education at an introductory level, focused on a methodology that has been recently adopted by the authors, with a considerable degree of success. It is based on a *hands-on* strategy, focused on the vulnerability analysis of a commercial, off-the-shelf consumer IoT device (namely, an IP camera). Starting with a presentation of the Mirai botnet incidents, the authors proceed with a step-by-step dissection of the target device to identify its vulnerabilities and showcase their criticality.

Keywords: Home Network Security, Distributed Security, Smart Homes, Home appliances.

1. Introduction

Despite the fact we have come a long way in terms of cybersecurity, here is still a lot of work to be done, as IT, IoT, SCADA and Smart City environments pose new challenges, as well as the ever-evolving landscape of classic threats and attack vectors. As modern warfare moves to the cyber domain, the demand for well-trained professionals is on the rise (ISACA, 2020), with the industry requiring more specialists than the educational establishment can train and prepare - including cybersecurity academies and other specialized training organizations, corporate training, universities or even the military, in some cases.

This situation is worsened by the fact that, for most ICT undergraduate students, cybersecurity is a specialization field heavily competing with other specialization, such as data science or software engineering. This issue can only be adequately tackled by means of a differentiated educational offer, both in terms of quality and didactic approach, focused on recruiting, engaging and preparing students to become capable professionals with a mindset geared towards excellence.

Also, and regardless of any interests they might pursue, there is also the need to foster cybersecurity awareness among students. Not necessarily to kickstart the interest on pursuing a security-oriented career path, but rather because the topic is orthogonal to most specialization profiles. For instance, software engineering projects often forego cybersecurity aspects to later development stages, instead of addressing them in the initial requirement specifications. Also, and unlike what happens with the group immunity effect, there is no guarantee that having a sufficiently large number of cybersecurity-aware individuals may be enough to avoid an incident within an organization (quite often, phishing attacks only need to deceive a single victim to be successful).

The authors believe that the investment in favour of the modernization and promotion of good pedagogical practices can be one of the potentially differentiating aspects of a University as a school for advanced studies. This effort should not be limited to the implementation of quality management policies based on surveys and the adoption of supervisory methodologies that, despite their relevance, often have little more effect than providing evidence to reinforce a diagnosis that has long been known: traditional teaching methods are often inadequate when it comes to motivating students.

At a time when online content abounds, it is up to the teacher to provide differentiation, at least when compared to these resources and contents that, despite being convenient, are often passive (as is the case with some video content). While there is no universal solution to the problem, the commitment to a logic of complementarity and proximity has proven to bring good results. The main motivation behind this approach is based on the idea of promoting the proximity and accessibility of the teacher as an added value in view of the characteristic distance from asynchronous online learning models.

Moreover, there has been an increasing prevalence of the so-called maturity deficit, often pointed out as one of the main causes of dropout, indiscipline and failure among undergraduate students. In fact, there are recent studies (Sawyer et al., 2018), often reinforced by the widespread opinion of colleagues, which point to an extension of the period covered by adolescence. From the teacher's point of view, this situation is based on the perception that there is a generalized instability at one or more levels of the students' behaviour and/or personality, such as emotional, autonomy, attention/concentration capacity, diligence and organization. However, the same students often have a very high hidden potential for creativity, often repressed by factors prior to entering the University, but which may tend to worsen when they access higher education.

Over the last four years of pedagogical practice, the solution adopted by the authors has been to identify a set of clues that allow characterizing the most frequent imbalances in the groups of students, in order to reinforce the specific teaching strategies. By acknowledging that learning is, by nature, a social process, the instructors implement selective scaffolding procedures in order to make content accessible, while focusing on providing a quick route enabling students to become involved in *hands on* activities as quickly as possible. When used in the context of a course, the authors often prefer to encourage students to work with greater autonomy in terms of choosing project topics, planning strategies and development methodologies, with the counterpart of a greater demand in terms of the quality of the final result to be achieved.

This paper documents an example of a training module/class plan, focused on providing insights on IoT security and, more specifically, on the nature of a Mirai-like botnet (Venkatesan, 2018). It was designed to be easily adjustable to students from basic-to-intermediate skill level, mostly from second year undergrads onwards, in order to be applied in contexts which range from one-shot security awareness dissemination actions to introductory classes in postgraduate security specialization courses.

The remaining of this paper is organized as follows: the next section presents a synthetic state-of-the-art overview on didactic approaches and testbed designs conceived for cybersecurity training purposes, followed by the presentation of the motivation and approach for the training module. Next, the training module structure will be outlined and executed, step-by-step, with the final chapter taking care of presenting the relevant conclusions and results obtained from the proposed approach.

2. Cybersecurity training: a brief overview

This section provides an overview of several related efforts on the topic of cybersecurity didactics, as well as testbed designs which may be of interest in the context of the objectives hereby outlined – nevertheless, it should be noted that this overview is not, by any means, an exhaustive analysis about the matter.

Several universities and education/training organizations are incorporating specialized cybersecurity modules in their curricula, covering aspects such as network security, cyber threat intelligence or risk assessment, among others. In some cases, cyber security training and simulation platforms are used to support learning activities, as it is the case for cyber-ranges, providing students with the opportunity of becoming involved in realistic scenarios (Hallaq et al., 2018), assuming specific roles accordingly to the nature of the exercises (as it is the case for red/blue team drills). This is the case, for instance, for the cybersecurity curricula taught at the De Montfort (UK) (Maglaras et al., 2020) and Coimbra (PT) (Frazão et al., 2018) Universities.

Considering other approaches towards teaching cybersecurity-related topics, (Trabelsi and Saleous, 2018) proposed an approach for introducing fundamental concepts such as network keylogging and eavesdropping attacks. The authors described a group-based approach in which students have access to machines with the CommView tool installed for network analysis and packet sniffing purposes. Regarding security auditing techniques, (Zseby et al., 2015) presented an approach for teaching network traffic anomaly detection techniques using an IP darkspace monitor, providing support to a series of exercises to be solved using the

MATLAB, tcpdump, corsaro, and RapidMiner tools. They also introduced a concept of network security laboratory designed for educational purposes.

Regarding purpose-specific testbeds, (Eliot et al., 2018) present the design of a flexible laboratory for teaching cybersecurity skills, where scenarios can be easily customizable. Such a honeypot-based laboratory was also designed with hackathon events in mind, being used at the Northumbria University (UK) to support undergraduate and postgraduate activities. Other competitive approaches, such as the NetSecLab proposed by (Lee et al., 2010) have also been used for teaching systems and network security, enabling students to acquire the basic skills required to deal with cybersecurity concepts and issues.

In (Teixeira et al., 2018), the development of a SCADA testbed (water tank storage control system) for cybersecurity research was presented. This testbed was designed to provide the means to execute cyberattacks both for impact analysis and dataset generation, the latter being used for studying the effectiveness of using machine-learning (ML) algorithms for intrusion detection. Despite the research-oriented focus, the testbed design is also promising for educational purposes. The same rationale may be equally applied to the Hybrid Environment for Development and Validation (HEDvA) testbed design presented by (Cruz et al., 2017), as it provides a convenient blueprint for implementing a hybrid testbed using a mix of physical equipment, simulation models, emulated components and virtual machines – this approach provides a convenient way to implement virtual labs by selecting and customizing components from inventory pools.

Among the proposals which were reviewed, the HEDvA was probably the most influential, inspiring several architectural decisions affecting the design of the laboratory testbed supporting the training module hereby presented, which will be presented into detail in the next section.

3. Motivation and high-level approach

This section provides an outline of the action plan that is followed to introduce the subject of IoT security, as well as the reasons for its relevance. First, students are introduced to the specific nature of IoT technologies – while it may be needless to provide a rationale for their popularity, instructors must take care of addressing the knowledge gaps that may exist within heterogeneous audiences.

Along other use cases and incidents, the Mirai botnet (ISACA, 2018) is presented, as it provides a good example of the potential consequences of IoT insecurity in a real-world situation. One of the most visible incidents involving Mirai was the DDoS attack that targeted the *Krebs on Security* site on September 20, 2016. This attack reached 620 Gbit/s (a 1TBbit/s attack also targeted the OVH hosting provider). Mirai was also used to deploy several DDoS attacks against DNS services of a DNS provider, on October 21, 2016, resulting in the unavailability of high profiles websites such as GitHub or Twitter, among others.

Synthetically, the Mirai botnet operation follows a simple intrusion workflow that can be mimicked in an approximate way. Following this perspective, the *hands-on* part of the lesson plan was designed to showcase, in an interactive and manual basis, how the infection procedure could have been undertaken (this is further true if we consider that Mirai infections often start by firing a shell script on a vulnerable device which downloads and executes the infectious payload).

The training module/lesson hereby documented takes place (*in situ* or by remote access) in a laboratory equipped with a testbed environment, designed for educational purposes, in which the IoT camera was deployed. Students are provided access by means of Virtual Private Network (VPN) connections to a set of Virtual Machines, preloaded with Kali Linux (see Figure 1). This setup provides students with equal resources and rights regarding testbed access, while shielding the environment from the Campus LAN. Each VM is configured with 2 virtual cores, 2GB RAM, two network interfaces and 100GB of virtual persistent storage. Both hypervisor nodes have the same configuration. The testbed network encompasses the 172.27.224.0/24 range of IP addresses.

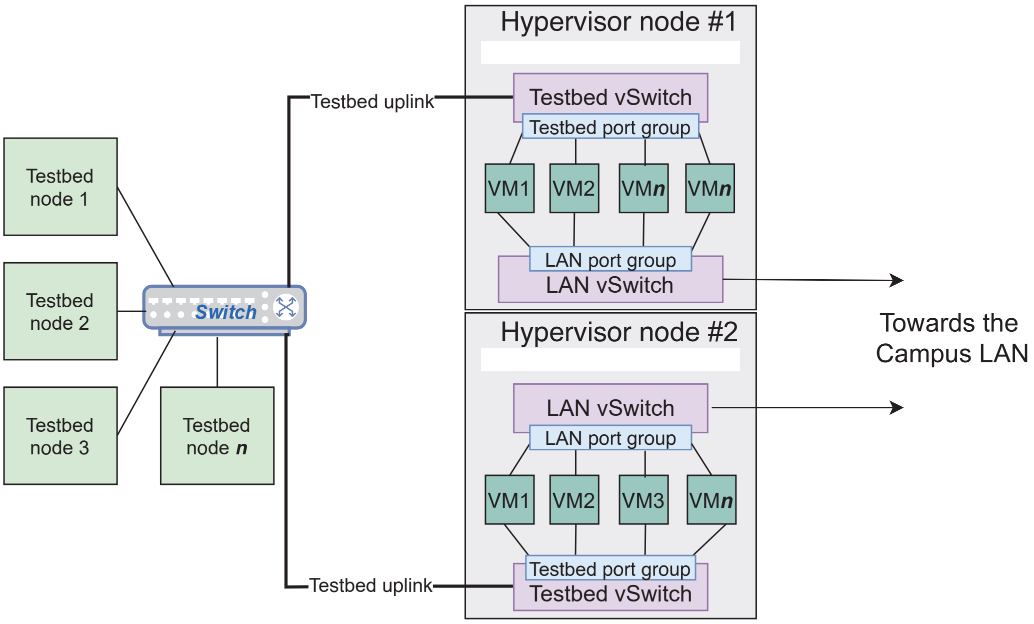


Figure 1: Testbed organization

As it has already been mentioned, the testbed architecture has the added benefit of isolating the testbed from the campus network. This means that cybersecurity-related training activities such as scouting and experimentation, are kept away from the production networks. More than convenience reasons, this option has to do with the potential impact of such activities, which range from service exhaustion to unauthorized traffic steering, as well as unlawful data interception, eventually exposing sensitive activity evidence or information.

The basic module/lesson structure is organized as such:

- Step 1 - Scouting procedures: presentation of the toolset to be used, as well as the techniques to be deployed in the field;
- Step 2 - IoT/target device analysis: this step will attempt to identify the hardware architecture and resources available on the device;
- Step 3 – Exploitation: identification of the most interesting vulnerability vectors and discussion of how they can be exploited and leveraged to gain control of the device.

This strategy intends to enhance the students' knowledge on network security, while also providing them with insights on exploitable intrusion vectors, good development practices and proper handling of security issues.

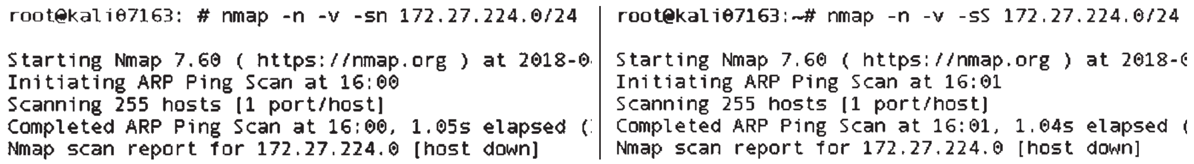
4. Execution of the training module plan

This section will detail the various steps undertaken within the training module/lesson, also presenting the expected outcomes for each one. Moreover, specific hints regarding the didactic strategies adopted for each step are also provided, based on the authors' experience.

4.1 Initial scouting/reconnaissance procedures

Network/range scouting is often one of the first steps of a Mirai-like infection process. From a more generic perspective, this step is required to gather information about all the components of the target environment, to discover and identify topologies, hosts and services. For instance, a host may be profiled using information about its MAC address (usually only possible when the attacker is in the same Layer 2 domain), OS versions or TCP/IP open ports, using techniques such as SYN of FIN scans, together with OS fingerprinting. Combined with Open-Source Intelligence (OSINT) techniques, the information obtained from these sources can be further enriched with additional elements about manufacturers and models, as well as firmware versions or known vulnerabilities. For the network scanning steps, the NMAP (Lyon, 1997) tool was used. NMAP is a well-maintained open-source and portable tool providing a complete feature set and a powerful command-line interface, also being able to integrate with other toolchains.

The first step consists of scanning the entire IP range (see Figure 2), in order to detect active hosts. In real-world conditions, a seasoned attacker would opt to perform a slow scan, in order to go unnoticed by intrusion detection techniques based on volumetric network traffic analysis. For this stage, the NMAP tool will be used to implement a ping scan, with no DNS reverse resolution – this is a fast (albeit not error-proof) way to scan an IP range. Students are also taught to implement SYN and FIN scans, to acquire information about open ports and potentially available network services.



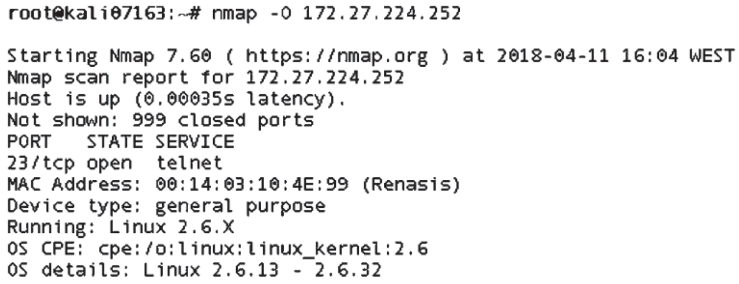
```
root@kali07163: # nmap -n -v -sn 172.27.224.0/24

Starting Nmap 7.60 ( https://nmap.org ) at 2018-0
Initiating ARP Ping Scan at 16:00
Scanning 255 hosts [1 port/host]
Completed ARP Ping Scan at 16:00, 1.05s elapsed (
Nmap scan report for 172.27.224.0 [host down]
```

```
root@kali07163:~# nmap -n -v -sS 172.27.224.0/24

Starting Nmap 7.60 ( https://nmap.org ) at 2018-0
Initiating ARP Ping Scan at 16:01
Scanning 255 hosts [1 port/host]
Completed ARP Ping Scan at 16:01, 1.04s elapsed (
Nmap scan report for 172.27.224.0 [host down]
```

Figure 2: Preliminary NMAP scans (left: ping scan, right: SYN scan)

Having access to the same Layer 2 domain as the testbed network (something that, in most cases, would not be possible in the open Internet), students are able to get information about Ethernet MAC addresses, which provides extra insights about the nature of the devices found on the network. Eventually, a MAC address with prefix "00:14:03" (mapped to Renasis LLC – an apparently out-of-business manufacturer) is found on the network, which hints at the possibility of that node being some sort of IoT device – also, SYN scans report that TCP port 23 is open, which further reinforces this diagnostic (port 23 is usually associated to the legacy telnet protocol, an insecure remote console service which is often active on IoT equipment, for maintenance and management purposes).

Furtherly, extra information is gathered from the device, using NMAP's OS detection feature (see Figure 3). NMAP's OS detection works by using TCP/IP stack fingerprinting techniques – sending a series of TCP and UDP packets to the target, and gathering evidence about features such as response patterns, TCP Initial Sequence Numbers, initial window sizes, protocol options support, TOS fields or IP ID (for the MDL period) predictability, among others. This information is compared to a database of known OS fingerprints.



```
root@kali07163:~# nmap -O 172.27.224.252

Starting Nmap 7.60 ( https://nmap.org ) at 2018-04-11 16:04 WEST
Nmap scan report for 172.27.224.252
Host is up (0.00035s latency).
Not shown: 999 closed ports
PORT   STATE SERVICE
23/tcp open  telnet
MAC Address: 00:14:03:10:4E:99 (Renasis)
Device type: general purpose
Running: Linux 2.6.X
OS CPE: cpe:/o:linux:linux_kernel:2.6
OS details: Linux 2.6.13 - 2.6.32
```

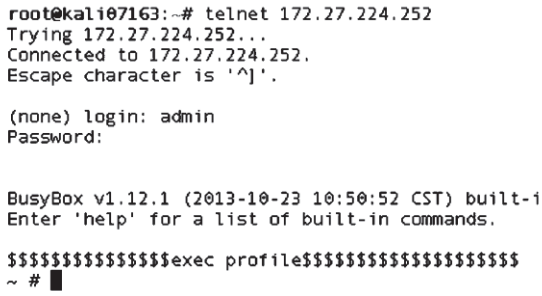
Figure 3: NMAP OS fingerprinting results

The results for this step (see Figure 3) further reinforce the assertion that the IP under analysis belongs to an IoT device, as it seems to host a Linux 2.6 kernel version (from a rather old branch), used in several SDKs for System-on-Chip (SoC) platforms and Board Support Packages (BSPs), between 2007 and 2015. This also hints at the nature of the embedded systems firmware, but on the other hand, leaves a lot of open possibilities regarding the specific CPU Instruction Set Architecture (ISA), as the Linux kernel has been ported to many platforms.

During this step, and accordingly with the background and skill level of the course attendees, instructors are advised to fill eventual knowledge gaps (for instance reviewing TCP/IP connection state diagrams or other required technical concepts), encouraging (and giving time for) students to explore these concepts on their own. For course groups with a more advanced skillset, this step offers an interesting opportunity to explore OSINT techniques, which may be leveraged by students to make the most of information that was gathered.

4.2 Insider profiling of the execution environment and hardware architecture

During the reconnaissance stage it was revealed that the device under scrutiny (IP: 172.27.224.252) presented what apparently seemed to be a telnet remote console service. A direct connection attempt (see Figure 4) confirmed the suspicion, with a login prompt being presented. Mirai-like infections frequently take place by means of these exposed and insecure services.

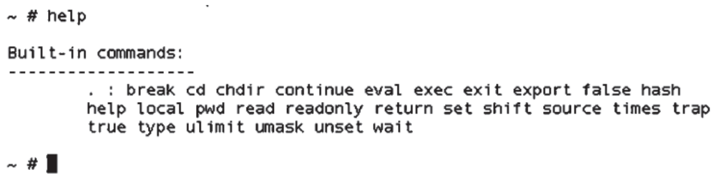


```
root@kali07163:~# telnet 172.27.224.252
Trying 172.27.224.252...
Connected to 172.27.224.252.
Escape character is '^]'.

(none) login: admin
Password:


BusyBox v1.12.1 (2013-10-23 10:50:52 CST) built-in shell (ash)
Enter 'help' for a list of built-in commands.

$$$$$$$$$$$$$$$exec profile$$$$$$$$$$$$$$$$$$$$
~ # █
```

Figure 4: Telnet session and *busybox* shell on the device

In this specific case, it didn't take too long to get shell on device – simple guessing revealed the access credentials to be *admin/admin*. For this purpose, Mirai deploys a dictionary-based attack, pretty much along the same lines of the technique used by the Morris worm in 1982. Not only this proves that some old tricks never go out of fashion, but it also gives the instructor the opportunity to revisit the Morris worm incident, also providing an excuse for recommending some related readings, namely Clifford Stoll's *The Cuckoo's Egg* (Stoll, 1989), whose epilogue tells the story of the author's efforts to fight the Morris worm.

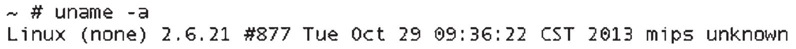
Once the session goes past the login prompt, users are greeted with a *busybox* prompt, further hinting at the nature of the device. *Busybox* is static binary which can be compiled to provide the functionality of many UNIX commonplace utilities (albeit in a more lightweight version), therefore allowing to replace a set of libraries and binaries with a single executable, providing a complete environment for embedded system firmware images. Further exploration has revealed that the *busybox* version that was compiled for this specific device has a rather reduced set of functionalities, which seem mostly oriented towards the execution of shell scripts (see Figure 5).



```
~ # help

Built-in commands:
-------------------
        . : break cd chdir continue eval exec exit export false hash
        help local pwd read readonly return set shift source times trap
        true type ulimit umask unset wait

~ # █
```

Figure 5: Functionality of the embedded *busybox* shell

Exploration (using the `uname` command) confirms a Linux 2.6.21 kernel, running on a MIPS ISA CPU (see Figure 6). This RISC CPU architecture still holds a relevant share of the embedded systems marked, with several OEMs producing licensed designs. These findings are also used as the motto for jumpstarting a more in-depth analysis of the underlying hardware platform of the device.



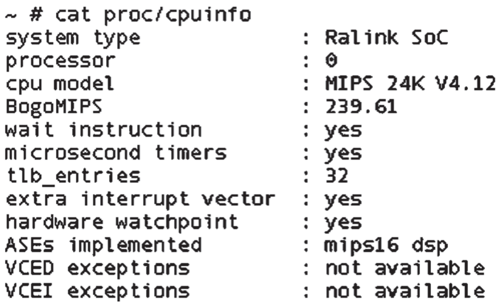
```
~ # uname -a
Linux (none) 2.6.21 #877 Tue Oct 29 09:36:22 CST 2013 mips unknown
```

Figure 6: Output of the *uname* command

The *procfs* filesystem provides a good starting point to search for specific system information – through a set of virtual files and folders, this filesystem exposes information about hardware, processes and device subsystems in a convenient way. In the case of a Linux system, places such as `/proc/meminfo` (memory information), `/proc/interrupts` (interrupt assignments), `/proc/bus` (for system buses, USB included) or `/proc/devices` (major numbers and device groups), `/proc/kmsg` (kernel message ring buffer) provide access to valuable information.

Particularly, `/proc/cpuinfo` (see Figure 7) and the whole structure of the `/proc` filesystem provide more information about the CPU, which students find to be a Ralink SoC (RT5350) (Ralink, 2010) supporting a MIPS 24K ISA (inspection of `/proc/meminfo` and `/proc/kmsg` also confirmed 32MB RAM and a CPU frequency of 360MHz). The RT5350 SoC consists of a highly integrated architecture incorporating a 32-bit MIPS CPU core, a 5-port fast ethernet switch, an 802.11n MAC and a USB host device interface, together with several hardware interfaces (UART, PCM, SPI, among others). One important feature of the MIPS family (which will be particularly relevant later on) has to do with endianness (the byte order used for storing multibyte numbers in memory): except for the R8000, MIPS processors can be configured to run in big or little-endian modes (Linux-MIPS, 2018).

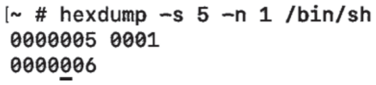
When dealing with basic to intermediate skilled audiences, this provides a convenient opportunity to refresh fundamental concepts on computer architectures.



```
~ # cat proc/cpuinfo
system type             : Ralink SoC
processor               : 0
cpu model               : MIPS 24K V4.12
BogoMIPS                : 239.61
wait instruction        : yes
microsecond timers      : yes
tlb_entries             : 32
extra interrupt vector  : yes
hardware watchpoint     : yes
ASEs implemented        : mips16 dsp
VCED exceptions         : not available
VCEI exceptions         : not available
```

Figure 7:`/proc/cpuinfo` dump

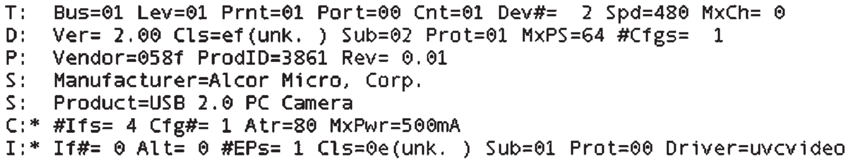
Discovering the endianness mode under which the RT5350 is operating is a task for OSINT. From the official data sheet, it can be found that the RT5350 endianness can be configured at bootstrap (with little endian being the default mode), so doubt still remains. That can be solved by studying the file header structure of the Executable and Linking Format (ELF) binary format used in Linux – the 5th offset byte contains the value 1 or 2 whether it was compiled for little or bin endian targets, respectively. By dumping the first five bytes of a firmware binary (`/bin/sh`) it is possible to confirm a 32-bit MIPS CPU Application Binary Interface (ABI) configured in little endian mode (see Figure 8).



```
[~ # hexdump -s 5 -n 1 /bin/sh
0000005 0001
0000006
```

Figure 8: ELF file header inspection

Analysis of the USB bus structure provides students with evidence for the USB host OHCI and EHCI controllers (for USB 1.x and 2.0 support) , as well as an embedded USB UVC (USB Video Class) device (see Figure 9).



```
T:  Bus=01 Lev=01 Prnt=01 Port=00 Cnt=01 Dev#=  2 Spd=480 MxCh= 0
D:  Ver= 2.00 Cls=ef(unk. ) Sub=02 Prot=01 MxPS=64 #Cfgs=  1
P:  Vendor=058f ProdID=3861 Rev= 0.01
S:  Manufacturer=Alcor Micro, Corp.
S:  Product=USB 2.0 PC Camera
C:* #Ifs= 4 Cfg#= 1 Atr=80 MxPwr=500mA
I:* If#= 0 Alt= 0 #EPs= 1 Cls=0e(unk. ) Sub=01 Prot=00 Driver=uvcvideo
```

Figure 9: USB UVC device characteristics and ID

This finding confirms the nature of the device: most likely an IP cam based on a OEM module (RT5350) attached to a standard USB webcam (of the same kind often found integrated in 201x-contemporary laptops – a search on `/proc/kmsg` retrieved the driver initialization messages and a list of supported resolutions and framerates). This finding can be, in fact, quite disturbing (something the instructors should stress with particular emphasis) as it means the device not only can be easily accessed, but it would be equally easy to extract a video feed from it, as all the standard Linux APIs (Schimek, 2009) are available – it just takes the time to cross-compile a binary for the target platform to make it possible to invade someone's else privacy.

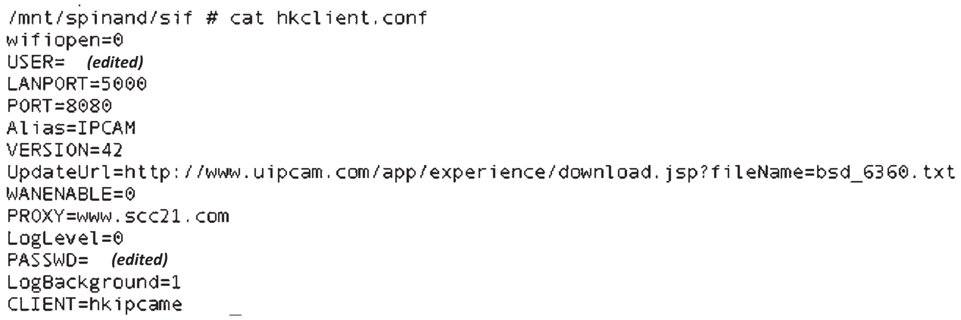
4.3 Exploitation and identification of vulnerable infection vectors

This specific step requires some knowledge about the traditional Unix SystemV-like initialization process, as well as basic process management and associated commands. The exploration process begins with the execution of simple `ps -ax` command, revealing a series of processes with the same command line invocation pattern:

`bin/hkipc` `<STRING> <STRING> www.scc21.com 8080`

A simple DNS query, together with the `dig` and `whois` tools made it possible to get information about the associated IP address for the *www.scc21.com* URL, the (Chinese) registrar for the domain and associated contacts. The site IP address was found to be hosted in the USA under the ownership of a Chinese local cloud provider branch, accordingly with a GeoIP database query. Since 2009 the associated domain was allocated to a company from the Chinese Guandong province (where the Chinese *Silicon Valley* of Shenzhen is located, also being adjacent to Hong Kong), whose site went offline as of 2019 – another site was found to be associated to the same IP address, later found to be also referred in the configuration files.

A configuration file was retrieved, containing several parameters appearing as arguments in the process listing, revealing their true purpose (see Figure 10). When comparing to the process listing arguments, it was found that first string corresponded to the USER (camera ID), the second string to the PASSWD and the URL corresponded to the PROXY configuration parameters, respectively.



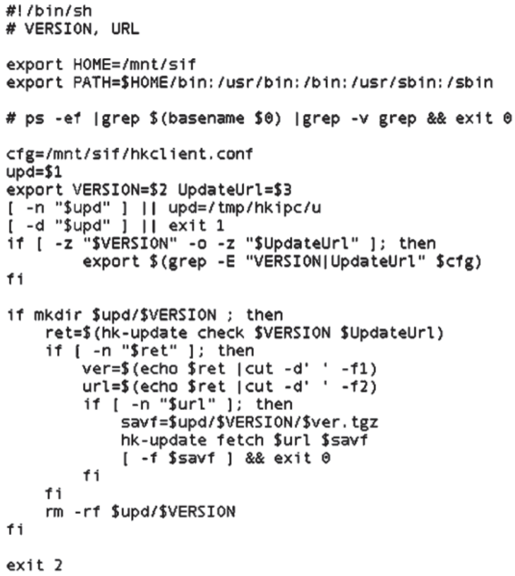
```
/mnt/spinand/sif # cat hkclient.conf
wifiopen=0
USER=  (edited)
LANPORT=5000
PORT=8080
Alias=IPCAM
VERSION=42
UpdateUrl=http://www.uipcam.com/app/experience/download.jsp?fileName=bsd_6360.txt
WANENABLE=0
PROXY=www.scc21.com
LogLevel=0
PASSWD=  (edited)
LogBackground=1
CLIENT=hkipcame    _
```

Figure 10: Internal configuration file

All evidence points towards the aforementioned processes being used to proxy stream the video feed to an external service (a list of open TCP and UDP ports obtained using `netstat`). The rationale for this is quite simple: since most IP Cams are deployed behind NAT firewalls (at least in residential LANs), the IP camera also sends a feed to a server on the Internet that can be accessed with a special application that allows to receive the video stream from anywhere. Unfortunately, sensitive data such as passwords and device IDs are stored in cleartext in the configuration file, together with an update URL that hints at the possibility of a remote update procedure (moreover, it resolves to the same IP as the proxy).

Often at this stage, and when the skill level of the students is adequate, there is an opportunity for instructors to challenge them to explore the inner workings of the device, with a special focus on update procedures. For this purpose, is not uncommon to suggest a teamwork-based approach, creating groups of two students with the intent of fostering collaboration, but also to encourage constructive discussions about relevant findings.

The update URL that was identified in the configuration raised suspicions about the existence of some sort of firmware update procedure, something that was confirmed after further investigation. Both during the boot process and at regular intervals (8 minutes past every 6th hour), the update script (see Figure 11) is executed by a `cron` job – the download process uses a wrapper script executing the `wget` tool to download the firmware payload over a HTTP cleartext connection.

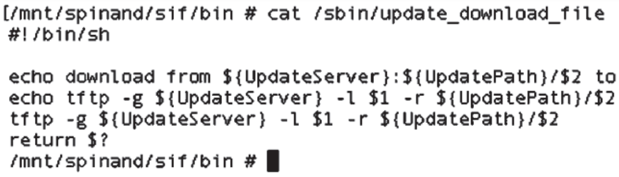


```
#!/bin/sh
# VERSION, URL

export HOME=/mnt/sif
export PATH=$HOME/bin:/usr/bin:/bin:/usr/sbin:/sbin

# ps -ef |grep $(basename $0) |grep -v grep && exit 0

cfg=/mnt/sif/hkclient.conf
upd=$1
export VERSION=$2 UpdateUrl=$3
[ -n "$upd" ] || upd=/tmp/hkipc/u
[ -d "$upd" ] || exit 1
if [ -z "$VERSION" -o -z "$UpdateUrl" ]; then
        export $(grep -E "VERSION|UpdateUrl" $cfg)
fi

if mkdir $upd/$VERSION ; then
    ret=$(hk-update check $VERSION $UpdateUrl)
    if [ -n "$ret" ]; then
        ver=$(echo $ret |cut -d' ' -f1)
        url=$(echo $ret |cut -d' ' -f2)
        if [ -n "$url" ]; then
            savf=$upd/$VERSION/$ver.tgz
            hk-update fetch $url $savf
            [ -f $savf ] && exit 0
        fi
    fi
    rm -rf $upd/$VERSION
fi

exit 2
```

Figure 11: Update script (`hk-check-update.sh`)

A TFTP-based update procedure was also discovered. This procedure allows for downloading and updating the firmware from the LAN. It is integrated within the SysV *init* boot process, being triggered by the execution of the

12chkupdate.sh service initialization script that, by its turn, executes a downloader script (see Figure 12) – the default TFTP server IP address is 192.168.1.13.

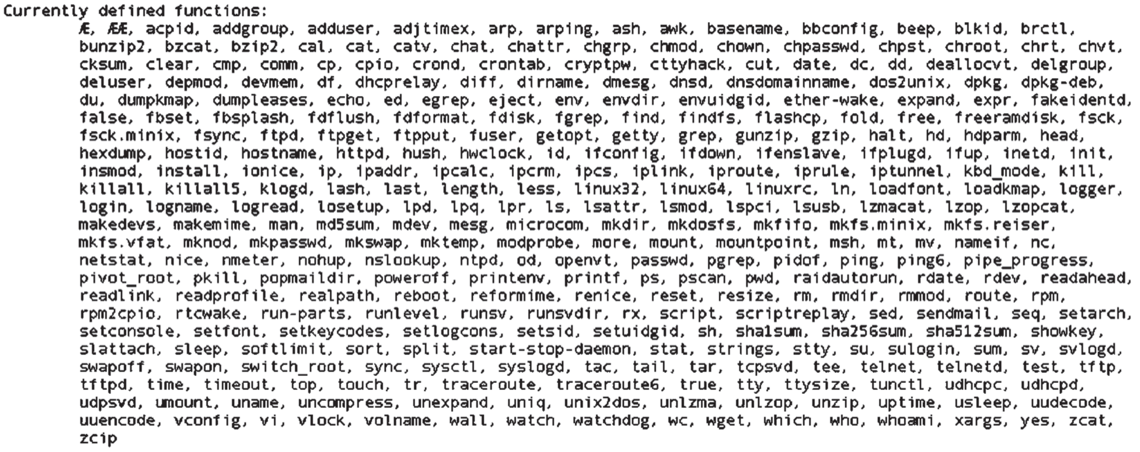


```
[/mnt/spinand/sif/bin # cat /sbin/update_download_file
#!/bin/sh

echo download from ${UpdateServer}:${UpdatePath}/$2 to $1
echo tftp -g ${UpdateServer} -l $1 -r ${UpdatePath}/$2
tftp -g ${UpdateServer} -l $1 -r ${UpdatePath}/$2
return $?
/mnt/spinand/sif/bin # █
```

Figure 12: TFTP update script

Overall, both the HTTP (wget) and TFTP procedures are insecure by nature, relying on cleartext data transfers – moreover if we consider the ease of intercepting and manipulating the data moving inflight for such connections. Also, DNS poisoning attacks or the hijacking an expired DNS record may be used to massively infect thousands of devices, without even needing to contact them directly.

Going back to the Mirai infection workflow, it is important to remind students that quite often an infection script is used to download and execute a malicious binary payload. To illustrate this, a cross-compiler setup was deployed and used to compile a fully-fledged *busybox* binary for a *mipsel* (MIPS - *endianness little*) target – this is the moment where the research about the endianness mode of the CPU paid off. Once the compiled binary was ready it was made available on a webserver, from which it was downloaded using wget. After a simple permissions adjustment (to provide execution rights), it was possible to execute the new payload, customized with a wide selection of built-in functionalities (see Figure 13).



```
Currently defined functions:
        Æ, ÆÆ, acpid, addgroup, adduser, adjtimex, arp, arping, ash, awk, basename, bbconfig, beep, blkid, brctl,
        bunzip2, bzcat, bzip2, cal, cat, catv, chat, chattr, chgrp, chmod, chown, chpasswd, chpst, chroot, chrt, chvt,
        cksum, clear, cmp, comm, cp, cpio, crond, crontab, cryptpw, cttyhack, cut, date, dc, dd, deallocvt, delgroup,
        deluser, depmod, devmem, df, dhcprelay, diff, dirname, dmesg, dnsd, dnsdomainname, dos2unix, dpkg, dpkg-deb,
        du, dumpkmap, dumpleases, echo, ed, egrep, eject, env, envdir, envuidgid, ether-wake, expand, expr, fakeidentd,
        false, fbset, fbsplash, fdflush, fdformat, fdisk, fgrep, find, findfs, flashcp, fold, free, freeramdisk, fsck,
        fsck.minix, fsync, ftpd, ftpget, ftpput, fuser, getopt, getty, grep, gunzip, gzip, halt, hd, hdparm, head,
        hexdump, hostid, hostname, httpd, hush, hwclock, id, ifconfig, ifdown, ifenslave, ifplugd, ifup, inetd, init,
        insmod, install, ionice, ip, ipaddr, ipcalc, ipcrm, ipcs, iplink, iproute, iprule, iptunnel, kbd_mode, kill,
        killall, killall5, klogd, lash, last, length, less, linux32, linux64, linuxrc, ln, loadfont, loadkmap, logger,
        login, logname, logread, losetup, lpd, lpq, lpr, ls, lsattr, lsmod, lspci, lsusb, lzmacat, lzop, lzopcat,
        makedevs, makemime, man, md5sum, mdev, mesg, microcom, mkdir, mkdosfs, mkfifo, mkfs.minix, mkfs.reiser,
        mkfs.vfat, mknod, mkpasswd, mkswap, mktemp, modprobe, more, mount, mountpoint, msh, mt, mv, nameif, nc,
        netstat, nice, nmeter, nohup, nslookup, ntpd, od, openvt, passwd, pgrep, pidof, ping, ping6, pipe_progress,
        pivot_root, pkill, popmaildir, poweroff, printenv, printf, ps, pscan, pwd, raidautorun, rdate, rdev, readahead,
        readlink, readprofile, realpath, reboot, reformime, renice, reset, resize, rm, rmdir, rmmod, route, rpm,
        rpm2cpio, rtcwake, run-parts, runlevel, runsv, runsvdir, rx, script, scriptreplay, sed, sendmail, seq, setarch,
        setconsole, setfont, setkeycodes, setlogcons, setsid, setuidgid, sh, sha1sum, sha256sum, sha512sum, showkey,
        slattach, sleep, softlimit, sort, split, start-stop-daemon, stat, strings, stty, su, sulogin, sum, sv, svlogd,
        swapoff, swapon, switch_root, sync, sysctl, syslogd, tac, tail, tar, tcpsvd, tee, telnet, telnetd, test, tftp,
        tftpd, time, timeout, top, touch, tr, traceroute, traceroute6, true, tty, ttysize, tunctl, udhcpc, udhcpd,
        udpsvd, umount, uname, uncompress, unexpand, uniq, unix2dos, unlzma, unlzop, unzip, uptime, usleep, uudecode,
        uuencode, vconfig, vi, vlock, volname, wall, watch, watchdog, wc, wget, which, who, whoami, xargs, yes, zcat,
        zcip
```

Figure 13: Custom, cross-compiled, *busybox*

Curiously, it is not unfeasible to use the functionality provided by the custom *busybox* binary to build scripted malware with relative ease (it only takes some ingenuity and creativity). For instance, it could be possible to create a simple script, designed to receive instructions from a command and control center and perform specific attacks against selected targets, or to perform horizontal scans within the LAN.

5. Conclusion and Future Work

This work documents an approach for a cybersecurity training module that designed with flexibility in mind. Instead of constituting a curricular monolith, the plan hereby presented intends to demonstrate how a real-world use case can be adapted for audiences possessing different backgrounds and skill sets. In fact, with the right didactic approach (such as the establishment of student groups with unbalanced skill levels), it may even be possible to deal with heterogeneous technical backgrounds within the same group of attendees.

The criteria for the selection of the use case hereby presented obeyed to three fundamental requirements, namely: to be relevant; to be current and to allow for intersections with other areas which are akin to different cybersecurity practitioner profiles. By choosing a topic (IoT) and a device (IP Camera) which are both relatable to the audience, the authors intended to maximize interest for the widest possible group of individuals, from freshman and undergraduate level, up to more skilled and experienced graduate students. Finally, it should be stressed out that the plan hereby presented has been put into practice by authors on several occasions, including classes and training sessions, with considerable success and positive feedback from students, with the latter being instrumental for the continuous improvement of the didactic approach.

References

Cruz, T. and Rosa, L. and Proença, J. et al. (2016) A cybersecurity detection framework for supervisory control and data acquisition systems. IEEE Transactions on Industrial Informatics 2016; 12(6): 2236–2246. DOI: 10.1109/TII.2016.2599841

Eliot, N. and Kendall, D. and Brockway, M. (2018) A Flexible Laboratory Environment Supporting Honeypot Deployment for Teaching Real-World Cybersecurity Skills, in *IEEE Access*, vol. 6, pp. 34884-34895, 2018.

Frazão, I. and Abreu, P. and Cruz, T. and Araújo, H. and Simões, P. (2018) Denial of Service Attacks: Detecting the Frailties of Machine Learning Algorithms in the Classification Process. In: International Conference on Critical Information Infrastructures Security, Springer, 2018: 230–235. DOI: 10.1007/978-3-030-05849-4_19

G. Lyon ("Fyodor"), "Nmap: the network mapper - free security scanner." [Online]. Available: https://nmap.org/

Hallaq, B. and Nicholson, A. and Smith, R. and Maglaras, L. and Janicke, H. and Jones, K. (2018) "CYRAN: a hybrid cyber range for testing security on ICS/SCADA systems. In: Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications. IGI Global (pp. 622–637).

Information Systems Audit and Control Association (ISACA), State of Cybersecurity 2020 [online], https://www.isaca.org/go/state-of-cybersecurity-2020

Lee C. amd Uluagac, A. and Fairbanks, K. and Copeland J. (2010) The design of NetSecLab: A small competition-based network security lab. IEEE Transactions on Education 2010; 54(1): 149–155.

Linux-MIPS project wiki [online]. Available at: https://www.linux-mips.org/wiki/Endianness

Maglaras, L. and Cruz, T. and Ferrag, MA and Janicke, H. (2020) Teaching the process of building an Intrusion Detection System using data from a small-scale SCADA testbed. *Internet Technology Letters*. DOI: 10.1002/itl2.132

Ralink (now Mediatek), RT5350 Preliminary Datasheet [online], November 25, 2010. Available at: https://cdn.sparkfun.com/datasheets/Wireless/WiFi/RT5350.pdf

Sawyer, Susan M et al. (2018) The age of adolescence, The Lancet Child & Adolescent Health, 2018. DOI: 10.1016/S2352-4642(18)30022-1.

Schimek, M. et al. (2009) Video for Linux Two API Specification [online]. Available at: https://www.linuxtv.org/downloads/legacy/video4linux/API/V4L2_API/spec-single/v4l2.html

Stoll, C. (1989) The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage, ISBN: 1416507787, ed. Pocket Books.

Teixeira, M.A. and Salman, T. and Zolanvari, M. and Jain, R. and Meskin, N.; and Samaka, M. (2018) SCADA System Testbed for Cybersecurity Research Using Machine Learning Approach. *Future Internet* **2018**, *10*, 76.

Trabelsi, Z and Saleous, H. (2018) Teaching keylogging and network eavesdropping attacks: Student threat and school liability concerns. In Proc of IEEE Global Engineering Education Conference 2018: 437–444.

Venkatesan, D. (2018) New Wave of Mirai Leverages Open-Source Project for Cross Platform Infection Technique [online], Symantec Threat Intelligence Blog. Available at: https://symantec-blogs.broadcom.com/blogs/threat-intelligence/mirai-cross-platform-infection

Zseby, T. and Vázquez, F. and King, A. and Claffy, K. (2015) Teaching network security with IP darkspace data. IEEE Transactions on Education 2015; 59(1): 1–7.

Cyber Crisis Management and Leadership

Didier Danet and Claude Weber
Saint-Cyr Military Academy, France
didier.danet@st-cyr.terre-net.defense.gouv.fr

DOI: 10.34190/EWS.20.032

Abstract: The trend toward greater use of networked systems is going on quickly, transforming existing public and private structures in digitized organizations. These digitized organizations are more than ever exposed to vulnerabilities and threats by criminal groups or state sponsored attackers. Reinforcing crisis management knowledge and practices is required. The paper addresses a classical dilemma in crisis management: should digital crisis management leaders seek to comply with pre-established formal procedures (continuity plans / recovery plans) which have been tested and approved or should they promote innovative solutions which may be needed to tackle unprecedented challenges? Referring to popular movies which provide many examples of this dilemma, should the cyber crisis management leader look like the Lt Gorman (bureaucratic leader in Alien) or Ripley (charismatic and innovative leader) ? Should Sully be prosecuted because he saved the plane by improvising a derogatory and risky solution (ditching the aircraft on the Hudson River)? In the paper, we consider that several answers are possible. According to the nature of the crisis universe that digital organization may face, the relevant leadership style may be the Gorman or the Ripley's one. The question is then: How to choose one style? How to decide to switch from one to the other? Which cyber crisis management structure could facilitate this choice? Two lessons can be learned from our paper: 1) The relevant leadership style mostly depends on the characteristics of the crisis universe (as defined by Snowden and Bonne) and 2) The choice may be enlighted by separating the unit which runs the crisis plans and the unit which monitors the results.

Keywords: Digital Transformation, Crisis Management, Leadership, CYNEFIN Framework, Crisis Management Organization

1. Introduction

Digital Transformation can be defined as a set of "changes associated with the application of digital technology in all aspects of human society." (Wikipedia) More than just a technological improvement, Digital Transformation is a cultural revolution which provides significant opportunities for cost killing, efficiency gains, networking... (Andal-Ancion, Cartwright and Yip, 2003) But if executives or employees lack the adequate mindset to take advantage of these opportunities, Digital Transformation will simply magnify existing weaknesses and flaws of the former organizations. Here is the dark side of Digital Transformation: The trend toward greater use of networked systems exposes digitized organizations to vulnerabilities and threats by criminal groups or state sponsored attackers. Reinforcing crisis management knowledge and practices is required even if a “Cyber Pearl Harbor” or a “Digital Chernobyl” remain hypothetical. But many digitized organizations had to face severe damages in the recent years because of malicious breaches or ransomwares.

The paper addresses a classical dilemma in crisis management topic: should digital crisis management leaders seek to comply with pre-established formal procedures (continuity plans / recovery plans) which have been tested and approved or should they prioritize innovative solutions which may be needed to tackle unprecedented challenges ? Which kind of leader and leadership should be promoted? In times of crisis, the choice of a charismatic and innovative personality is often acclaimed and perceived as the best option. But, most of the accidents and attacks in cyberspace will better be solved through a quick and compliant application of pre-established emergency plans by a bureaucratic leader. In other words, we don't (necessarily) need another hero to manage crisis in a digital world. The paper tries to define which style of leadership is relevant according to the specific characteristics of crisis in a world of digitized organizations.

2. From Cyberspace to Digital Age: a growing concern for crisis management

Crisis management in cyberspace has been a topic of interest for both practitioners and academics since a long time. As soon as economic, cultural or military activities became cyber dependent, experts underlined the multiple threats associated with the generalized interconnection of computers systems, especially for industrial plants remotely controlled or interconnected financial institutions.

Since the beginning of the 1990s, spying, sabotage or terrorists attacks are considered as major threats due to the huge damages which could result of a successful cyber-attack (Dunn Cavelty, 2010) Cyber crisis management has then been defined and conceived primarily as a way of dealing with accidents or attacks aiming at destroying or paralyzing information systems, creating disruptions in economic or social activities and, possibly, threatening human lives. Crisis management focuses on early detection of cyber threats, making sense of unclear events, stop the spread of the attacks, and maintain or restore the normal functioning of information systems. (Trimintzios, 2014) This approach of crisis management relies on computer and networks expertise. The goal is to understand vulnerabilities of interconnected information systems, identify evolving threats, establish strong barriers towards multiple and sometimes sophisticated attacks. In case of an attack, the answer relies on technical expertise, anticipation and prompt application of best practices and pre-defined procedures.

Today, the development of cyberspace has been involved in a much broader transformation encompassing for example cloud technology, Big Data, advanced analytics, artificial intelligence, machine learning, IoT, 5G... This set of intertwined innovations is giving rise to a whole "Digital Transformation" which impacts each and every dimensions of social bodies:

- Organizational ecosystems: digital technologies are key to the creation of extended networks of partners and stakeholders, giving rise to co-opetitive ecosystem approaches in which data and intangibles become major innovation assets;
- Organizational processes: one or more interconnected sets of operations and activities designed to achieve a specific organizational goal in which process management take advantage of new technologies such as autonomous robotics, AI or 3D vision;
- Organizational functions: Every function is concerned by Digital Transformation: communication, operations, human resources, administration, procurement, R&D, etc.

Such a transformation is not easy to define and understand. According to Morakanyane et alii, the Digital Transformation is a global process with certain characteristics (radical, disruptive) that is driven by digital technologies, capabilities and models, in order to create certain impacts (added value, relationships and coordination, saving, acceleration...) on certain aspects of the organization (ecosystem, business models, operational processes...) (Morakanyane, Grace and O'Reilly, 2017)

In such a digital world, threats and vulnerabilities do not disappear. On the contrary, they dramatically grow in number and complexity. Criminal organizations and military powers both enhance their hacking capabilities, build cyber weapons, deliver offensive operations. Digital crisis tend to be more and more likely and their consequences increasingly severe. Developing our digital crisis management capabilities may become one of the great issues in the years to come.

3. Who should take the lead in case of a digital crisis ?

Crisis management is often perceived as an art more than a science due to the mysterious alchemy which leads to success or failure. One of the core elements of successful crisis management lies in the people who are in charge. At the very heart of the debates is the nature of leadership which is best suited to exceptional situations where tough decisions need to be taken under high level of constraints and with enormous stakes. According to the common sense, the personal traits of the leader seem to be widely recognized as of crucial importance. Famous novels or movies willingly describe heroic characters, courageous, open minded, loyal and so much more. In most of these movies, crisis is the privileged moment in which exceptional characters emerge, gifted with charismatic personality, capable of thinking "outside the box", struggling against bureaucratic executives mired in absurd rules and procedures.

Alien (by Ridley Scott) is a famous exemplification of movies dealing with leadership in times of crisis. (Harrington and Griffin, 1990) As Harrington and Griffin underline it, the movie portrays different styles of leadership. Lt Gorman is described as an inexperienced officer who is bureaucratically referring to the legitimate authority and strictly complies with the rules of engagement even if they are inappropriate in the context of a life struggle against hostile extra-terrestrial creatures. Needless to say that standard procedures do not match the challenge of Xenomorphs attacks and that Gorman's panicking behavior leads to the Marines slaughtering. To the contrary, Ripley is a charismatic emergent leader. She is both technically capable (she impresses the Marines by her mastering of assault weaponry), gifted with interpersonal skills required to build a unified team, fighting against an almost invincible enemy thanks to a strategic vision and innovative tactical capabilities.

Beyond Alien, a very long list of famous movies promotes the charismatic leader as the best person to put in charge when crisis strikes and the situation seems desperate. Only charismatic leaders can recognize that a problem exists, seeking for help and trusting others to succeed (The King's Speech, Twelve O'clock High) They are exceptionally persuasive and dedicated to achieving common goals (Lawrence of Arabia, Twelve Angry Men) They do not submit to defeat, even in the face of it, enable others to act and work as a team, to innovate and reach impossible goals (Apollo 13, Thirteen Days) In many of these movies, the charismatic leader is exposed to incompetent, coward or antipathetic bureaucratic leaders (Paths of Glory)…

"Sully" by Clint Eastwood, starring Tom Hanks as the main character, is a very interesting example of blockbuster movie which promotes the charismatic and innovative leader as a victim of bureaucratic administration. The film follows the famous emergency landing of a plane on the Hudson River, in which all 155 passengers and crew survived with only minor injuries, and the subsequent publicity and investigation. Based on real facts, the movie reflects the public opinion that values a man, Captain Sullenberger, who gets an extraordinary result thanks to a mix of experience, initiative mind, ethics… and to the detrimental of the prescribed rules in case of an engine failure. The film director makes us share his anger towards the Federal Air Administration who sues the heroic captain, arguing that simulations demonstrates he could have land in a proper airport instead of taking maximum risks because of an improvised and unprecedented procedure. He clearly portrays the investigators as trying to smear the pilot. In the Guardian, Stephan Cass states: "In depicting government investigators as petty and clueless, the Hudson plane crash film trumpets a libertarian worldview at the expense of passenger safe." (Cass, 2016) This is precisely the question. In case of a crisis, what is the best for the exposed stakeholders. Should the leader comply with the rules and strictly follow the predefined procedures or should he be judged by the results whatever the means and the risks? Should have Sully been confronted to the investigators and compelled to justify his decision to break the rules even if passengers and crew were safe? Is the pilot some kind of “hero” who can deliberately infringe security rules to obey his guts or an obeying employee who cannot put aside the emergency rules pre-established by teams of security experts?

This popular preference for charismatic leaders would be of no importance if it was just a recipe for writing novels or films. But, in an interesting master thesis, Margaret K. Hanslik shows that this feeling is shared by police professionals who intervene in hostage situations. (Hanslik, 2018) According to the author, a positive correlation between preferred leadership traits in crisis management and charismatic leadership traits can be observed. On the contrary, a positive correlation between the leaders lacking those traits and the least preferred leaders in a crisis management situation is to be noted.

The academic literature is more nuanced than the popular approaches. The “One size fits all” approach is rejected: "A leader well suited to manage a particular crisis in a given organizational culture may not be fit to manage a different crisis in a different organizational culture". (Bowers, Hall and Srinivasan, 2017) We share the conclusion that crisis management requires some kind of contingent solution which depends on the nature of the crisis, the organizational culture and the personal style of the leader in charge of managing the crisis. Considering three different types of organizational culture (Elitist / Hierarchic / Adhocratic) and two sorts of crisis (internal / external), the authors conclude that a "directive leader" in a company whose culture is based on hierarchy should be able to manage the three main challenges of a crisis situation : time pressure, balance between communication and problem solving, implementation of a crisis plan. (1982 Tylenol Case / James Burke / Johnson & Johnson) But, this "directive leader" would probably not fit with crisis management in an adhocratic company such as Netflix where a “transformation leader” is more in line with the environment and gets better chances of success (2011 Product pricing problem / Reed Hastings / Netflix)

4. The specific characters of digital crisis and the choice of the good leadership style

Who could be the good leader in case of digital crisis? Should we rely on computing systems experts complying with rules and procedures such as Lt Gorman did? Would it be better to trust in charismatic leaders (Ripley or Sully) able to face unprecedented threats and imagine disruptive solutions ? That question cannot be answered without taking into account the specific characters of digital crisis to come.

The Digital Transformation is associated with a fast moving, technology driven process which impacts the very core functions, organizational culture and environment of all social ecosystems. From this basic observation follow the idea that the digitization is associated with an increasing vulnerability of our societies. Anticipating threats and designing relevant responses are more complicated than ever since an attack (or an incident) can take many forms and hit many layers of digital space: a ransomware deprives an organization of its files and data, a computer’s virus can erase disks or even destroy machines, a campaign on social medias may alter the reputation of an organization and the confidence of its partners. The cyber-attacks do not only threaten the physical infrastructure but also the data and, more generally, all intangible assets which are the main drivers of performance in contemporary world: intellectual property, trust, privacy...

From a theoretical point of view, we could argue that the digitization of organizations checks all the criteria Charles Perrow outlined to identify sectors prone to "normal accidents". (Perrow, 2011) Perrow considered accidents as "normal and inevitable" when three cumulative reasons are gathered:

- The systems are extremely complex;
- The systems are tightly coupled;
- The systems pose catastrophic potential in case of an accident.

Within a digitized organization, multiple forms of complexity develop due to the nature of the man-machine systems or the general interconnexion of more and more numerous systems (IoT for example) or, simply, because of the evolving configuration of global networks. The same finding prevails for the growing degree of coupling in modern information technology environments. Interactive complexity and tight coupling could easily and quickly aggregate to bring down other system functions, creating a cascade of catastrophic damages all over the spectrum of social activities. According to Perrow's typology, digitized societies should be considered as a privileged domain of "normal accidents".

In a digitized society, crisis are more frequent, more diverse and more complex. This means that economic, military or political organizations have to deal with a wide variety of attacks and threats. Most of them can be defeated by complying with pre-established procedures. No special leadership capabilities are required. The best crisis managers are talented experts able to understand the technical issue and fix it with appropriate known response. However, in a limited number of cases, attacks are driven by sophisticated actors (states sponsored), hidden behind faked identities and using unknown vulnerabilities. In these cases, complying with the procedures is not enough; the crisis managers need to try new solutions which results cannot be forecast with certainty. No single leadership style fits all situations. The crucial question is then: how to decide whether Gorman or RIpley is the good leader to deal with crisis in digital age?

5. Ordered and Unordered Crisis Universes : Gorman or Ripley ?

In order to solve the problem, Dave Snowden and Mary J. Boone suggest to consider the characteristics of the universe that the crisis manager is dealing with. "Truly adept leaders will know not only how to identify the context they're working in at any given time, but also how to change their behavior and their decisions to match that context". (Snowden and Boone, 2007)

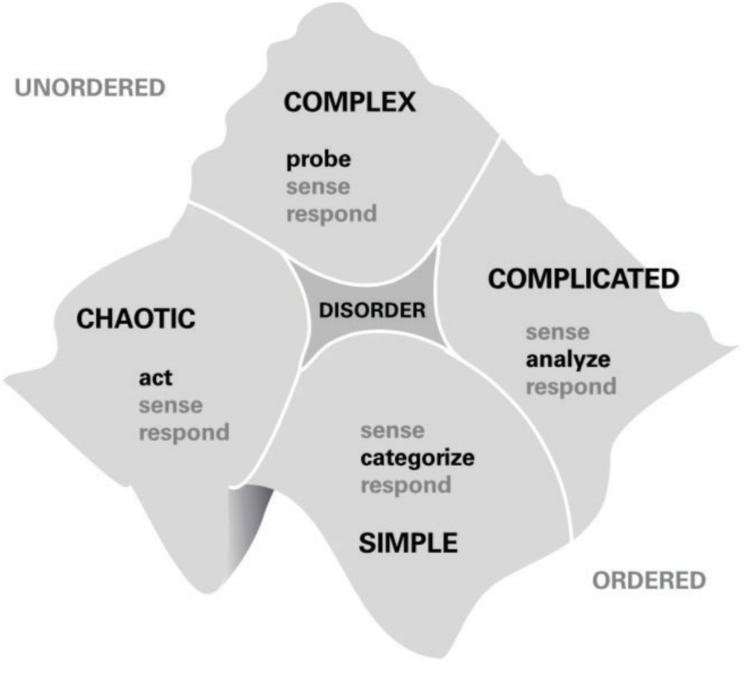


Figure 1: The CYNEFIN Framework, Snowden & Boone, 2007, p.4

In most cases, the events that the organization is confronted with obey a certain logic; they proceed from causes to consequences; they follow identified patterns and their results can be anticipated. In such cases, the leader is supposed to restore the ex-ante state of the system, maintain activity along the routines and established procedures and mobilize expertise in order to deal with the causes of the incident and stop the spiral of its development. A good example of this "Ordered Universe" is the famous cyber-attack against a French TV company whose website was defaced and broadcasting interrupted. (Corera, 2016; Nocetti, 2018) In an ordered universe, the person in charge of crisis management must ensure that: proper measures are implemented according to emergency plans, technical experts use best practices, different stakeholders cooperate efficiently, right messages are communicated through the relevant channels to the different partners, authorities and opinion... In the most complicated cases, this manager is supposed to identify and mobilize external competences and arbitrate conflicting expertises. The leader must sense, categorize or analyze the problem and respond in the proper way. The exploitation of known solutions according to pre-established emergency procedures is the key condition to face a crisis in an ordered universe. Gorman would perfectly match this profile.

But, in a certain number of cases, the crisis events do not follow the usual anticipated patterns. The crisis management unit is overloaded with flows of irrelevant or even contradictory data. Things that should not happen do occur. Standard procedures do not fix the problems. Causal relations cannot be discerned. The nature or the severity of the crisis itself may be questioned. For the leader, the main issue is to understand what is happening rather than applying the "one best solution" to an identified problem. In such unordered environment, the organization is threatened with dislocation and collapse. The leader must escape from crisis plans and act so that new patterns and solutions can emerge, be tested and adapted. Contrarily to the basic principle of crisis management - run the plan according to the pre-established scheme validated by experts and in cooperation with the few group of relevant stakeholders - the leader impulses experiments that allow novel patterns, actors and solutions to emerge. He sustains high levels of interaction and communication. He encourages dissent opinions and diversity. He looks for what works instead of what is complying with standards and expertise. According to Snowden and Boone, dealing with an unordered universe implies to break up with the logic of best practices, to put aside experts and crisis plans in order to let untested and collective solutions emerge from the chaos and provide the adequate response to the accident or the attack. In this case, the good leader to manage crisis is Ripley.

Since the two leadership styles are at the opposite one from the other, the key problem for the leader is to identify the characteristics of the context (ordered / unordered environment) so that he can adapt his style by switching from exploitation to innovation if necessary or, at least, put the emphasis on proven solutions or on experimentations.

6. How to choose between Gorman and Ripley?

In case of a digital crisis, the main question is: which style of leadership should prevail to face the threat? Should the leader comply with the best practices or should he innovate in disruptive solutions? Is the universe ordered or unordered? The answer is not that simple. The organization and the crisis management units are under high pressure and must decide very quickly in spite of incomplete information.

Fortunately, the greatest part of cyber attacks or digital crisis refers to well-known patterns and they perfectly fit with ordered universe. Solutions are to be found in emergency plans (ordered universe) Triggering crisis procedures and structures according to pre-established plans is then the relevant decision. Most often, the threat is neutralized thanks to automated procedures. For example, "spams" which were a major issue a few years ago are now automatically identified, categorized and isolated by email software. Applying by default the emergency plans and manage bureaucratically could be considered as a good way to face "ordinary crisis" in a digitized world.

But exceptions will remain. A complex or even chaotic universe may arise from a major accident so that the leader sometimes has to switch to the logic of discovery and innovation. This decision is from special importance because it implies to derogate from the reference framework and standard procedures without any certitude that the untested solutions will work better than the pre-established ones.

Due to the high chances of failure in putting aside the emergency plans, the decision cannot be abandoned to the "feeling" of the person in charge of crisis management structure. Nunamaker et alii suggest an interesting solution: implementing a specific organizational crisis management system. (Nunamaker Jr, Weber and Chen, 1989) In this model, the crisis command center is supposed to apply the emergency plans, following the logic of known solutions which is adequate to ordered universes.

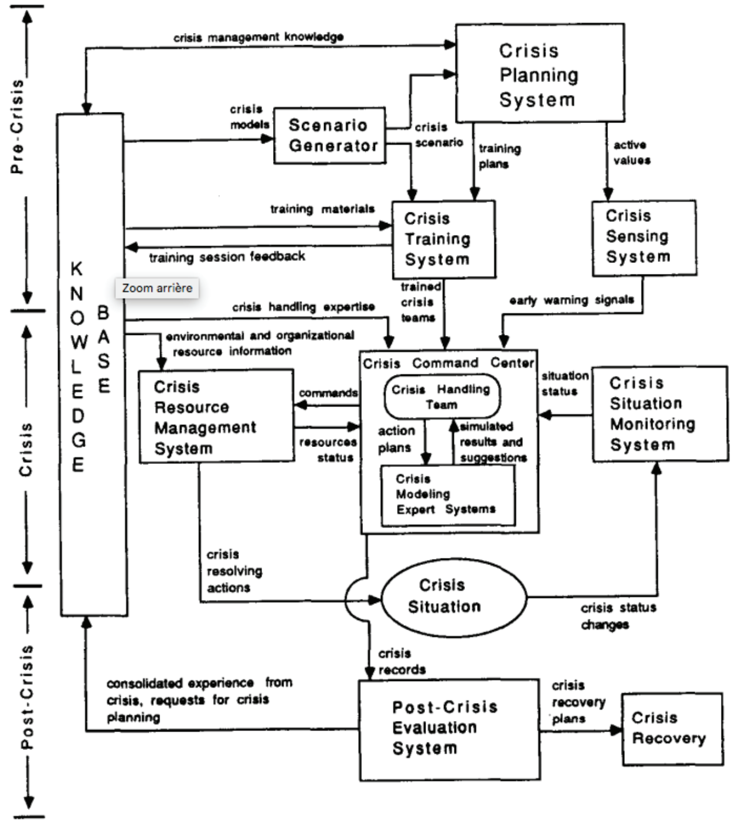


Figure 2: Model of A Crisis Management Environment (Nunamaker et alii, 1988, p.29)

The Crisis Command Center is backed by the "Crisis Situation Monitoring System". "Unaided, the crisis team stands little chance of tracking the development of events, their effects, or the actions of various decision makers. The crisis situation monitoring system... tracks and reports to the crisis team... so that the decision group does not miss any important signals in the confusion of events." In this organizational framework, the unit who leads the crisis management operations can concentrate on the implementation of pre-established crisis plans, considering by default that the crisis occurs in an ordered universe. The monitoring unit can concentrate on the relevance of this assumption. If the “One Best Solution” does not work, a switch towards disruptive solutions may be required. Changing from compliance to innovation implies to change the leadership style and, in a certain number of cases, the person in charge. Gorman may give place to Ripley.

In digital organizations facing numerous and heterogeneous accidents or attacks, crisis management leadership needs to master and combine both logics of exploitation (complying by standard procedures relevant in ordered universes where technical expertise is of primary importance) and innovation (implementing disruptive solutions in response to unknown or emerging patterns that defines unordered universe where multifaceted charismatic leaders are required) The distinction between the team who runs the emergency plans and the team who monitors the situation is a good solution to deal with the leadership dilemma.

7. Conclusion

This paper aims to explain the two leadership styles that need to be mastered and combined in order to manage crisis which could arise from the numerous and sometimes intentional and sophisticated threats associated with the ongoing Digital Transformation. More than ever, the relevant style of leadership in times of crisis needs to fit with a set of variables such as the personal traits of the leader in charge of the crisis management unit but also the organizational culture and the environmental context. Two main conclusions are to be considered.

First, the relevant leadership style may take two opposite values according to the crisis universe: ordered or unordered. In the case of an ordered universe, the leader in charge of the crisis management unit must strictly comply with pre-established emergency plans and the required capabilities are mostly relying on technical expertise; any deviation from the standard procedures will introduce delays, uncertainty among stakeholders and higher failure chances. In reference to popular movie characters, a good example of this kind of leadership is Lt Gorman. On the opposite side, in unordered universes, the leader must step aside from the pre-established procedures in order to let unknown causal relations emerge, understand new patterns and conceive disruptive solutions. The good leader in that case is Ripley or Sully.

It results from this first conclusion that promoting a leadership style and, possibly, switching from a leadership style to the other, is a key but difficult question in crisis situations. The implementation of a Crisis Situation Monitoring Unit, such as suggested by Nunameker et alii, is certainly an interesting and efficient formula. A distinction is established between the Crisis Command Center (who runs the plans and stays focused on emergencies) and the Crisis Situation Monitoring Unit (who assess the results) The Command Center focuses on the best practices and pre-established procedures while the Monitoring Unit appraises their effectiveness. Based on these results, the Board may characterize the crisis universe and decide to switch from best practices to disruptive solutions, from technical expertise to collective intelligence, from Lt Gorman to Ripley or Sully.

References

Andal-Ancion, A., Cartwright, P. A. and Yip, G. S. (2003) 'The digital transformation of traditional business', *MIT Sloan Management Review*, 44(4), p. 34.

Axelrod, R. and Iliev, R. (2014) 'Timing of cyber conflict', *Proceedings of the National Academy of Sciences*, 111(4), pp. 1298–1303.

Bowers, M. R., Hall, J. R. and Srinivasan, M. M. (2017) 'Organizational culture and leadership style: The missing combination for selecting the right leader for effective crisis management', *Business Horizons*, 60(4), pp. 551–563.

Cass, S. (2016) 'Sullied: with Sully, Clint Eastwood is weaponizing a hero', *The Guardian*. Available at: https://www.theguardian.com/film/2016/sep/12/sully-clint-eastwood-hudson-river-plane-crash-ntsb.

Corera, G. (2016) 'How France's TV5 was almost destroyed by "Russian hackers"', *BBC News*.

Dunn Cavelty, M. (2010) 'Cyberwar', *The Ashgate research companion to modern warfare The Ashgate Research Companion to Modern Warfare / ed. by George Kassimeris and John Buckley, ISBN 9780754674108*, pp. 123–144.

Hanslik, M. K. (2018) *The Use of Charismatic Leadership in Crisis Management in Policing*. Master of Science. Graduate Council of Texas State University.

Harrington, K. V. and Griffin, R. W. (1990) 'Ripley, Burke, Gorman, and friends: Using the film" Aliens" to teach leadership and power', *Organizational Behavior Teaching Review*, 14(3), pp. 79–86.

March, J. G. and Augier, M. (2004) 'James March on Education, Leadership, and Don Quixote: Introduction and Interview.', *Academy of Management Learning & Education*, 3(2).

Morakanyane, R., Grace, A. A. and O'Reilly, P. (2017) 'Conceptualizing Digital Transformation in Business Organizations: A Systematic Review of Literature.', in. *Bled eConference*, p. 21.

Nocetti, J. (2018) 'Géopolitique de la cyber-conflictualité', *Politique étrangère*, (2), pp. 15–27.

Nunamaker Jr, J. F., Weber, E. S. and Chen, M. (1989) 'Organizational crisis management systems: planning for intelligent action', *Journal of management information systems*, 5(4), pp. 7–32.

Perrow, C. (2011) *Normal accidents: Living with high risk technologies*. Princeton University Press.

Riasanow, T. *et al.* (2018) 'Clarifying the Notion of Digital Transformation in IS Literature: A Comparison of Organizational Change Philosophies', *Available at SSRN 3072318*.

Snowden, D. J. and Boone, M. E. (2007) 'A leader's framework for decision making', *Harvard business review*, 85(11), p. 68.

Trimintzios, P. ; H., Roger ;. Koraeus, Mats ;. Uckan, Boris ;. Gavrila, Razvan ;. Makrodimitris, Georgios (2014) *Report on Cyber Crisis Coooperation and Management*. Bruxelles: European Union Agency for Network and Information Security (ENISA), p. 52.

Warusfel, B. (2010) 'La protection des réseaux numériques en tant qu'infrastructures vitales', *Securite et strategie*, 4(2), pp. 31–39.

Water Distribution Network Leak Detection Management

Arno de Coning[1,2] and Francois Mouton[3,4]
[1]North West University, Potchefstroom, South Africa
[2]University of Pretoria, Pretoria, South Africa
[3]Noroff University College, Oslo, Norway
[4]University of Western Cape, Bellville, Western Cape, South Africa
arnodeconing@gmail.com
moutonf@gmail.com

DOI: 10.34190/EWS.20.088

Abstract: Water is becoming an increasingly scarce commodity and must be effectively managed to ensure minimal losses. Water is a basic human resource and during both a physical and a cyber war, limiting, or interrupting the supply can have a devastating effect on a country and its population. It is proposed to have an integrated systems approach to determine losses through leaks within a water distribution system by using monitoring equipment combined with occupancy data. The integrated system allows both the detection of cyber attacks, as well as, a potential reduction in losses by performing early leak detection. This paper covers an overview of current water distribution networks and explores the current cyber threat landscape. Furthermore, this paper also proposes a design for an anomaly leak detection system on an open-ended water distribution network. Having such a system in place can potentially mitigate the disastrous effect of a cyber attack on an organisation's water distribution network by acting as an alert trigger. This alert trigger can also serve as an input to existing security information and event management system.

Keywords: Cyber Security, Cyber Attacks, Water Distribution Networks, System Optimization, Anomaly Detection, Artificial Intelligence, Intrusion Detection, Leak Detection

1. Introduction

Water is a limited resource with a prediction that 40% of the world's population will live in water-stressed areas by 2035 (Guppy & Anderson, 2017), (The Water Project, 2020), (United Nations Department of Economic and Socials Affairs, 2020). The available fresh water has reduced by 55% from the 1960s and demand is projected to increase by 50% by 2030 (Guppy & Anderson, 2017). This has been equated to US$ 500 billion per annum due to water insecurity (Guppy & Anderson, 2017). The United Nations (UN) has developed, due to this scarcity and projections, the Sustainable Development Goal 6 (SDG6) to work towards water security to the world that is affordable to the masses (United Nations, 2017).

South Africa is currently been struck with droughts in several of its provinces. Cape Town was been declared a crisis zone with severe water restrictions to the urban communities of 50 litres of water per day (Baker, 2020), (Welch, 2020). These severe restrictions have been introduced to assist that there is water supply to the community and additional measures have been imposed to secure natural spring area to ensure water theft is minimised (Welch, 2020). The fact unfortunately, is that around 30% of water is lost through leakage within distribution networks and the Durban area has losses of 35% due to theft through illegal connections (The Water Project, 2020), (United Nations, 2017). Reduction in these losses will assist in alleviating some of the stresses in the water supply. This can be a focus from the clients' side to reduce usage through losses which in turn will reduce monthly bills as well.

The problem is that current water monitoring systems do not have a mechanism to perform early leak detection, and in turn, there is no ability to detect cyber attacks on the water system. It is unfortunate that current monitoring products still lack the ability to intelligently alert on high consumption while technology and commonly available techniques can greatly improve the reaction time on water leaks. This can have a devastating impact on the water distribution network when there is a cyber attack on the environment and have severe water supply implications. A proposal is therefore made to improve monitoring systems to detect leak flows and send alerts to minimise losses while implementing systems to mitigate impact of cyber attacks.

This paper focuses on a specific client site, an University Campus, to illustrate the impact of water leaks and how early detection would impact the environment. The campus already has a monitoring system in place, however, in its current state it does not perform early leak detection. The current system only has a high water flow alarm combined with a low water pressure alarm. The alarm is sent via a Short Message Service (SMS) messaging protocol. Currently, the system requires fixed threshold values to send alerts without intelligent alerting threshold adjusting. The value on high flow, for example, is set up based on the historic maximum flow of the specific site but does not take into account that other hours may have leaks while there is lower usage that will not be detected. A second system has been installed to collect usage data per hour and can be displayed on a web interface. This system has a vast amount of information but requires manual intervention to view each site on the system to determine if there are irregularities in water usage. Large enterprises may have several of these monitoring sites and the one being reviewed has close to 300 locations information that need to be reviewed as the system has no alerting mechanism installed.

The remainder of the paper is structured as follows. Section 2 addresses how an existing monitoring system operates with the current limitations. Section 3 proposes an intelligent leak detection approach for client-side water distribution networks by making use of the current monitoring system data. Section 4 discusses the cyber security risk and impact with design elements to mitigate the impact of cyber attacks. Section 5 provides a high-level overview of the functional units of the proposed system. The purpose of the high-level overview is to have a model which can be implemented in order to minimise losses through leaks within a water distribution network. The losses through leaks can be improved through the means of implementing a client-side anomaly leak detection algorithm.

2. Background

Water distribution networks management has four distinct approaches. The minimalistic approach will be to have a water distribution network without any installed monitoring equipment. This will then entirely require physical leak detection or reports from the clients when a leak has been detected. Improving upon this will be to install monitoring equipment. The one method with the monitoring equipment will be to have an alerting system on high flows or low-pressure alerts. Trend analysis of the usage is the next usage of the monitoring equipment where it will send the data to a central point. This can then further be expanded to add intelligence to improve the alerting on leak flows on the system.

The University uses two of these systems that are used to assist in bulk water supply monitoring across several campuses. The first is a dated system used to send alerts when a threshold limit has been triggered. The system is known as PMAC and is dependent on the installed equipment on the water supply but can alert on water flow and pressure data. The front page of the software will give an overview of all the sites with monitoring equipment as seen in Figure 1. A drill-down for specific sites will allow the user to see flow and pressure information for the sites.

Data is sent via SMS on a specific time interval that is set at 12 hours to limit data costs. This does provide a historic overview but is problematic if real-time data is required to be analysed to potentially detect small leaks that do not trigger the threshold setup. Figure 2 indicates the function available to view historic data of a site. This data is currently stored in several text files that make it difficult to perform queries and analyses the usage. It has been identified that converting the text files into a database structure, will have some added benefits but proprietary encryption on these text files complicates the implementation of this approach while adding data export functionality.

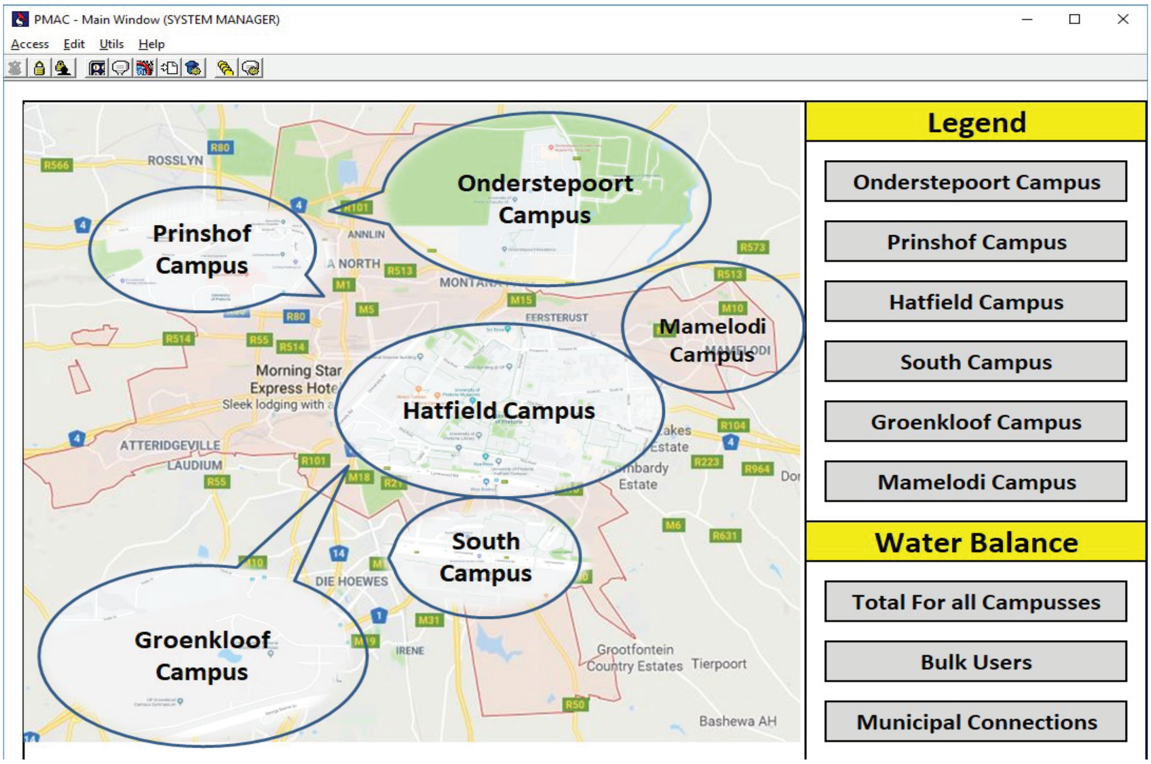
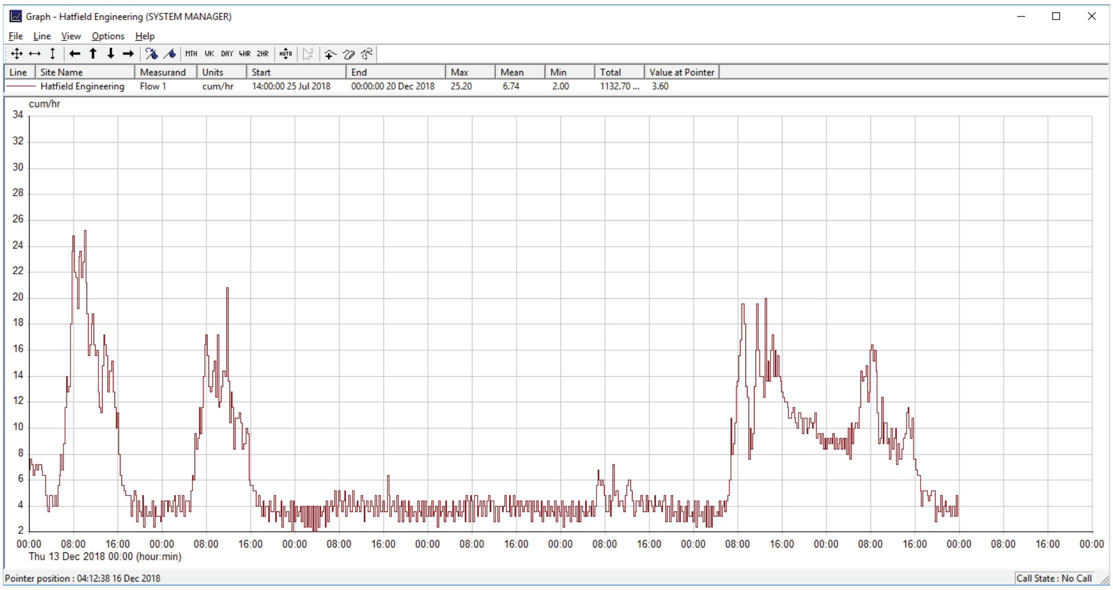


Figure 1: PMAC system overview



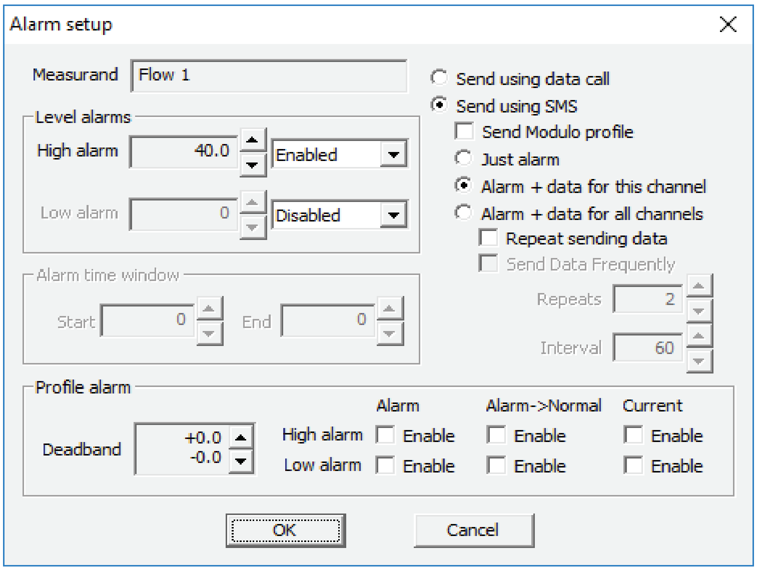
Line	Site Name	Measurand	Units	Start	End	Max	Mean	Min	Total	Value at Pointer
——	Hatfield Engineering	Flow 1	cum/hr	14:00:00 25 Jul 2018	00:00:00 20 Dec 2018	25.20	6.74	2.00	1132.70 ...	3.60

Pointer position : 04:12:38 16 Dec 2018

Call State : No Call

Figure 2: Data example

The alarm functionality is a key feature used on this platform as the daily usage is labour intensive to analyse. Alarm setup is available for specific threshold monitoring on the installed equipment. Figure 3 provides an overview of what is available to be set up for the alarm threshold of the PMAC system. Configuration of the alarm's threshold requires manual analysis of the historic data to set the limit that will be accurate to indicate a leak and not give false alarms to react on. The system will send data and alerts via SMS if the threshold limit is reached on a 15-minute interval.



Measurand: Flow 1

Level alarms

High alarm: 40.0 Enabled

Low alarm: 0 Disabled

Alarm time window

Start: 0 End: 0

○ Send using data call
● Send using SMS
□ Send Modulo profile
○ Just alarm
● Alarm + data for this channel
○ Alarm + data for all channels
□ Repeat sending data
□ Send Data Frequently

Repeats: 2

Interval: 60

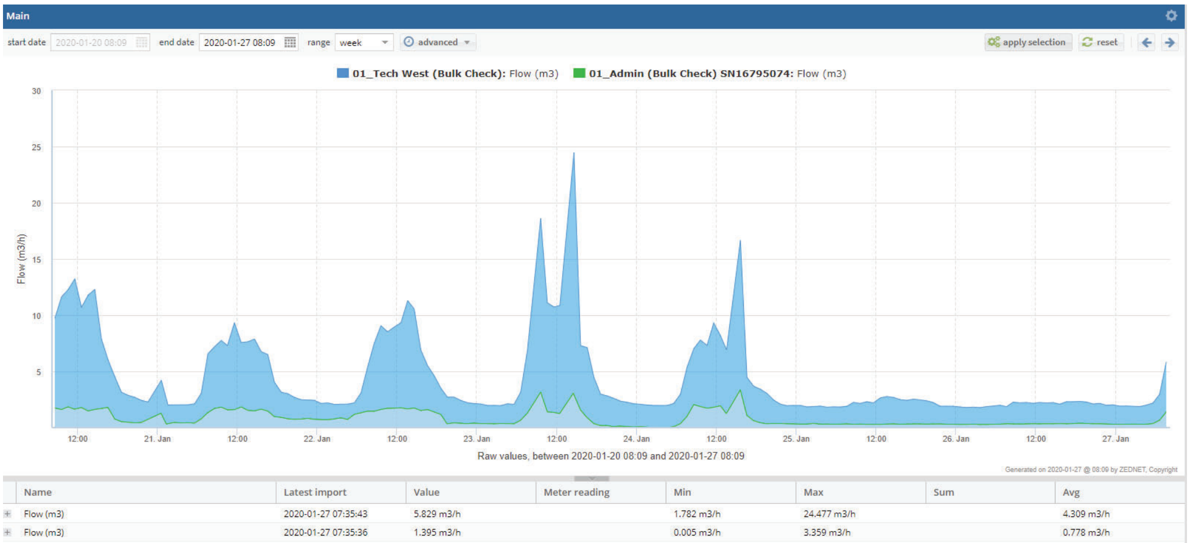
Profile alarm

Deadband: +0.0 -0.0

	Alarm	Alarm->Normal	Current
High alarm	□ Enable	□ Enable	□ Enable
Low alarm	□ Enable	□ Enable	□ Enable

Figure 3: Alarm setup example

The second system is that is used is called ZEDNET for monitoring of sites' water flow rates across several campuses at the University. It is a web-based system that requires extensive manual viewing if analysis has to be conducted to find leak flows. Figure 4 shows two sites overview for a week's worth of data with minimum, maximum, and average flow data. The system has a report feature to make daily of monthly leak flow reports that is emailed, but this also has either a massive time delay to find leaks or extensive manual work to review all the reports. A benefit of the ZEDNET system is that it makes use of a MySQL database that ensures easier access directly to the data to be analysed and automate analysis and interrogation of the database.


Main
advanced
apply selection
reset
01_Tech West (Bulk Check): Flow (m3)
01_Admin (Bulk Check) SN16795074: Flow (m3)
30
25
20
15
10
5
Flow (m3/h)
12:00
21. Jan
12:00
22. Jan
12:00
23. Jan
12:00
24. Jan
12:00
25. Jan
12:00
26. Jan
12:00
27. Jan
Raw values, between 2020-01-20 08:09 and 2020-01-27 08:09
Generated on 2020-01-27 @ 08:09 by ZEDNET, Copyright

start date 2020-01-20 08:09 end date 2020-01-27 08:09 range week

Name	Latest import	Value	Meter reading	Min	Max	Sum	Avg
⊞ Flow (m3)	2020-01-27 07:35:43	5.829 m3/h		1.782 m3/h	24.477 m3/h		4.309 m3/h
⊞ Flow (m3)	2020-01-27 07:35:36	1.395 m3/h		0.005 m3/h	3.359 m3/h		0.778 m3/h

Figure 4: ZEDNET system overview

These systems do have limitations that can be improved upon and the advantages and disadvantages need to be understood to assist in this process. Table 1 below indicates the primary advantages and disadvantages for the alerting focussed system in use at the university and the monitoring system. The alerting and monitoring system does not have intelligence included for leak detection and a manual threshold is required for high flow alerting on the alert system. The monitoring system has easy access to the data in a database while the alerting system data cannot easily be queried from the text files. Alerts on the current system are SMS based and can be costly when alerting numerous personnel, but the monitoring system has no alerting. Both of these systems require manual intervention to detect alerts from historic data with the attempt to detect the water leaks.

A proposal is made in section 3 to use the monitoring data and intelligently alert based on the multiple factors. While this will have a benefit, it may have security-related risks as any system may have and the risk and impact will be discussed in section 4 and the proposed system overview in section 5.

Table 1: Advantages and Disadvantages of the Current Systems

Alerting System	
Advantages	**Disadvantages**
Alerts on high leak flow	Fixed threshold to alert on
Historic data viewable on system	Difficult to analyse text files
	Manual intervention required to detect leaks
	Billed SMS based alerting
Monitoring System	
Advantages	**Disadvantages**
Easily viewable through web interface	Manual intervention required to detect leaks
Data stored in database	No alerting mechanism

3. Intelligent leak detection for client side

The focus of this paper is on the client-side of the water distribution system as the client has control to repair their water network. This process is assisted by having a quicker response to water leaks by making use of the monitor equipment on the water network on their own premises. No infrastructure can be expected to last forever and a failure will occur at some time. Once a leak is detected, a repair process can be started to reduce losses. Quicker response time on a leak will logically reduce losses of water within the distribution network. This proposed approach in principle should be user, supply or client-side, independent and only need the required monitoring equipment. Implementation can easily be expanded to the supply side distribution network as the same aspects is present within a larger network.

Water supply side requires a pumping scheme that needs to provide water flow into the system while pressurising the system as well. The performance of these pumps requires industrial control systems to ensure the correct functionality of the water supply into the distribution network. Maintenance on the pumping system is also essential to ensure that breakdowns do not occur, and that they are not run to failure. Condition monitoring has been implemented on multiple industrial equipment and 36% improvement on maintenance processes can be realised (Marais & Black, 2018). The same principle can be implemented on the pumps within the water network at the client or supply-side to assist in preventative maintenance instead of the run to failure method.

Artificial intelligence (AI) has gained great traction over recent years. The suggestion is, therefore, to use AI-based methods to assist in the detection of water leaks. Construction of the artificial neural network (ANN) will firstly require considering the data to be used as input. The monitoring system functions on a flow rate per hour with the equipment sending it to a base station with a radio frequency (RF) connection. This data is sent via a data connection to the bay station for processing. The usage of these around 300 sites require manual analysis to determine if there is an excessively high flow rate. The ANN implementation is to assist small teams to respond quickly without the need to manually check the data.

The first step is to easily have a report generated to detect night leak flows without having to manually go through the system daily. This analysis is done from 00:00 to 05:00 when the site should be empty and the flows that are detected have a high likelihood of being leak flows in buildings or bulk water supply lines. Once a high flow is detected a repair process can be started. The leak flow report includes the site name where the monitoring equipment is placed. The kilolitre per hour (kl/h) rate is then included with it converted to a South African Currency (ZAR) per day and the equivalent pipe size in millimetres (mm) that would make such a leak. An example of the report can be seen in Table 2 with the values sorted from high to low that will assist in prioritising the repair process to the highest water losses. Some knowledge of the network is required to understand if the reported leak is on bulk water supply line or buildings.

Table 2: Example of night leak flow report

Site	Night leak flow (kl/h)	ZAR per day	Equivalent pipe size (mm)
Site 1	1.3148	993.36	13.64
Site 2	1.2695	959.13	13.4
Site 3	1.2533	946.89	13.32
Site 4	0.1199	90.59	4.12
Site 5	0.1011	76.38	3.78
Site 6	0.9863	745.17	11.81
Site 7	0.9626	727.26	11.67

Seasonality is an observable phenomenon in several industries. Travel patterns change of a region based on vacations and fruit available is dependent on the time of year are some examples. This principle also applies with an open-ended water system, like the sites that are being investigated, and has different occupancy due to time of day and the time of year. The time of year classifier will, thus, assist in understanding the water usage when combining the time of year. Classifying the time of year is simple as this is a University and there are fixed schedules well in advance for the academic year. The first input as classifier are the following seven-day types for the academic year:

- Class weekday
- Class weekend
- Exam weekday
- Exam weekend
- Recess weekday
- Recess weekend
- Public holidays

Determining the occupancy of a specific area can be attempted by several methods. University has an access control system on the perimeter of the campus that can be utilised to obtain information regarding the number of personnel and students within the site. The data from this access control system is already stored in a SQL database and can be queried to determine the amount of entries and exits onto the campus. It should be noted that the data from this system does not always include absolute values, as vehicles that may have more than one person when entering, only requires a single participant to validate onto the access control system. This occupancy data should assist in seasonality for the bulk infrastructure and may have some correlation to the building level but will be required to be tested.

Occupancy based on cell phone movement on campus may give supplementary data on occupancy on the University campus. WiFi networks are open to all students and personnel on campus and can be used as an indicator of occupancy on campus and can be drilled down to building level as well (Ihsan & Mutia, 2019). The tracking of GSM for movement can also be used for devices that are not on the WiFi network (Ihsan & Mutia, 2019). This method of occupancy detection relies heavily on the occupants of the campus having devices on the WiFi or GSM network in the area. Inaccuracies in this method may also go over in the cases when one person has several devices on the WiFi or GSM network. Combining the two datasets may assist in reducing the inaccuracies in the datasets.

The third proposed method to detect occupancy is specifically for building level. CCTV cameras usage has increased exponentially over the past decade. This is primarily for security purposes but can be used in conjunction with some image processing to become people counting tools. Placement of the cameras has to be strategic to ensure entrances and exits are covered to accurately track the movement of the people. This dataset will only be used on building level to be combined with the leak flow data.

4. Security risk and impact

There is, unfortunately, risk involved in implementing the proposed systems. These systems require network communication to receive information from monitoring equipment which exposes it to cyber attacks (Mouton, et al., 2017). The first step to take is to analyse what is the risk if the systems do not function as intended due to cyber attacks.

Client-side proposed systems have several interfaces that may be open to potential attacks if it is not designed to minimise the risk. Monitoring equipment requires network communication that can be interrupted to ensure no data is sent to the base station. Communication can be attacked by a commonly used Distributed Denial-of-Service (DDoS) attack to overwhelm the communication of the monitoring equipment. This same technique can be implemented on the server-side to ensure that data cannot be received and the web interface with alerts cannot be accessed.

The web-based system will make use of a connection to a backend database. This can be vulnerable to SQL injection attacks if appropriate measures are not taken (Masango , et al., 2017). Data is essential in any of the proposed improvements and a severe SQL attack can delete all the data in the database and halt the operations completely. Alerting on high flows will in turn also not occur if the data is not available to be sent for the relevant personnel to respond.

The water supply side makes use of industrial control systems for the correct operations of the system. Stuxnet is one of the first attacks specifically targeting these industrial control networks and has crippled operations. This cyber attack has the intention to target the industrial networks but started with infection on windows computers which is another interface that needs to be defended against these attacks. Impact on the control systems on the water supply side can have devastating consequences. Water supply to a precinct can be halted if an attack alters or stops operations of the equipment pumping the water to the system. Combining these attacks with physical malicious actions can increase the impact. An example is that the firefighting water supply could be interrupted whilst there is an ongoing fire within the building.

Attacks on a water distribution can have a severe impact. The question is thus what steps can be taken to attempt to mitigate cyber attacks. Firstly, the web-based system can make use of simple yet effective methods to mitigate SQL injection attacks. This will greatly decrease the risk of unwanted alterations to the database that is essential for the detection of the leaks from the monitoring equipment data. Additional steps can be taken on the server to ensure ports are not open to being accessed or only accessed by authorised IP addresses. Logging of activity on the website can be used for analysis after a potential breach has occurred and indicated in Figure 5 as functional unit *logging*.

DDoS attacks on client and supply side is a potential threat. This method of attack has been used commonly but has changed from brute force large scale attacks to more sophisticated methods. The methods have evolved to include small attacks to determine vulnerabilities within the victims' system. It is therefore essential to monitor network traffic in these essential systems. Understanding and analysing the network traffic can assist in the detection of these attacks and take appropriate actions and indicated as *traffic analysis* functional unit in Figure 5. Attacks on the monitoring equipment will decrease the interval on data received on the system. A monitor can be built to periodically check whether data has indeed been received from the monitoring equipment and alert if they have not and this will be the *data interval analysis* functional unit. This will initiate a process to investigate potential attacks on the monitoring equipment and decrease time to restore correct data acquisition from the equipment.

Guidelines and best practices have been developed over the past decade to improve control systems cyber security (Stewart, et al., 2012), (Zetter, 2020). A wide variety of measures can be taken and include network segmentation, patch management, system monitoring, and awareness training (Stewart, et al., 2012), (Zetter, 2020). These networks can also benefit from network traffic monitoring as there is most likely a computer on the same network that can be a target for a cyber attack with the end goal to stop operations of equipment.

5. Proposed system design

The improvement can be made to address the inefficiencies of the current systems and a functional unit representation of the proposed system design can be seen in Figure 5. Leak detection from the monitoring system can be achieved by making use of AI techniques such as supervised training classifiers and then anomaly detection AI techniques. Input into the model will use water usage data, classification of the day, and the occupancy information with it visually represented as the *usage analysis* functional unit. Implementation of the technique is specifically based on the unsupervised training approach. This will assist in minimal input required by small teams to start detection of leaks from the monitoring system. The focus of the model will be to analyse real-time data with a high flow data point as the anomaly. The output of the anomaly detection will then be used to alert appropriate personnel to respond to leaks that have been detected.

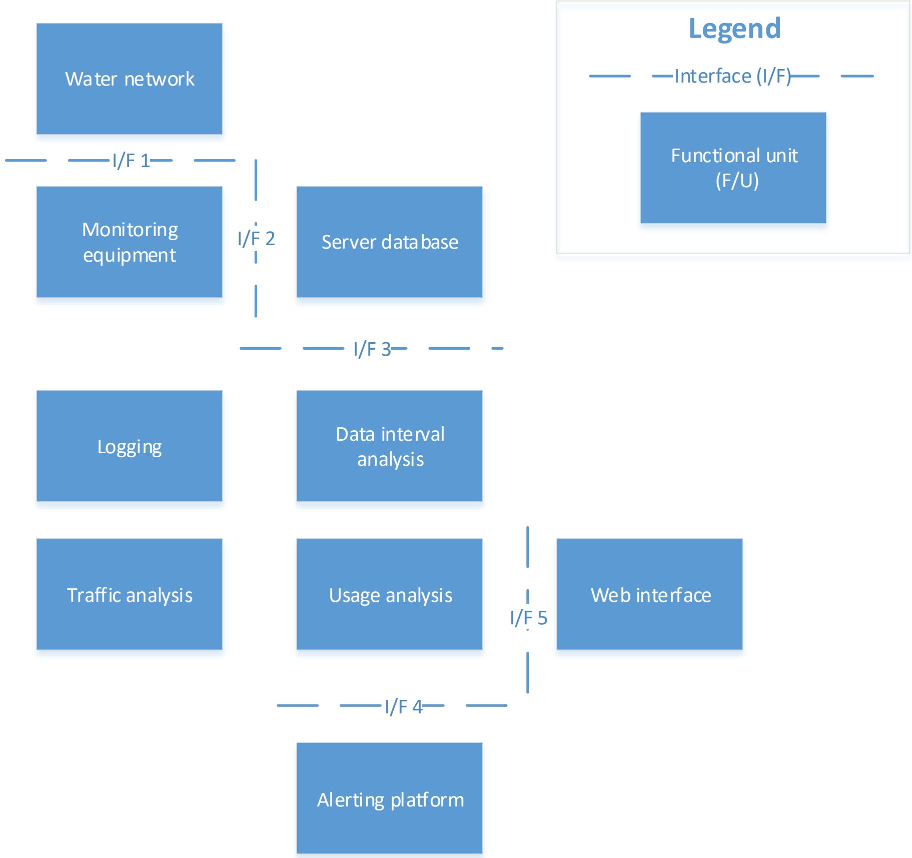


Figure 5: System functional units

This proposal will assist in only alerting on actual high flow usage while making use of day classification and the occupancy information. This requires an alternative to simplify receiving high leak flow reports through a secure channel. The first suggestion is to make use of a messaging application like Telegram messenger. This platform has an open bot application programming interface (API) that can be coded easily to automatically send alerts once a high flow has been triggered. Alerts will thus be able to be sent to a channel with the appropriate personnel to respond accessible through an Android, iOS, or computer version of the Telegram messenger and is indicated as the *alerting platform* function unit. Implementation of an alerting platform will decrease the manual intervention required that the current system requires. This process can be complemented by developing a web-based dashboard or *web interface* functional unit to only display sites with current high flow alerts. Limiting alerts only to what has triggered by the proposed methods will assist in decreasing false alarms while reporting on leaks that would not trigger fixed thresholds that the current system sends.

6. Conclusion

Water as a resource is becoming more scares and 40% of the world's population will live in water-stressed areas by 2035. Fresh water supply has reduced by 55% from the 1960s and projections indicate that the worlds demand will increase by 50% by 2030. Water distribution networks are an essential part to deliver water to communities but 30% of water is lost through leaks during this process. Current monitoring systems have some drawbacks that inhibit quick response times to leaks within the networks. These limitations include fixed threshold reporting on alerting systems and monitoring systems that has no alerting mechanism available.

Background on the water network monitoring on a University site has been discussed with the current inefficiency of the two systems currently in use. A proposal was made to implement an intelligent leak detection design for client-side operation to improve reaction time on leaks that will, in turn, reduce losses in the water network. This system will make use of the monitoring system data and combined with AI model to alert on leaks within the network. Alerts will then feed directly to a Telegram channel and a web-based dashboard so appropriate action can be taken to fix the leak. Achieving the proposed system require a design that will mitigate the impact of cyber attacks. A proposal is made to reduce this risk by combining several industry standards into the design of the system.

Future work will be to implement the proposed prototype. This will then be tested on real-time data on the University site while documenting the reduction in losses within the water network. System design optimisation will be conducted to further improve leak detection and reduce the risk of a cyber attack.

References

Baker, A., 2020. *What It's Like To Live Through Cape Town's Massive Water Crisis.* [Online] Available at: https://time.com/cape-town-south-africa-water-crisis/

Guppy, L. & Anderson, K., 2017. Global Water Crisis: The facts. *Hamilton, UNU-INWEH.*

Ihsan, A. & Mutia, E., 2019. Wi-Fi and GSM Based Motion Detection in Smart Home Security System. *IOP Conference Series: Materials Science and Engineering*.

Marais, H. & Black, J., 2018. *A low-cost condition monitoring solution for industrial bakery equipment.* Rustenburg South Africa, s.n., pp. 347-354.

Masango , M., Mouton, F., Antony, P. & Mangoale, B., 2017. *Web Defacement and Intrusion Monitoring Tool: WDIMT.* Chester, International Conference on Cyberworlds (CW), pp. 72-79.

Mouton, F., Teixeira, M. & Meyer, T., 2017. *Benchmarking a mobile implementation of the social engineering prevention training tool.* Johannesburg, Information Security for South Africa (ISSA), pp. 106-116.

Stewart, J. M., Chapple, M. & Gibson, D., 2012. *Certified Information Systems Security Professional Study Guide Sixth Edition.* Indianapolis: John Wiley & Sons, Inc..

The Water Project, 2020. *Water in crisis - South Africa.* [Online] Available at: https://thewaterproject.org/water-crisis/water-in-crisis-south-africa

United Nations Department of Economic and Socials Affairs, 2020. *International Decade of Actions 'Water for Life' 2005 - 2015.* [Online] Available at: https://www.un.org/waterforlifedecade/scarcity.shtml

United Nations, 2017. *Goal 6: Ensure access to water and sanitation for all.* [Online] Available at: https://www.un.org/sustainabledevelopment/water-and-sanitation/ [Accessed 29 November 2019].

Welch, C., 2020. *Why Cape Town Is Running Out of Water, and Who's Next.* [Online] Available at: https://www.nationalgeographic.com/news/2018/02/cape-town-running-out-of-water-drought-taps-shutoff-other-cities/

Zetter, K., 2020. *An Unprecedented Look at Stuxnet, the World's First Digital Weapon.* [Online] Available at: https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/

Cyber as a Hybrid Threat to NATO's Operational Energy Security

Arnold Dupuy[1], Ion Iftimie[2,3], Daniel Nussbaum[4] and Stefan Pickl[5]
[1]Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]NATO Defense College, Rome, Italy
[3]European Union Research Center, George Washington School of Business, Washington, D.C., USA
[4]Energy Academic Group, Naval Postgraduate School, Monterey, CA
[5]Institut für Theoretische Informatik, Mathematik und Operations, Research an der Universität der Bundeswehr München
acdupuy@vt.edu
iftimie@gwu.edu
danussba@nps.edu
stefan.pickl@unibw.de
DOI: 10.34190/EWS.20.044

Abstract: This paper discusses the evolving role of NATO in addressing cyber as a hybrid threat to operational energy security. Cyber threats to NATO's operational energy security, and particularly to interconnected liquid fuel supply chains and grids, have the capacity to directly impact NATO's enhanced Forward Presence and the success of forward deployed units. Despite this, the academic literature on cyber warfare currently does not address NATO's role in combating the cyber threats to its operational energy security. We posit that, in practice, NATO defines energy security as the assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements. Starting from this definition, we propose that Baldwin's analytic framework can be used by NATO to prioritize the protection of critical energy systems in the cyber domain. This effort is needed to advance future research by the NATO Science and Technology Organization related to prototype-operating models providing early warning on cyber-attacks against NATO's energy security.

Keywords: NATO, Cyber Defense, Operational Energy Security, Enhanced Forward Presence, Forward Deployed Units

1. Introduction

In this paper, we discuss the evolving role of the North Atlantic Treaty Organization (NATO) in addressing cyber as a hybrid threat to the critical energy infrastructure (CEI) that NATO's enhanced Forward Presence and forward deployed units rely upon. Over the past two decades, cyber-attacks against industrial control systems (ICS) of NATO member states energy supply chains have grown exponentially. Between 2012 and 2014, the number of reported cyber-attacks against the supervisory control and data acquisition (SCADA) systems of the United States alone, have increased by 636 percent (IBM, 2015: 5). In an effort to assert the new role of NATO in defending this CEI from cyber-attacks, the 2014 Wales NATO Summit "added the cyber dimension to the dimensions in which NATO exercises collective defence, as it has been noticed that cyber-attacks can have the same effect on the territorial integrity of NATO countries as conventional attacks" (Flamant and Parrein, 2017: 53). Not long after, on Dec. 23, 2015, Russia conducted a cyber-attack against the electrical grid of Ukraine, a member of the North Atlantic Cooperation Council and the Partnership for Peace programme, causing "a six-hour blackout for hundreds of thousands of customers" (Sullivan and Kamensky, 2017: 30). Since then, Russia and other state and non-state actors have been consistently targeting the CEI of the United States and its NATO allies, as well as the CEI of NATO partners that provide the energy requirements for forward deployed units (CISA, 2018; Coats, 2019).

Cyber threats to the interconnected liquid fuel supply chains and grids that support NATO operations have the capacity to directly impact the NATO mission, highlighting the need for unified civil-military responses in the cyber domain. Vulnerable energy systems range from the refinery, bulk transport, storage, transit and delivery of fuels critically needed by forward deployed units to the electrical grids that power NATO critical infrastructure. These critical energy systems can be challenged through any number of attack vectors that could have devastating results on the military effectiveness of the Alliance and could negatively impact mission success. This highlights the importance for NATO to secure these energy systems and to treat them as critical components across all aspects of the battlespace. Any impediment to access affordable and acceptable energy supplies by NATO infrastructure and forward deployed units, or the so-called 'last tactical mile', via power grids, pipeline, road, rail, air or barge, will inhibit NATO's ability to operate seamlessly across the five domains (land, sea, air, space and cyber) and could severely undermine NATO's enhanced Forward Presence in the eastern part of the Alliance.

In the following sections, we propose a working definition of energy security for NATO, then address the cyber threats to information technology (IT) and operational technology (OT) of the energy systems vital for maintaining NATO operations. We then discuss how Baldwin's analytic framework could help NATO in addressing the raising number of attacks from both state and non-state actors in the cyber domain. This research is needed to advance new research by the NATO Science and Technology Organization related to prototype-operating models providing early warning on cyber-attacks against NATO's energy security.

2. Energy security and NATO

The energy security threats until the end of the Cold War have traditionally been defined by the robustness, or ”long-term access to” (Deane et al., 2015), energy supplies, in particular, to oil (Yergin, 1988, 2006). This is embodied in the energy security definition used by the International Energy Agency (IEA), which was created in response to the 1973 oil crisis, as “the uninterrupted availability of energy sources at an affordable price” (IEA, 2019). With the Heads of State and Government and High Representatives of the United Nations committing the international community in 2015 to adopt the sustainability principle during “low-carbon energy transitions” (Bridge et al., 2013: 331; Geels, 2014: 30; Goldthau and Sovacool, 2012: 232), the IEA also emphasizes that long term energy prospects must be in line with “environmental needs” (IEA, 2019). In effect, the IEA adopted the Asia Pacific Energy Research Centre's ‘4 As’ of energy security: availability, accessibility, affordability and acceptability (Centre, 2007).

NATO's interest in energy security emerged during the 2008 Bucharest Summit (NATO Heads of State and Government, 2008). NATO, however, does not have a working definition for energy security, despite the fact that NATO military operations are dependent on the availability of energy supplies and despite the fact that NATO stated in 2008 its commitment to “first and foremost ensuring a steady supply of fuel to its military forces” (Khamashuridze, 2008: 53). Instead, every NATO member state uses its own definition of energy security. For instance, the National Defense Authorization Act for Fiscal Year 2018 defines the terms energy security for the United States Department of Defense as “having assured access to reliable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements” and energy resilience as “the ability to avoid, prepare for, minimize, adapt to, and recover from anticipated and unanticipated energy disruptions in order to ensure energy availability and reliability sufficient to provide for mission assurance and readiness, including task critical assets and other mission essential operations related to readiness, and to execute or rapidly reestablish mission essential requirements” (*National Defense Authorization Act for Fiscal Year 2018*, 2017: 2831).

After looking at the various definitions of energy security used by the 29 member states, we propose that in practice, NATO defines energy security as having assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements. Particular attention needs to be paid here to the addition of the acceptability principle, which encompasses key environmental considerations acknowledged by NATO's 2010 Strategic Concept. Using this working definition, we can assert that NATO's energy security is threatened by a wide variety of hybrid threats that must be addressed by NATO in order to meet mission essential requirements. The European Center of Excellence for Countering Hybrid Threats (Hybrid CoE) defines a hybrid threat as an action by state or non-state actors aimed to “undermine or harm the target by influencing its decision-making at the local, regional, state or institutional level” (Hybrid COE, 2020). Within the context of NATO energy security, hybrid threats are, thus, action by state or non-state actors aimed to undermine or harm NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements by influencing its decision-making at the local, regional, state or institutional level.

3. Cyber as a hybrid threat to energy security

The level of cyber threats against NATO energy security is “capable of holding multiple dimensions and taking on different specificities depending on the country” (Chester, 2010/2: 887). In other words, each NATO member state handles cyber-attacks against their CEI using cyber capabilities that vary vastly from member state to member state. Furthermore, both NATO and the European Union are developing their own means and ways to respond to the growing number of cyber-attacks against both national and regional/interconnected CEI. The discrepancy of how NATO member states prioritize cyber threats and the growing competition between NATO and the EU on the ways and means to address these cyber threats against CEI has significant implications for NATO's operational energy imperative. For the purpose of this paper, a cyber-attack is "an attack on a computer and network system, consisting of computer actions (e.g., remote or local connection, computer file access, program execution, etc.)" (Chi et al., 2001: 320) to compromise the secure operation of "automated systems for storing, processing, and distributing information" (van den Hoven et al., 2018) and "computer-controlled physical processes such as industrial control systems or other types of control systems" (Bryant, 2016: 7).

A cyber-attack against IT and OT of CEI represents one of the cheapest and most available forms of hybrid threats to the energy security of NATO member states and to the Alliance as a whole. These types of attacks against CEI are not a new phenomenon. For example, as early as December of 2002, "hackers were able to penetrate the SCADA system responsible for tanker loading at a marine terminal in eastern Venezuela, [preventing] tanker loading for eight hours" (Byres and Eng, 2009: 58). In June of 2010, the 'Stuxnet' worm caused the centrifuges of an Iranian nuclear facility (Farwell and Rohozinski, 2011: 23) "to tear themselves apart" (Kushner, 2013), slowing down the progress of Iran's nuclear program. Examples of cyber-attacks against CEI around the world are plentiful, and the success of NATO forward military operations is threatened by them. As such, NATO needs a framework for assessing the cyber threats to NATO energy security.

4. The proposed framework for assessing the cyber threats to NATO energy security

According to Baldwin, security is defined as the "low probability of damage to acquired values" and requires national security professionals address three basic questions: security for whom, for which values, and from what threats? (Baldwin, 1997). Cherp and Jewell restated these questions as 1) what critical systems to protect, 2) from which risks, and 3) by which means? (Cherp and Jewell, 2014). In the context of cyber threats to NATO energy security, answering the 'what critical systems to protect?' question will define the critical energy systems that guarantee NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements. Answering the 'from which risks to protect?' question will identify the cyber threats from state or non-state actors aimed to undermine or harm these energy systems by influencing its decision-making at the local, regional, state or institutional level. Finally, answering the 'by which means to protect?' question will delineate the available means that NATO has at its disposal to address these cyber threats to the critical energy systems that guarantee NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements.

4.1 What to protect?

Availability and access to affordable and acceptable energy sources by NATO in the post-Cold-War security environment "depends on a complex system of global markets, vast cross-border infrastructure networks, a small group of primary energy suppliers, and interdependencies with financial markets and technology" (Chester, 2010/2: 887). This means that NATO's military activities are increasingly dependent on energy infrastructure owned and/or operated by public and private entities. These critical energy systems are located both within the territory of NATO member states and along the routes supplying NATO's forward deployed forces (see Figure 1).

Additionally, within much of old NATO across Europe, NATO fuel requirements rely on supplies from the NATO Pipeline System (NPS). The NPS is comprised of eight national—the Greek Pipeline System (GRPS), Icelandic Pipeline System (ICPS), Northern Italy Pipeline System (NIPS), Norwegian Pipeline System (NOPS), Portuguese Pipeline System (POPS), United Kingdom Government Pipeline and Storage System (UKGPSS), and the Turkish Pipeline Systems (TUPS)—and two regional—the Central Europe Pipeline System across Belgium, France, Germany, Luxembourg and the Netherlands (CEPS) and the North European Pipeline System across Denmark and Germany (NEPS)—storage and distribution systems (NATO, 2017). The geography of old NATO is thus, characterized as a 'developed theater' (Garrett, 1993: 11), where, with the exception of CEPS, national organizations manage the operations and security of these NATO pipeline systems, and often also use them to supply their civilian national airports with fuel (NATO, 2017).

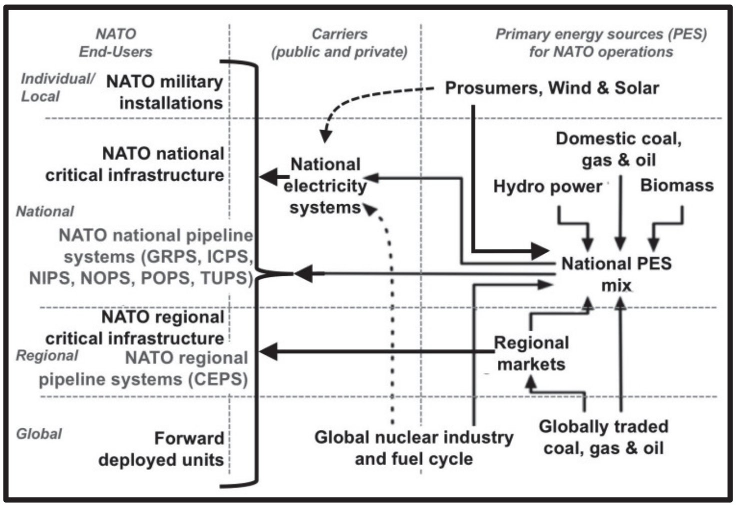
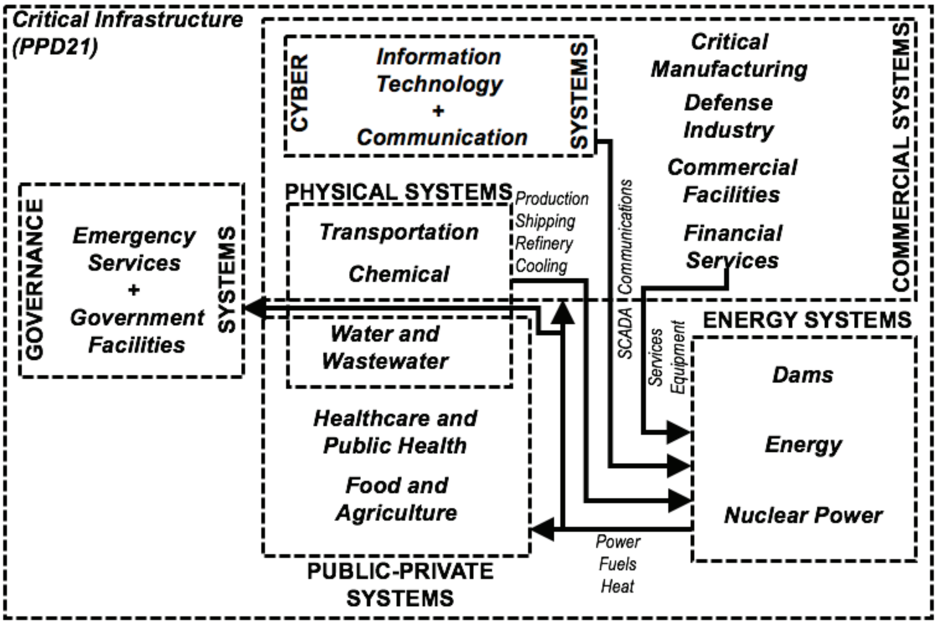

NATO End-Users
Carriers (public and private)
Primary energy sources (PES) for NATO operations
Individual/ Local
NATO military installations
Prosumers, Wind & Solar
NATO national critical infrastructure
National electricity systems
Domestic coal, gas & oil
Hydro power
Biomass
National
NATO national pipeline systems (GRPS, ICPS, NIPS, NOPS, POPS, TUPS)
National PES mix
NATO regional critical infrastructure
Regional
NATO regional pipeline systems (CEPS)
Regional markets
Global
Forward deployed units
Global nuclear industry and fuel cycle
Globally traded coal, gas & oil

Figure 1: Critical energy systems that guarantee NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements

Source: Adapted from GEA (Cherp et al., 2012)

Beyond the 13 NATO member states listed above (primarily located in Western and Southern Europe, old NATO), there are no NATO pipelines fulfilling the NATO fuel requirements in Central and Eastern Europe, resulting in a regional 'fuel desert' within the Alliance. As a result, NATO fuel requirements across new NATO (the Central European nations that joined NATO after the Cold War was over) are supplied from national primary energy sources (PES) by entities that are more often than not, privately owned and operated. New NATO in Central and Eastern Europe is, thus, not only still an underdeveloped theater, but also a geographic area where the refinery, bulk transport, storage, transit and delivery of fuels relies on privately owned and operated energy suppliers that are focused more on profit than security. This is of major concern for NATO, whose future operations may be hindered as a result of failure by privately owned critical infrastructure to fulfill NATO fuel and energy requirements.

It is important to note that the critical energy systems that guarantee NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements (as seen in Figure 1) also rely on secondary and tertiary critical infrastructure "that supports primary mission essential functions" (Presidential Policy Directive 21, 2013). These include critical functions played by the chemical sector, commercial facilities sector, communications sector, critical manufacturing sector, defense industrial base sector, emergency services sector, financial services sector, food and agriculture sector, government facilities sector, healthcare and public health sector, information technology sector, transportation systems sector, and the water and wastewater systems sector (see figure 2). Due to dependability of energy systems on these governance, commercial and public-private systems, a cyber attack against any sector listed above could also negatively impact NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements.



Source: Adapted from PPD21 Critical Infrastructure Areas and Interactions (Presidential Policy Directive 21, 2013)

Figure 2: Interaction between critical energy systems and other critical infrastructure that could impact impact NATO's assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements

4.2 From what threats to protect?

The critical energy systems—and the secondary and tertiary systems they depend on—delineated in the previous section are vulnerable to numerous cyber threats from both state and non-state actors. This ability of state and non-state adversaries to negatively affect military operations through a hybrid warfare posture is of major concern for the Alliance. Potential state adversaries of NATO include nation states like Russia, China and Iran, with advanced cyber capabilities and knowledge of CEI of NATO member states; while non-state adversaries include terrorist groups, criminal organizations and other entities that "have one common element, namely the willingness of the non-state actors to [hurt the United States], be it for monetary or ideological reasons"(Czosseck and Ziolkowski, 2013: 25).

Russia has been pursuing "a policy of all-out assertion (Eastern flank, the Mediterranean Sea, Syria, the Balkans) and across all domains" (Macron, 2017: 41). In the cyber domain, it remains "a highly capable and effective adversary, integrating cyber espionage, attack, and influence operations to achieve its political and military objectives" and "is now staging cyber attack assets to allow it to disrupt or damage US civilian and military infrastructure during a crisis and poses a significant cyber influence threat" (Coats, 2019: 5). Russia also has the ability "to execute cyber attacks in the United States [and other NATO member states] that generate localized, temporary disruptive effects on critical infrastructure—such as disrupting an electrical distribution network for at least a few hours—similar to those demonstrated in Ukraine in 2015 and 2016" and is currently mapping the critical infrastructure of the United States and its allies "with the long-term goal of being able to cause substantial damage" (Coats, 2019: 6). In March of 2018, for example, the United States Department of Homeland Security and the Federal Bureau of Investigations issued a joint technical alert revealing that Russian Government cyber activities are targeting energy, nuclear and other critical infrastructure sectors (CISA, 2018). The report states that Russian advanced persistent threat activities are specifically targeting critical energy systems operating in English, French, German, Italian and Spanish languages—primary languages of the United States and key European NATO allies—with the goal of identifying vulnerabilities pertaining to Industrial Control Systems (ICS). With regards to secondary and tertiary systems, the communication sector, in particular, has been vulnerable to Russian malign influence and propaganda. China presents "a persistent cyber espionage threat and a growing attack threat to [NATO] core military and critical infrastructure systems" and "has the ability to launch cyber attacks that cause localized, temporary disruptive effects on critical infrastructure—such as disruption of a natural gas pipeline for days to weeks—in the United States" and other NATO member states (Coats, 2019: 5). It also controls much of the manufacturing sector for equipment used by critical infrastructure in NATO member states, or in states that supply energy to NATO forward deployed units. Iran has also been preparing “for cyber attacks against the United States and our allies [and] it is capable of causing localized, temporary disruptive effects—such as disrupting a large company’s corporate networks for days to weeks—similar to its data deletion attacks against dozens of Saudi governmental and private-sector networks in late 2016 and early 2017” (Coats, 2019: 6). Finally, non-state adversaries, such as terrorist groups, criminal organizations and other entities have a “growing availability and use of publicly and commercially available cyber tools” that could be used against the critical infrastructure of NATO member states (Coats, 2019: 7).

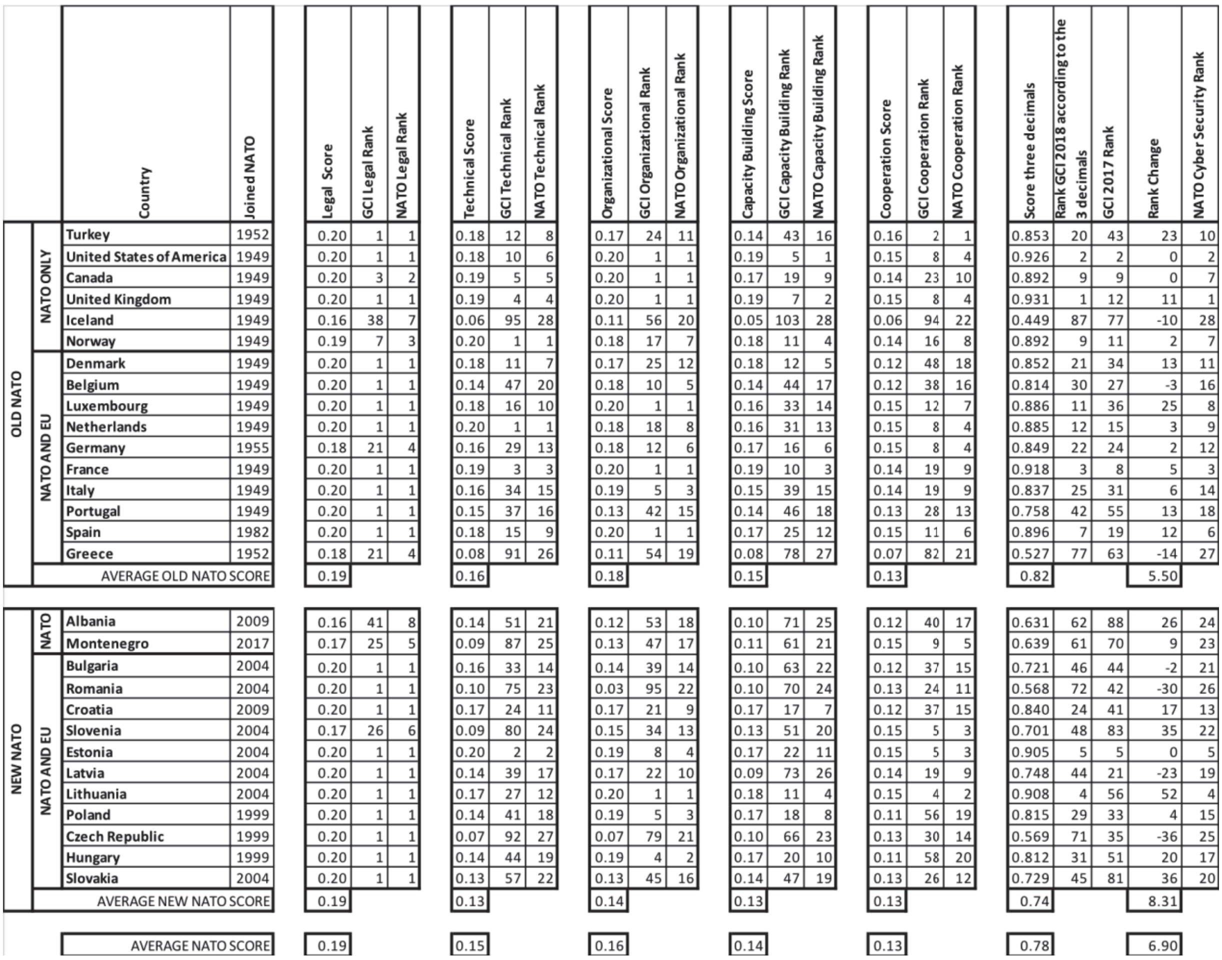
These threats work either independently, or jointly, to counter NATO capabilities and operations. Situations where states are "outsourcing cyber attacks to non-attributable third parties, including criminal organisations" (Farwell and Rohozinski, 2011: 23) have also been identified, making attribution of cyber attacks harder for NATO and its member states. This means that the ability of NATO to guarantee access to affordable and acceptable supplies of energy and the protection and delivery of sufficient energy to meet mission essential requirements is hindered by a variety of hybrid threats in cyberspace with varying capabilities and hostile intent—where attribution to specific state or non-state actors is critical in determining both.

4.3 By what means to protect?

At the 2016 NATO Summit in Warsaw, the Heads of State and Government agreed “to develop NATO's capacity to support national authorities in protecting critical infrastructure, as well as enhancing their resilience against energy supply disruptions that could affect national and collective defense, including hybrid and cyber threats. In this context, [NATO] will include energy security considerations in training, exercises, and advance planning” (NATO, 2016: 135). At the 2018 NATO Summit in Brussels, the Heads of State and Government further agreed that “energy security plays an important role in our common security. A stable and reliable energy supply, the diversification of routes, suppliers and energy resources, and the interconnectivity of energy networks are of critical importance and increase [NATO] resilience against political and economic pressure. While these issues are primarily the responsibility of national authorities, energy developments can have significant political and security implications for Allies and also affect [NATO] partners. Consequently, [NATO] will continue regular Allied consultations on issues related to energy security. [NATO] will refine [its] role in energy security in accordance with established principles and guidelines, and continue to develop NATO’s capacity to support national authorities in protecting critical infrastructure, including against malicious hybrid and cyber activity” (NATO Heads of State and Government, 2018). In other words, there is an agreement among the member states that NATO as an alliance must have a clear role in guaranteeing its assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements.

In the 21st Century, NATO has the prerogative to ensure the critical energy infrastructure protection (CEIP) in the cyber domain of its member states and partners supplying energy for NATO forward military operations. Cyber threats against CEI differ, however, from nation to nation; and so do the capabilities of these nations to defend their CEI against these threats. For example, energy systems fulfilling NATO fuel requirements in new NATO (Central Europe) are also more susceptible, on average, to successful cyber-attacks than in old NATO. This is exemplified by the International Telecommunication Union’s (ITU) Global Cybersecurity Index (GCI), which ranks the preparedness of states to tackle cyber threats (ITU, 2019). On average, new NATO is less prepared to tackle cyber threats than old NATO both overall and across three of the five GCI categories - technical, organizational, and cooperation. Both old and new NATO ranked equally on the legal and capacity building categories (as seen in table 1). What is of interest here is that some new NATO member states (like Estonia and Lithuania) are more prepared than most old NATO member states to tackle cyber threats. This indicates that national preparedness to respond to cyber threats is less about means (financial power and access to resources) and more about ways and ends (strategies and political will).

Table 1: Preparedness of New NATO and Old NATO to respond to cyber threats against Critical Energy Infrastructure



		Country	Joined NATO	Legal Score	GCI Legal Rank	NATO Legal Rank	Technical Score	GCI Technical Rank	NATO Technical Rank	Organizational Score	GCI Organizational Rank	NATO Organizational Rank	Capacity Building Score	GCI Capacity Building Rank	NATO Capacity Building Rank	Cooperation Score	GCI Cooperation Rank	NATO Cooperation Rank	Score three decimals	Rank GCI 2018 according to the 3 decimals	GCI 2017 Rank	Rank Change	NATO Cyber Security Rank
OLD NATO	NATO ONLY	Turkey	1952	0.20	1	1	0.18	12	8	0.17	24	11	0.14	43	16	0.16	2	1	0.853	20	43	23	10
		United States of America	1949	0.20	1	1	0.18	10	6	0.20	1	1	0.19	5	1	0.15	8	4	0.926	2	2	0	2
		Canada	1949	0.20	3	2	0.19	5	5	0.20	1	1	0.17	19	9	0.14	23	10	0.892	9	9	0	7
		United Kingdom	1949	0.20	1	1	0.19	4	4	0.20	1	1	0.19	7	2	0.15	8	4	0.931	1	12	11	1
		Iceland	1949	0.16	38	7	0.06	95	28	0.11	56	20	0.05	103	28	0.06	94	22	0.449	87	77	-10	28
		Norway	1949	0.19	7	3	0.20	1	1	0.18	17	7	0.18	11	4	0.14	16	8	0.892	9	11	2	7
	NATO AND EU	Denmark	1949	0.20	1	1	0.18	11	7	0.17	25	12	0.18	12	5	0.12	48	18	0.852	21	34	13	11
		Belgium	1949	0.20	1	1	0.14	47	20	0.18	10	5	0.14	44	17	0.12	38	16	0.814	30	27	-3	16
		Luxembourg	1949	0.20	1	1	0.18	16	10	0.20	1	1	0.16	33	14	0.15	12	7	0.886	11	36	25	8
		Netherlands	1949	0.20	1	1	0.20	1	1	0.18	18	8	0.16	31	13	0.15	8	4	0.885	12	15	3	9
		Germany	1955	0.18	21	4	0.16	29	13	0.18	12	6	0.17	16	6	0.15	8	4	0.849	22	24	2	12
		France	1949	0.20	1	1	0.19	3	3	0.20	1	1	0.19	10	3	0.14	19	9	0.918	3	8	5	3
		Italy	1949	0.20	1	1	0.16	34	15	0.19	5	3	0.15	39	15	0.14	19	9	0.837	25	31	6	14
		Portugal	1949	0.20	1	1	0.15	37	16	0.13	42	15	0.14	46	18	0.13	28	13	0.758	42	55	13	18
		Spain	1982	0.20	1	1	0.18	15	9	0.20	1	1	0.17	25	12	0.15	11	6	0.896	7	19	12	6
		Greece	1952	0.18	21	4	0.08	91	26	0.11	54	19	0.08	78	27	0.07	82	21	0.527	77	63	-14	27
		AVERAGE OLD NATO SCORE		0.19			0.16			0.18			0.15			0.13			0.82			5.50	
NEW NATO	NATO	Albania	2009	0.16	41	8	0.14	51	21	0.12	53	18	0.10	71	25	0.12	40	17	0.631	62	88	26	24
		Montenegro	2017	0.17	25	5	0.09	87	25	0.13	47	17	0.11	61	21	0.15	9	5	0.639	61	70	9	23
	NATO AND EU	Bulgaria	2004	0.20	1	1	0.16	33	14	0.14	39	14	0.10	63	22	0.12	37	15	0.721	46	44	-2	21
		Romania	2004	0.20	1	1	0.10	75	23	0.03	95	22	0.10	70	24	0.13	24	11	0.568	72	42	-30	26
		Croatia	2009	0.20	1	1	0.17	24	11	0.17	21	9	0.17	17	7	0.12	37	15	0.840	24	41	17	13
		Slovenia	2004	0.17	26	6	0.09	80	24	0.15	34	13	0.13	51	20	0.15	5	3	0.701	48	83	35	22
		Estonia	2004	0.20	1	1	0.20	2	2	0.19	8	4	0.17	22	11	0.15	5	3	0.905	5	5	0	5
		Latvia	2004	0.20	1	1	0.14	39	17	0.17	22	10	0.09	73	26	0.14	19	9	0.748	44	21	-23	19
		Lithuania	2004	0.20	1	1	0.17	27	12	0.20	1	1	0.18	11	4	0.15	4	2	0.908	4	56	52	4
		Poland	1999	0.20	1	1	0.14	41	18	0.19	5	3	0.17	18	8	0.11	56	19	0.815	29	33	4	15
		Czech Republic	1999	0.20	1	1	0.07	92	27	0.07	79	21	0.10	66	23	0.13	30	14	0.569	71	35	-36	25
		Hungary	1999	0.20	1	1	0.14	44	19	0.19	4	2	0.17	20	10	0.11	58	20	0.812	31	51	20	17
		Slovakia	2004	0.20	1	1	0.13	57	22	0.13	45	16	0.14	47	19	0.13	26	12	0.729	45	81	36	20
		AVERAGE NEW NATO SCORE		0.19			0.13			0.14			0.13			0.13			0.74			8.31	
		AVERAGE NATO SCORE		0.19			0.15			0.16			0.14			0.13			0.78			6.90	

Source: Build with 2018 data from the International Telecommunication Union's (ITU) Global Cybersecurity Index (GCI) (ITU, 2019)

5. Conclusions

According to Buzan's securitization theory, "a policy problem becomes a security issue if an agent manages to cast it as an 'existential threat', or a 'supreme priority' which requires treatment and intervention by extraordinary means" (Leung et al., 2014: 2). This paper proposes that cyber threats against NATO's energy security represent an existential threat to NATO's military operations. We posit that, in practice, NATO defines energy security as the assured access to affordable and acceptable supplies of energy and the ability to protect and deliver sufficient energy to meet mission essential requirements. Cyber threats to NATO's energy security, and particularly to interconnected liquid fuel supply chains and grids, have the capacity to directly impact NATO's enhanced Forward Presence and forward deployed units. At the national level, a cyber-attack against the critical infrastructure of NATO member states "might be regarded as armed aggression, due to [...] scale and severity. A major cyberattack may, given the damage it could cause, justify invoking legitimate defence under Article 51 of the UN Charter" (Macron, 2017: 33). But NATO member states may also rely on NATO's Article 5 to defend themselves against adversaries (Flamant and Parrein, 2017: 53), meaning that NATO must be able to provide "an adequate response" (Flamant and Parrein, 2017: 28) that guarantees uninterrupted access to affordable and acceptable supplies of energy, which are needed to meet mission essential requirements. Future research by the NATO Science and Technology Organization is needed to develop prototype-operating models providing early warning on cyber-attacks against NATO's energy security. For this to be achieved, however, more research is needed to asses less known vulnerabilities, such as cyber threats to land and sea-based energy chokepoints, as well as on NATO/Partner cyber requirements and capabilities, CEI interoperability issues, etc. that impact the enhanced Forward Presence mission. Additionally, more research is needed to assess the cyber threats to secondary and tertiary critical systems that impact NATO's operational energy (OE) security, for example, on the ability of cyber-attacks to deter contract trucks from supporting the Alliance's mission (also known as the white truck/green truck dilemma) in forward deployed environments.

References

Baldwin DA (1997) The concept of security. *Review of international studies* 23(1). Cambridge University Press: 5–26.

Bridge G, Bouzarovski S, Bradshaw M, et al. (2013) Geographies of energy transition: Space, place and the low-carbon economy. *Energy policy* 53. Elsevier: 331–340.

Bryant WD (2016) Mission assurance through integrated cyber defense. *Air & Space Power Journal* 30(4). Air Force Research Institute: 5–18.

Byres EJ and Eng P (2009) Cyber security and the pipeline control system. *Pipeline & Gas Journal*. tofinosecurity.com. Available at: https://www.tofinosecurity.com/sites/default/files/Cyber_Security_and_The_Pipeline_PGJ_Feb_2009.pdf.

Centre APER (2007) A Quest for Energy Security in the 21st Century: Resources and Constraints. Asia Pacific Energy Research Centre Tokyo.

Cherp A and Jewell J (2014) The concept of energy security: Beyond the four As. *Energy policy* 75. Elsevier: 415–421.

Cherp A, Adenikinju A, Goldthau A, et al. (2012) Energy and security. Cambridge University Press and IIASA. Available at: http://pure.iiasa.ac.at/id/eprint/10062/1/GEA%20Chapter%205%20Energy%20and%20Security.pdf.

Chester L (2010/2) Conceptualising energy security and making explicit its polysemic nature. *Energy policy* 38(2): 887–895.

Chi S-D, Park JS, Jung K-C, et al. (2001) Network Security Modeling and Cyber Attack Simulation Methodology. In: *Information Security and Privacy*, 2001, pp. 320–333. Springer Berlin Heidelberg.

CISA (2018) *Russian Government Cyber Activity Targeting Energy and Other Critical Infrastructure Sectors*. Alert (TA18-074A), 15 March. Department of Homeland Security. Available at: https://www.us-cert.gov/ncas/alerts/TA18-074A.

Coats DR (2019) *Worldwide Threat Assessment of the US Intelligence Community*. ODNI. Available at: https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf.

Czosseck C and Ziolkowski K (2013) State actors and their proxies in cyberspace. *Peacetime Regime for State Activities in*. Citeseer. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.645.1982&rep=rep1&type=pdf#page=36.

Deane JP, Gracceva F, Chiodi A, et al. (2015) Assessing power system security. A framework and a multi model approach. *International Journal of Electrical Power & Energy Systems* 73. Elsevier: 283–297.

Farwell JP and Rohozinski R (2011) Stuxnet and the Future of Cyber War. *Survival* 53(1). Routledge: 23–40.

Flamant R and Parrein P-J (2017) *The Strategic Vision for Defence*. D-2017/9376/1. Belgian Ministry of Defence.

Garrett RK (1993) *Is a Single Fuel on the Battlefield Still a Viable Option?* 94-07540. National Defense University. Available at: https://apps.dtic.mil/docs/citations/ADA276757.

Geels FW (2014) Regime Resistance against Low-Carbon Transitions: Introducing Politics and Power into the Multi-Level Perspective. *Theory, Culture & Society* 31(5). SAGE Publications Ltd: 21–40.

Goldthau A and Sovacool BK (2012) The uniqueness of the energy security, justice, and governance problem. *Energy policy* 41. Elsevier: 232–240.

Hybrid COE (2020) What is Hybrid CoE? Available at: https://www.hybridcoe.fi/what-is-hybridcoe/ (accessed 25 January 2020).

IBM (2015) *Security attacks on industrial control systems*. October. IBM Security. Available at: https://www.ibm.com/downloads/cas/4PE7DQ3G (accessed 15 March 2019).

IEA (2019) Energy Security. Available at: https://www.iea.org/topics/energysecurity/ (accessed 25 August 2019).

ITU (2019) *Global Cybersecurity Index (GCI)*. 978-92-61-28191-5. ITU.

Khamashuridze Z (2008) Energy Security and NATO: Any Role for the Alliance? *Connections* 7(4). JSTOR: 43–58.

Kushner D (2013) The real story of stuxnet. *IEEE Spectrum* 3(50): 48–53.

Leung GCK, Cherp A, Jewell J, et al. (2014) Securitization of energy supply chains in China. *Applied energy* 123. Elsevier: 316–326.

Macron E (2017) *Defence and National Security Strategic Review 2017*. Government of France.

National Defense Authorization Act for Fiscal Year 2018 (2017) Public Law 115–91. Available at: https://www.congress.gov/115/plaws/publ91/PLAW-115publ91.pdf.

NATO (2016) Joint declaration by the heads of state and government. Available at: https://franceintheus.org/spip.php?article668.

NATO (2017) NATO Pipeline System. Available at: https://www.nato.int/cps/en/natohq/topics_56600.htm?selectedLocale=ru (accessed 8 September 2019).

NATO Heads of State and Government (2008) Bucharest Summit Declaration. Available at: https://www.nato.int/cps/en/natolive/official_texts_8443.htm (accessed 1 September 2019).

NATO Heads of State and Government (2018) Brussels Summit Declaration. Available at: https://www.nato.int/cps/en/natohq/official_texts_156624.htm?selectedLocale=en (accessed 19 September 2019).

Ostrom E (2007) A diagnostic approach for going beyond panaceas. *Proceedings of the National Academy of Sciences of the United States of America* 104(39): 15181–15187.

Presidential Policy Directive 21 (2013) Critical infrastructure security and resilience. PPD-21, Released February 12, 2013.

Sullivan JE and Kamensky D (2017) How cyber-attacks in Ukraine show the vulnerability of the U.S. power grid. *Electricity Journal* 30(3). Elsevier: 30–35.

van den Hoven J, Blaauw M, Pieters W, et al. (2018) Privacy and Information Technology. Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2018. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/sum2018/entries/it-privacy/.

Yergin D (1988) Energy Security in the 1990s. *Foreign affairs* 67. HeinOnline: 110.

Yergin D (2006) Ensuring Energy Security. *Foreign affairs* 85(2). Council on Foreign Relations: 69–82.

Cyber Counterintelligence: Some Contours towards the Academic Research Agenda

Petrus Duvenage, Victor Jaquire and Sebastian von Solms
University of Johannesburg, South Africa
duvenage@live.co.za
jaquire@gmail.com
basievs@uj.ac.za

DOI: 10.34190/EWS.20.053

Abstract: In little more than a decade, Cyber Counterintelligence (CCI) has evolved from a near obscure academic research topic to a distinctive, interdisciplinary field of enquiry. Although in various respects still in its academic infancy, CCI now boasts a significant corpus of academic literature. At this juncture, CCI's progression as an academic sub-discipline can benefit from selectively taking stock of what has been done, reflecting on research challenges being experienced and prioritising some areas for future research. This paper advances such a high-level proposition on the CCI research agenda. This proposition consists of four parts. Firstly, we explain and qualify our approach to addressing the paper's theme. Since a CCI literature review has been undertaken in three recently published research works (Duvenage, Jaquire & von Solms 2018 a-b, Duvenage 2019), we categorically qualify our approach as not attempting a conventional literature review. Secondly, the paper examines CCI's research context and outline research challenges. Thirdly, the paper selectively 'contours' existing research. These 'contours' comprise some salient topics addressed in surveyed CCI literature. This assessment thus moves beyond a conventional literature review and instead follows a broad-brush approach that highlights some main areas of CCI research thus far conducted. Fourthly, the paper proposes some 'contours' for the CCI research agenda going forward. These 'contours' are presented as general proposals on priority areas for further research. It is hoped that these proposals will aid the focusing of academic CCI research conducted outside the realm of state security structures. The paper concludes with a summary of findings.

Keywords: Cyber security, cyber counterintelligence, risk management, offensive cybersecurity, threat intelligence

1. Introduction

In little more than a decade, Cyber Counterintelligence (CCI) has evolved from a near obscure academic research topic to a distinctive, interdisciplinary field of enquiry boasting a discernible corpus of academic literature. However, in comparison to other closely related fields such as cyber warfare and cyber intelligence, CCI's body of published academic literature is fast growing but still limited. This limitation in literature should be viewed within CCI's research context that includes CCI thus far being a niche academic field mainly focused on by those associated with nation-state security and related structures. Although CCI research by nation-state security structures in some areas is typically classified, there is abundant space and rich opportunities for public academic work in this field.

At this juncture, CCI's progression as an academic research field can therefore benefit from assessing the CCI research context, taking stock of what research has been done and by prioritising some areas for future research. Collectively, the trends characterising existing research and areas for future research constitute the 'contours' of the CCI academic research agenda. Accordingly, this paper advances such a proposition comprising contours of (i) some salient topics in existing research and (ii) proposed areas for further research. To this end, the rest of paper is structured as follow:

- Section 2 explains the approach to, and qualifies, the paper.
- Section 3 examines the CCI academic research context.
- Section 4 contours existing CCI academic research by means of describing some salient (research) themes.
- Section 5 proposes some contours (areas) for future research.

This section highlighted the need for a proposition on the CCI academic research agenda. We explained what is meant by the 'contours' of the research agenda and then proceeded to present the paper's structural outline. The next section describes and qualifies the approach we followed in designing our proposition on the research agenda.

2. Approach and qualification to the CCI academic agenda's design

As suggested by its title, this paper does not purport to advance comprehensive, detailed propositions on a *future research agenda*. In a nutshell, our proposition in this paper must *not* been seen as *an inventory*, but only as *some items on an agenda*. Accordingly, our aim is to provide high-level outlines (contours) that can serve as a soundboard for further stimulating the academic discourse.

In a similar vein, we employ a *broad-brush approach* to the *contouring* of *existing* CCI *academic research* through the identification of salient research topics. It must be emphasised that this *contouring does not purport to in any way represent a literature view* (i.e. who, where, what and why) *as this would duplicate previous research*. Our contouring in this paper is thus informed by, but does not replicate, a CCI literature review presented at the 17th European Conference for Cyberwarfare and Security (Duvenage, Jaquire & von Solms 2018a) as subsequently developed to an article in the *Journal for Information Warfare* (Duvenage, Jaquire & von Solms 2018b) and further expanded on in a recently completed doctoral thesis (Duvenage 2019). These literature reviews, and thus our contouring in this paper, are restricted to publically available works in the English language, that has CCI as a primary focus and were featured in selected platforms, namely: Scopus, EBSCO, Institute of Electrical and Electronics Engineers (I.E.E.E.E.) Explore, Springer Link, Google Scholar and Proquest. In line with the paper's aim to advance a proposition on the *academic* research agenda, the scope of literature considered was further narrowed to peer-reviewed contributions (e.g. articles, papers, post-graduate research outcomes) and books.

In this section, we discussed our approach and methodology to the design of the CCI research agenda. Building on this discussion, the next section explains CCI's research context.

3. Cyber Counterintelligence’s Academic Research Context

A credible proposition on the CCI research agenda should of course consider CCI's research context. The research context not only shaped the contours of existing research, but impacts the challenges for, and contours of, the future research agenda. Our discussion of the research context comprises of three parts namely (i) the criteria to which CCI as an emerging sub-discipline ought to aspire; (ii) challenges to CCI’s academic progression; and (iii) the implications of the said criteria and challenges for future research.

3.1 CCI and the criteria for recognition as a distinctive academic sub-discipline

While a distinctive research field, CCI is not in all respects a separate academic discipline as yet. Rather, and as will be elaborated on later in the paper, CCI is a Counterintelligence (CI) sub-discipline (subset) which overlaps with *inter alia* Intelligence Studies and Cyber Intelligence (Lee 2014; van Niekerk, Ramluckan & Duvenage 2019). Nonetheless, CCI’s consolidation as an academic sub-discipline and distinctive, interdisciplinary field could profit from CCI aspiring to attain academic disciplines’ attributes. In the main, an academic discipline has the following attributes (Krishnan 2009):

- A particular research object (that is not necessary exclusive and that could be shared with other disciplines);
- an accumulated body of specialist knowledge;
- theories and concepts that effectively structure accumulated knowledge;
- terminologies and/or “a specific technical language adjusted to their research object”;
- research methods congruent with its specific research requirements; and
- "Institutional manifestation" in the form of subjects taught at tertiary institutions such as universities and colleges.

3.2 Challenges to CCI’s academic evolvement

CCI’s evolvement has been, and for the immediate future is likely, to be influenced by three challenges namely (i) CCI’s complexity and technological dimension; (ii) restrictions in material publically available for academic research and outputs; and (iii) CCI’s perceived status a niche field. Since these challenges are interwoven, we now continue by discussing the challenges not separately but in an integrated manner.

Contrary to what may be expected, an appreciation and description of these three challenges does not start with CCI *per* se. CCI can be defined as that *sub-discipline of counterin*telligence (CI) "aimed at detecting, deterring, preventing, degrading, exploiting and neutralisation adversarial attempts to collect, alter or in any other way breach the C-I-A of valued information assets through cyber means” and proactively degrade such adversaries' intelligence capabilities (expanded from Duvenage, von Solms & Corregedor 2016 - emphasis added). From this definition, its flows that an explanation of CCI's research challenges should start with appraising counterintelligence (CI) more generally.

Over decades, CI has been acknowledged – in practice and academically – as arguably the most intricate and complex of the Intelligence disciplines (c*f*. Sims 2014, Godson 2011, Sims 2009). The following assertion by Miler (1980) four decades ago still rings true:

> *It is not easy, nor can one feel confident, to re-enter this world where, it has been said, the tortuous logic of counterintelligence prevails ... Unfortunately, there seems to be no easy way to explain counterintelligence ... Because effective counter-intelligence is a combination of so many aspects of the intelligence business and other political, military, economic and societal factors.*

Adding to this complexity, counterintelligence (CI) could in the academic realm be viewed as 'secretive' field reserved mainly for current and former state security practitioners. Published, freely available academic research is dominated by historical case studies of public known events (Prunckun 2012). With a few notable exceptions (Prunckun 2012, 2019; Duvenage 2010), little attention is given to research advancing CI as an academic discipline. CI's complexity and limitations in research material affect academic research and ultimately tertiary education. In this regard, a former Director of Counterintelligence in the United States, Michelle van Cleave (2009) asserts:

> *Over 40 years have passed, and intelligence studies are now a part of most serious international relations departments and are integrated into the curriculum at our nation's war colleges. But the role of counterintelligence remains little known or understood among scholars or practitioners of national security studies and policymaking."*

Practically, the above presents as challenges that influence both CI research interest and output. Although CI academic work was and remains limited, CCI's development has been unpinned by the application of time-tested CI concepts and constructs. Therefore, it is doubtful, if not impossible to academically understand and explain CCI's complexities without a sound appreciation of CI. For this reason, application of CI constructs to CCI runs as a thread throughout all three phases of CCI progression, namely (Duvenage, Jaquire & von Solms 2018 a, b):

- Foundational phase (pre-2009)
- Emergence as an academic research theme (2009 -2012)
- Crystallisation as a distinctive academic sub-discipline (2012 – present).

Throughout all these phases, the addition to CI of CCI's defining technical dimension ('cyber') has substantially added to its complexity. In other words, CCI is not only complex because CI is complex – it is substantially more complex because its technical (cyber) dimension further adds a twofold complexity. Firstly, owing to its cyber dimension, CCI has to compute and reflect in its academic research the fast evolving techno-societal landscape. While challenging, high-level research on this landscape and its impact on CCI is indeed academically feasible. However, open academic CCI research is for self-evident reasons unlikely to drill down to leading-edge 'technical CCI gold'. The intricacies, mechanics and even existence of some CCI tools and techniques, are by their very nature closely protected and classified.

Secondly, CCI's 'cyber' dimension adds several technical specialisations to CI. Bardin (2011) rightly describes CCI as "a conglomerate of several disciplines and skills" each requiring "its own specialized training." Therefore, the development of CCI specialists [and by extension academics] takes a long time (Black 2014). Generally, state security structures have a divide between CI/CCI's technical versus non-technical professional specialisations. Consequently, a significant body of CCI-relevant research and training material aims to both bridge this gap and enable specialisation. For the most part, however, also this research is classified and thus not available to the 'public' academic discourse. In a similar vein, and outside the USA, CCI is of mostly taught as an academic subject within the confines nation states' security apparatus.

Against the above background, CCI is understandably perceived as a niche academic field. Although publishing researchers' affiliation it is not always stated, it is clear from surveyed literature that the overwhelming bulk of

CCI published academic research was authored by (former/current) members of statutory security structures and scholars enrolled at research institutions with strong links to statutory security structures. In the authors' view, this further reinforces the perception of CCI as a niche academic field.

The preceding discussion substantiates our assertion at the start of this subsection (Section 3.2) on the threefold challenge to CCI evolvement as an academic sub-discipline, namely (i) CCI's complexity; (ii) restriction in material publically available for academic research and outputs; and (iii) CCI' perceived status a niche field.

3.3 Implications for future CCI academic research

The challenges that CCI academic research faces (Section 3.2) and the requirements for an academic sub-discipline (Section 3.1) hold the following threefold implications for the future CCI research agenda, namely:

1. Going forward, CCI research needs to be directed to those areas that can aid CCI's consolidation as an academic sub-discipline.
2. The identification of future research areas should be feasible in that they compute CCI's complexity and limitations in research material. Plainly put, the areas identified should be 'researchable' through open, unclassified research material.
3. Still being a niche academic field, future research should be sharply focused and duly consider the work that has already been done.

In line with the last requirement 3.3 (3), the next section contours existing academic research.

4. Contouring existing CCI academic research

This section contours existing CCI academic research by highlighting some salient research topics. As noted in Section 2, this paper does not purport to be informed by a rigorous and systematic literature review. Instead the paper follows a less a less formalistic and selective approach which is based on literature that came to the attention as part of a CCI research project at the University of Johannesburg (Duvenage, Jaquire & von Solms 2018 a-b; Duvenage 2019). Furthermore, works we cite in this section to substantiate assertions on research topics are merely illustrative and not by any measure an exhaustive list. With these caveats, and taking a bird's eye view of the academic literature thus far, we identified the following as some salient and overlapping research themes:

4.1 Fundamental concepts and CCI's relation with other fields

Being a relatively young field of academic enquiry, a substantial of academic research was, and remains, focussed on fundamental CCI concepts. Also works that primarily address other CCI research themes, devote significant attention to first unpacking fundamental concepts such as definitions, CCI's offensive and defensive dimensions as well as (in broad terms) CCI's interlock with fields such as cybersecurity, intelligence, counterintelligence, cybersecurity and, to a lesser degree, cyber warfare (Stone & Tucker 1988, Stone & Bluitt 1993, Davey & Armstrong 2002, French & Kim 2009, Knowles 2013, Duvenage, von Solms & Corregidor 2015, Boawn 2014, Fieber 2015, Duvenage & von Solms 2014-2015, O'Connor 2017).

4.2 The threat landscape, posture weaknesses and the case for effective CCI

Albeit to varying degrees, the vast majority of published research items surveyed describe the threat landscape. While reference is made to the threats confronting non-state actors, descriptions of the threat landscape are mostly presented from nation-states' (notably the USA's) perspective. Threat actors most often described are the People's Republic of China, Russia, North Korea and Iran (Justiano 2017, Boawn 2014, Fieber 2015). Typically, researchers use descriptions of the threat landscape to argue the case for a coherent and balanced CCI approach in which the offensive dimension features more prominently. (Snider 1987, Davey & Armstrong 2002, French & Kim 2009, Ferguson 2012, Sithole *et al* 2019) In some cases, the USA's intelligence posture is analysed and/or perceived systemic weaknesses identified to substantiate calls for effective CCI.

4.3 Cyber espionage as one of CCI's signature role

CCI, to recapitulate Subsection 3.2, aims at detecting, deterring, preventing, degrading, exploiting and neutralisation adversarial attempts to collect, alter or in any other way breach the C-I-A of valued information assets through cyber means and proactively degrade such adversaries. Inferring from this definition, the countering of cyber espionage is indeed an important, but only one, of CCI's signature roles. Yet, in comparison to CCI's other signature roles (such as denial/counter-denial and deception/counter-deception), counterespionage is by far the predominant research topic. Notable, exceptions in this regard include Bodmer *et al* (2012) and Stech & Heckman (2018).

4.4 Theory, frameworks and models

Fundamental concepts (mention in Section 4.2) are of course elementary building blocks of CCI theory. A substantial and fast-growing body of CCI research is moving beyond elementary concepts to advance 'higher order' conceptual constructs. These constructs vary greatly in nature and focus. On the macro-level contributions range from postulations to structure CCI as an academic field (Duvenage 2019) to a framework for a CCI maturity model (Jaquire & von Solms 2017a-c, Jaquire 2018). On the meso- and praxis levels, frameworks and models have been advanced for, to name but a few: the CCI posture (Stech & Heckman 2018, Duvenage, Jaquire & von Solms 2019), the CCI process (Fieber 2015), awareness and training (Black 2014, Sithole *et al* 2019), CCI all-source information flow and analysis (Sigholm & Bang 2013) and various valuable propositions on CCI's integration with hybrid warfare (Justiniano 2017).

4.5 CCI's role in safeguarding ICT and other critical infrastructure

CCI's role in safeguarding critical infrastructure is undoubtedly one of the most enduring research areas. With some exceptions, however, the protection of critical infrastructure is addressed as a subtheme and not the studies' primary focus (Stone & Bluitt 1993, Davey & Armstrong 2002, French & Kim 2009, Knowles 2013, Ferguson 2012).

4.6 CCI and the insider threat

Similar to the safeguarding of critical infrastructure, the insider threat is an enduring CCI research theme. With some exceptions, also the insider threat is addressed as a subtheme and is not the studies' primary focus.

4.7 Technical architecture and tools

Regrettably, but understandably, published academic research on the 'technical' CCI is limited to on the surface, generic contentions on architecture (e.g. honeynets), tools (e.g. sock puppets) and taxonomies of tools. Examples of published work addressing these aspects include Bodmer *et al* (2012), Loreto (2012) as well as Duvenage, Jaquire & von Solms (2016, 2019).

In this section, some salient topics of CCI research were discussed. Based on this discussion, and considering CCI’s research context (section 3), the next section proposes priority areas for further research.

5. Propositions on the CCI academic research agenda going forward

We noted in Section 3.3 that our proposition on the future CCI research agenda have to (i) consider existing CCI research, (ii) reflect CCI's research context (including the limits imposed by unclassified source material), and (iii) contribute to CCI's consolidation as a distinctive academic sub-discipline. Contributing to such a consolidation further requires us to consider academic disciplines' attributes (discussed in subsection 3.1). Moving from this premise, we can now postulate priority areas for future research. To aid the reader and avoid excessive referencing, we repeat some points made in Subsection 3.1 on the attributes of academic disciplines. Considering the above, we identified the following as the ten priority areas for further CCI academic research:

5.1 Confluence of accelerating technologies – impact on, and implications for, CCI

The confluence of technologies - such as artificial intelligence (AI), the Internet of Things (IOT), social media and Big Data - is a two-edge sword (cf. Sims 2014). On the one hand, the effective use thereof is indispensable to an organisation’s defensive counterintelligence efforts (through for example sensing systems, biometrics, surveillance, AI-driven defences) and essential to bolstering its offensive CCI endeavour in areas such deception, influencing, offensive collection and degrading adversarial intelligence. On the other hand, the confluence of the same technologies is a power maximizer to adversarial intelligence and CI efforts (cf. Sims 2014). This of course not only applies to nation states and large corporates. Technologies asymmetrically empower smaller role-players with limited resources.

Current academic research barely scratches the surface of the impact and implications of this dichotomy on CCI. In the view of the authors, this area should be a foremost point on the CCI research agenda.

5.2 Further rise of the insider threat

Societal shifts and technological innovation are likely to fuel “a rise in insider espionage over the decade to come. With the empowerment of the individual comes his mobilisation for causes of all kinds” and his ability to exfiltrate data and information (Sims 2014). This paper’s section 4.6 has noted the ‘insider threat’ as being mostly addressed in research as a subtheme and not as CCI studies' primary focus. The importance of future academic research on the insider threat, specifically from a CCI perspective, can hardly be over-emphasised.

5.3 CCI and non-state actors

The bulk of research outlined in Section 4, takes the nation-state as the primary referent object of national security. Accordingly, CCI's *raison d'etre* is advancing and safeguarding the nation-state's interests. At least in as far as consulted academic literature is concerned, non-state actors as CCI's referent object is vastly unexplored. More generally, the application of CI by non-state actors such as business enterprises and terrorist groups has been researched (Bodmer *et al* 2012, Mobley 2008). There is no reason why this cannot be extended to non-state actors’ application of CCI to safeguard and advance their interests. CCI is increasingly being adopted by larger corporates (*The Economist* 2015, Panda Security Labs 2018). Therefore, the academic exploration of CCI as part of Business Studies and related fields has great potential.

5.4 CCI and its interlock with related fields such as information/cyberwarfare and cyber intelligence

In Section 3, we observed that comparatively little work has been done on the interlock between, on the one hand CCI and, on the other hand, information/cyber warfare and cyber intelligence. In terms of further research, cognisance should be taken of the rich and expansive body of information warfare literature that spans more than two decades. A few examples of outstanding works are: Molander, Riddle & Wilson (1996), Denning (1999), Kopp (2000), Hutchinson & Warren (2001), Jones, Kovacich & Luzwick (2002), Hutchinson (2006), Armistead (2004), Clarke & Knake (2010), Armistead (2010), van Niekerk & Maharaj (2011), Warren (2013), Andress & Winterfeld (2014), Janczewski & Caelli (2016) and Buchanan (2016). These works contain highly useful concepts and constructs that can be examined for application to CCI. Furthermore, "the interface and dynamics between CCI and information/cyber warfare presents fertile, yet conceptually intricate, ground for further research" (Duvenage 2019). In a similar vein, the interlock and dynamics between CCI and cyber intelligence has been barely explored and present fertile ground for further research (Boawn 2014; Lee 2014; van Niekerk, Ramluckan & Duvenage 2019).

5.5 CCI Awareness, Education and Training (AET)

Accelerating trends within the threat landscape, such as the rise in the number and intelligence activities of threat actors, are fuelling the demand for CCI professionals and the addition of CCI elements to cybersecurity awareness programmes (Sithole *et al* 2019, Black 2014). Highly specialised training, in especially offensive CCI, is likely to remain within the realm of statutory state security structures and associated training institutions. For obvious reasons, such training material is of a classified nature and excluded from published academic outputs. Nonetheless, there is ample room for academic research on, for example, CCI awareness programs and curricula for CCI tertiary education on graduate and post-graduate level. This can bolster CCI's consolidation as an academic sub-discipline in promoting "Institutional manifestation" in the form of subjects taught at tertiary institutions (Krishnan 2009).

5.6 Implications on CCI of national and international law, conventions and guidelines

At least normatively, state and *bona fide* non-state actors' CCI endeavours ought to be within the parameters of national laws as well as international law, conventions and guidelines. In literature surveyed, we could find no explicit reference to 'cyber counterintelligence' in such law, conventions and guidelines. The *Tallinn Manual*, for example, contains only one such implicit reference in its description of terminology (Schmitt 2013). However, the *Tallinn Manual* does deal explicitly with fields and concepts closely related to CCI such as cyber warfare and cyber espionage. Consequently, the implications on CCI conventions and guidelines present a near unexplored area for the research agenda.

5.7 Implications of governance frameworks and standards

To the knowledge of the authors, there is no existing governance framework, outside the statutory security environment, that categorical refer to CCI. The same applied to well-known and standard such as ISO and COBIT. There is, in fact little, if any, reference in governance frameworks and standards to counterintelligence more generally. The imperatives governance frameworks pose or should impose on CCI are therefore an important research theme. Equally important are research and propositions on expanding governance frameworks to include CCI.

5.8 Theories for, and of, CCI

The existence of theory is not only a prerequisite for a field being recognised as a distinctive sub-discipline. Good theory is also a perquisite for sound practice. Within Intelligence Studies, a distinction is made between 'theories *for* intelligence' and 'theories *of* intelligence'. Theories for intelligence are typically on the (meso- and praxis levels), and "relate immediately to the needs of practitioners" (Gill 2006). A theories or intelligence are developed to "help academics research intelligence, come to understand it, and better explain it" (Gill 2006). Observing on the symbioses between these two theory classes, Gill (2006) states: "In one sense there is no conflict between these two. A good theory of intelligence should, by definition, be useful for intelligence."

The forgoing distinction between theories can be extended also to CCI. This paper's Section 4, shows that within CCI research most theorisation consists of theories *for* CCI (praxis and meso-theories), with a few contributions on the macro-level pitched as contributions to theories *of* CCI. In both respects, this research is only scratching the theoretical surface and there is fertile ground for further theorisation). Likewise, there is a pressing need for an overarching grand (high level) theory that coherently binds CCI theory on lower levels of abstraction (i.e. meso and praxis). On whatever level, all CCI theory in surveyed literature is underpinned by the Realist school of thought (within Intelligence Relations). CCI theorisation can thus benefit from exploring the application of other schools of thought such as liberalism, constructivism, radicalism and post-structuralism.

5.9 History of CCI

A "distinguishable tradition" and/or an "identifiable history" is critical for professional self-identity, understanding the present and anticipating the future (Krishnan 2009). On the matter of professional, self-identify Caelli, Liu & Longley (2013) state:

> *For any discipline to be regarded as a professional undertaking by which its members may be treated as true "professionals", practitioners must clearly understand that discipline's history as well as the place and significance of that history in current practice as well as its relevance to available technologies and artefacts at the time.*

Like other academic subjects, such history could contribute to consolidating CCI as a distinctive sub-discipline. We could find no contribution in surveyed literature that aims to present CCI's history. Given CCI research context (Section 3), it would be difficult if not impossible to attempt a credible history from an international perspective. More realistically, CCI's history within specific organisation's (such as statutory security structures), particular countries or academic communities (such as the Western Anglophone community) could be topics for future research.

The history of CCI does not need to be confined to the evolvement of the sub-discipline itself. **Case studies** of CCI-relevant incidents in some respect constitute CCI's unfolding history of successes and failures.

5.10 Anticipatory and Futures research

At various junctures in this paper, we alluded to accelerating trends and driving forces that are shaping the demands on CCI to a degree nearly unimaginable two decades ago. On a strategic level, proactive CCI needs to consider possible, probable and preferred futures. On the operational, tactical and technical level, the outcome of anticipatory research could guide the proactive sharpening of CCI-related technology, techniques and procedures. While some academic work has been done which reflect on societal-technical trends within CI, no systematic futures or anticipatory research on CCI has come the authors' attention.

This section proposed some areas for future CCI research. The next section concludes the paper with a summary of findings and observations on CCI's research future.

6. Conclusion

This paper presented an overview of some trends ('contours') in existing CCI academic research. Taking into account (i) trends in existing research and (ii) CCI research context, we advanced a proposition on some contours of an agenda for future CCI academic research. We argued that the proposed research agenda could contribute significantly to CCI's consolidation as an academic sub-discipline. We found CCI to be academically vastly underexplored and as presenting fertile, yet conceptually intricate, ground for further research. The contours we presented as points on the agenda should not be construed as an attempt to advance an exhaustive list. Similarly, the contours of existing CCI research are merely some outlines of existing work. In both respects, these contours are intended as a tentative soundboard for pointing future research.

CCI's fast growing adoption by non-state actors and its growth as topic of tertiary instruction are sure to fuel the demand for such research. Therefore, CCI presents a practically relevant and rewarding research areas for current and prospective researchers interested in this field.

References

Armistead, E. L. (2010) *Information Operations Matters: Best Practices*, Potomac Books, Dulles, United States (US).

Armistead, L. (ed.) (2004) *Information Operations: Warfare and the Hard Reality of Soft Power*, Potomac Books, Washington, D.C., US.

Andress, J. & Winterfeld, S. (2014) *Cyber* warfare – *techniques, tactics and tools for security pra*ctitioners, second edition, Elsevier, Waltham, US.

Bardin, J. (2011) 'Ten commandments of cyber counterintelligence', *CSO Magazine* (online), accessed on 09/01/2013 at http://www.csoonline.com/article/2136458/identity-management/ten-commandments-of-cyber-counterintelligence.

Black, J.M. (2014). *The complexity of cyber counterintelligence training,* unpublished Master of Science dissertation, Utica College. US.

Boawn, D.L. (2014) *Cyber counterintelligence, defending the United States' information technology and communications critical infrastructure from Chinese threats*, unpublished master's dissertation, Utica College, New York, US.

Bodmer, S.A. et al. (2012) *Reverse deception – Organized cyber threat counter-exploitation*, McGraw-Hill, New York, US.

Buchanan, B. (2016) *The cybersecurity dilemma – Hacking, trust, and fear between nations*, C. Hurst & Co. Publishers, London, United Kingdom (UK).

Caelli W., Liu, V. & Longley, D. (2013) 'Background to the development of a curriculum for the history of "cyber" and "communications security" ' in Dodge R.C. & Futcher L. (eds.) *Information Assurance and Security Education and Training*, IFIP Advances in Information and Communication Technology, vol. 406, Springer, Berlin, Germany.

Davey, J. & Armstrong, H (2002) 'Dominating the attacker: Use of intelligence and counterintelligence in cyberwarfare' in *Journal of Information* Warfare, Vol. 2. Nr. 1.

Denning, D.E. (1999) *Information warfare and security*, Addison-Wesley, Boston, US.

Duvenage, P.C. (2019) *A conceptual framework for cyber counterintelligence*, unpublished PhD (DCom) thesis dissertation, University of Johannesburg, South Africa.

Duvenage, P.C. (2010) Open-source environmental scanning and risk assessment in the statutory counterespionage milieu, unpublished D.Phil. thesis, University of Pretoria, Pretoria, South Africa.

Duvenage, P.C., Jaquire, V.J. & von Solms, S.H. (2016) 'Conceptualising cyber counterintelligence – Two tentative building blocks' *in Published Proceedings of the 15th European Conference on Cyber Warfare and Security*, Munich, Germany, June. Available at http://adam.uj.ac. za/csi/docs/ECCWS2016_DJVS_PDF.pdf

Duvenage, P.C., Jaquire, V.J. & von Solms, S.H. (2018a) 'A selective literature review on cyber counterintelligence' in *Published Proceedings of the 17th European Conference on Cyber Warfare and Security*, Oslo, Norway, June.

Duvenage, P.C., Jaquire, V.J. & von Solms, S.H. (2018b) 'Towards a literature review on cyber counterintelligence' in *Journal of Information Warfare*, 17(4), 11-25.

Duvenage, P.C. & von Solms, S.H. (2014) 'Putting counterintelligence in cyber counterintelligence' in *Published Proceedings of the 13th European Conference on Cyber Warfare and Security*, Piraeus, Greece, July.

Duvenage, P.C. & von Solms. S.H. (2015) 'Cyber counterintelligence: Back to the future' in *Journal of Information Warfare*, 13(4):42–56. Available at http://adam.uj.ac.za/csi/docs/Journal%20of%20Information%20Warfare%20Duvenage_VonSolms%202015.pdf

Duvenage, P.C., von Solms, S.H. & Corregedor, M. (2015) 'The cyber counterintelligence process – A conceptual overview and theoretical proposition' in *Published Proceedings of the 14th European Conference on Cyber Warfare and Security*, Hatfield, UK, July. Available at http://adam.uj.ac.za/csi/docs/ECCWS2015_Duvenage%20Von%20Solms%20Corregedor.pdf

The Economist (2015) "*Counter-intelligence techniques may help firms protect themselves against cyber-attacks", [online],* http://www.economist.com/news/business/21662540-counter-intelligence-techniques-may-help-firms-protectthemselves

Ferguson C. J. (2012) *Increasing the effectiveness of U.S. counterintelligence: Domestic and international micro-restructuring initiatives to mitigate cyberespionage*, unpublished master's dissertation, US Naval Postgraduate School, California, US.

Fieber, T. J. (2015) *The Iranian computer network operations threat to U.S. critical infrastructures*, Master of Science (capstone project), Utica College, US.

French, G.S. & Kim, J. (2009) 'Acknowledging the revolution: The urgent need for cyber counterintelligence' in *National Intelligence Journal*, 1(1):71–90.

Godson, R. (2001) *Dirty tricks or trump cards – U.S. covert action and counterintelligence*, Transaction Publishers, New Brunswick, US.

Hutchinson, W. & Warren, M. (2001) 'Principles of information warfare', *Journal of Information Warfare* Volume, 1 (1): 1-6 IBM (2016) 2016 Cyber Security Intelligence Index, accessed at https://www.ibm.com/security/data-breach/threat-intelligence-index.html .

Janczewski, L.J. & Caelli, W. (2016) *Cyber conflicts and small states*, Routledge, New York, US.

Jaquire, V.J. (2018) *A framework for a cyber counterintelligence maturity model*, unpublished PhD (D.Com) thesis, University of Johannesburg, Johannesburg, South Africa.

Jaquire, V.J. & von Solms, S.H. (2017*a*) 'Towards a cyber counterintelligence maturity model' in *Published Proceedings of the 12th International Conference on Cyber Warfare and Security*, Wright State University, Air Force Institute of Technology, Dayton, US, March.

Jaquire, V.J. & von Solms, S.H. (2017*b*) 'Developing a cyber counterintelligence maturity model for developing countries' in *Published Proceedings of the 2017 IST–Africa Conference, Windhoek,* Namibia, May–June.

Jaquire, V.J. & von Solms, S.H. (2017c) 'Cultivating a cyber counterintelligence maturity model' in *Published Proceedings of the 16th European Conference on Cyber Warfare and Security*, Dublin, Ireland, June.

Jones, A., Kovacich, G.L. & Luzwick, P.G. (2002) *Global information warfare*. Auerbach, Boca Raton, US.

Justiniano, J.E. (2017) *Advancing the capacity of a theater special operations command (TSOC) to counter hybrid warfare threats in the cyber gray zone*, Master of Science (capstone project), Utica College, US.

Kopp, C. (2000) 'A fundamental paradigm of infowar', *Systems*, accessed on 2011/10/25 at http://www.ausairpower.net/OSR-0200.html.

Lee, R.M. (2014) "Cyber Counterintelligence: From Theory to Practice," *Tripwire Blog*, 5 May, [online], accessed 10 August 2016, https://www.tripwire.com/state-of-security/security-data-protection/cyber-counterintelligence-from-theory-to-practice/

Knowles, J. A. (2013). *Applying computer network operations for offensive counterintelligence*, Master of Science (capstone project), Utica College, US.

Loreto, J (2014) *The effectiveness of honeypots*, Master of Science (capstone project), Utica College, US.

Miler, N.S. (1980) 'What is counterintelligence – Discussants' in Godson, R. (ed.*), Intelligence requirements for the 1980s: Counterintelligence*. National Strategic Information Center, Washington D.C., US.

Mobley, B. W. (2008) *Terrorist group counterintelligence*, unpublished PhD dissertation, George Town University, Washington, DC.

Molander, R.C., Riddle, A.S. and Wilson. P.A. (1996) *New face of war: strategic information warfare*, RAND Institute, Santa Monica, US.

Molander, R.C., Wilson, P.A., Mussington, D.A. & Mesic ,R.F. (1998) *Strategic information warfare rising*, RAND Institute, Santa Monica, US.

O'Connor C.J. (2017) 'Cyber counterintelligence: Concept, actors and implications for security' in Colarik, J, Jang-Jaccard, J & Mathrani, A (eds.) *Cybersecurity and Policy – A substantive dialogue*, Massey University Press.

Panda Security (2018) *The hunter becomes the hunted: How cyber counterintelligence works*, accessed on 06/11.2018 at https://www.pandasecurity.com/mediacenter/panda-security/cyber-counterintel ligence/

Prunckun, H. (2012) *Counterintelligence: Theory and practice*, Rowman & Littlefield Publishers, Plymouth, UK.

Prunckun, H. (2019) *Counterintelligence: Theory and practice*, second edition, Rowman & Littlefield Publishers, Plymouth, UK.

Putnam, R. T. (2015) *Digital mirrors casting cyber shadows - the confluence of cyber technology, psychology, and counterintelligence*, Master of Science (capstone project), Utica College, US.

Schmitt, M. N. (ed.) (2013) *Tallinn Manual on the International Law Applicable to Cyber W*arfare, Cambridge University Press, Cambridge, UK.

Sigholm, J. & Bang, M. (2013) 'Towards offensive cyber counterintelligence: Adopting a target-centric view on advanced persistent threats' in *Proceedings of the European Intelligence and Security Informatics Conference* (EISIC), I.E.E.E, Uppsala, Sweden.

Sims, J.E. (2009) 'Twenty-first-century counterintelligence' in Sims, J.E. &Gerber, B. (eds) *Vaults, mirrors and masks – Rediscovering U.S. counterintelligence,* Georgetown University Press, Washington D.C., US.

Sims, J.E. (2014) 'The future of counterintelligence: the twenty first century challenge' in Duyvesteyn, I, de Jong, B & van Reijn, R (eds.) *The future of Intelligence*, Routledge, Oxon, United Kingdom.

Snider, L.D (1987) *Counterintelligence and security: systemic weaknesses in the U.S. approach* IEEE ,Reference number CH2493 -5/87/0000-0844 $1.00 1987

Sithole, T.S., Duvenage, P.C., Jaquire, V.J. & von Solms, S. H. (2019) 'Eating the elephant - a structural outline of cyber counterintelligence awareness and training' in *Published Proceedings of the 17th European Conference on Cyber Warfare and Security*, Stellenbosch, South Africa, February.

Stech F.J. & Heckman K.E. (2018) 'Human Nature and Cyber Weaponry: Use of Denial and Deception in Cyber Counterintelligence' in Prunckun, H (ed.) *Cyber Weaponry Issues and Implications of Digital Arms*, Springer, Cham, Switzerland.

Stone, G.M. & Bluitt, K. (1993) 'Future law enforcement and internal security communications architecture employing advanced technologies', *IEEE Publication CH3372-0/93*, pp. 194 -202.

Stone, G.M. & Tucker R.S. (1988) 'Counterintelligence and unified technical security programs' in Security Technology, proceedings of the *IEEE International Carnahan Conference on Security Technology: Crime Countermeasures*, New York, US, October.

Van Cleave, M.K. (2007) *Counterintelligence and national strategy*, School for National Security Executive Education, National Defense University Press, Washington D.C., US.

van Niekerk, B., Ramluckan, T. & Duvenage, P.C. (2019) 'An analysis of selected *cyber intelligence texts' in Published* Proceedings of the 18th European Conference on Cyber Warfare and Security, Coimbra, Portugal.

van Niekerk, B. & Maharaj, M.S.(2011) 'The Information Warfare Life Cycle Model', *SA Journal of Information Management* 13(1).

Warren, M. (ed.) (2013) *Case Studies in Information Warfare and Security for Researchers, Teachers and Students*, Academic Conferences and Publishing International Limited, Reading, UK.

An Ontology for Cyber ISTAR in Offensive Cyber Operations

Tim Grant[1], Carien van 't Wout[2] and Brett van Niekerk[3]
[1]R-BAR, Benschop, The Netherlands
[2]Council for Scientific and Industrial Research (CSIR), South Africa
[3]University of KwaZulu-Natal, South Africa
tim.grant.work@gmail.com
cvtwout@csir.co.za
vanniekerkb@ukzn.ac.za

DOI: 10.34190/EWS.20.066

Abstract: The purpose of this paper is to propose an ontology for intelligence, surveillance, target acquisition, and reconnaissance (ISTAR) activities in offensive cyber operations (OCO). An ontology is a formal model of the domain objects and the relationships between them. Ontology languages are available that are both human-readable and computer-executable. Formal ontologies are used to automate the control of operations, often using symbolic Artificial Intelligence (AI) techniques. Both attack and defence can be speeded up by automating cyber operations, although the emphasis here is on attack. Previous work has shown that there are no ontologies in the literature that cover footprinting and reconnaissance, target selection, finding an access path to targets, collecting intelligence once inside the target, or post-attack evaluation. This paper fills the gap on footprinting and reconnaissance. The ontology proposed in this paper was developed following a widely-accepted methodology. Object-classes and relationships in the cyber ISTAR domain were extracted from selected texts in the open literature that describe the intelligence gathering processes in cyber warfare or other forms of hacking. To make ontology development tractable, several scoping assumptions were made. In particular, the ontology assumes that the attacker is outside the target. Further work is needed to extend the ontology to gathering intelligence from insiders or once the attacker has penetrated the target. After an introduction, the paper summarizes relevant theory on the ISTAR process, on OCO, and on ontology development. It describes how the ontology was developed, including outlining the source material. The resulting ontology is presented. Finally, the paper draws conclusions and recommends further work.

Keywords: Military intelligence, surveillance and reconnaissance, target acquisition

1. Introduction

1.1 Motivation

Automating cyber operations is a hot topic in research. An attack occurs so quickly that defenders' only hope of responding to it in real time is to automate pre-planned defensive actions. Conversely, attackers seek to apply automation to overwhelm the defence by coordinating faster, more complex, and distributed attacks.

Cyber warfare is likely to degenerate into sequences of retaliatory operations alternating between the adversaries. An incoming attack triggers a counter-attack, which then results in a counter-counter-attack, and so on. To perform a counter-attack successfully, the (counter-)attacker must combine defensive and offensive actions. The incoming attack must be detected quickly, incident response actions executed in real time, forensic investigation instigated promptly to identify what happened and who did it, before counter-attack options can be evaluated, approved, planned, and executed. It would take at least a year to mount a counter-attack (Grant, 2018), calling into question its legitimacy, let alone its effectiveness. Ideally, four or five orders of magnitude speed-up are needed.

Researchers are studying the application of Artificial Intelligence (AI) techniques to automate cyber operations. The current focus is on sub-symbolic AI techniques for learning from big data. However, big-data AI requires thousands of validated examples to learn effectively. In cyber operations this would restrict application to simple attacks, such as DDoS, defacing, and the actions of script kiddies. Advanced, state-level attacks, such as Stuxnet, Shamoon, Industroyer/Crash Override, BlockEnergy/KillDisk, Petya, and NotPetya, while devastating, do not occur in sufficient numbers to employ big-data AI. Instead, it is necessary to resort to older, symbolic AI techniques.

Symbolic AI requires the acquisition of knowledge about the application domain, with the acquired knowledge being engineered into a formal representation of the domain. While knowledge acquisition and engineering may be tool-supported, this is primarily a manual process. A well-founded method in knowledge engineering is to develop an ontology, defined as a "*formal, explicit specification of a shared conceptualisation*" (Gruber, 1993). An ontology models the objects found in the domain, together with the relations between these objects. Standard ontology languages have been developed that express ontologies in a form that is both human-readable and computer-executable.

Previous work (Grant, 2019) has shown that many cyber attack ontologies can be found in the open, scientific literature. The survey showed that these ontologies have been developed overwhelmingly for defensive purposes (Grant, 2019), such as vulnerability scanning, ethical hacking (a.k.a. penetration testing), or cyber red teaming. They all have a limited scope. They centre on the process of exploiting vulnerabilities to penetrate a target system. While many also cover the subsequent process of violating the target's security, this is invariably represented at a high level of abstraction. None cover battle-damage assessment and after-action review. Only one ontology covers attack processes prior to penetration, and this sole example is restricted to social engineering. In particular, there are no ontologies covering the collection, processing, and analysis of the intelligence needed to conduct cyber operations. It is this gap in the scientific literature that this paper aims to fill.

1.2 Purpose, scope, and paper structure

The purpose of this paper is to propose a logical ontology for intelligence, surveillance, target acquisition, and reconnaissance (ISTAR) activities in offensive cyber operations (OCO). The ontology is based on published descriptions of ISTAR in cyber warfare, and has been developed following steps 1 to 5 of Noy and McGuinness' (2001) methodology.

To make ontology development tractable, its scope was limited by making the following assumptions:

- Operations are performed by a professional, state-level actor, such as a national intelligence agency, the police, or military force, compliant with national and international law.
- The assumed context is that intelligence is being gathered for a counter-attack, in retaliation for a major, incoming cyber attack (e.g. one paralyzing critical national infrastructure).
- The counter-attack will be purely cyber in nature, i.e. not involving physical ("kinetic") action.
- Intelligence is gathered externally, i.e. not using insiders and prior to target penetration.
- The ontology is intended to serve as a checklist for intelligence-gathering at the tactical level.
- Intelligence collection may be all-source.

This paper consists of five sections. After this introductory section, Section 2 summarizes relevant theory on the ISTAR process, on OCO, and on ontology development. Section 3 describes how the ontology was developed. Section 4 presents the resulting ontology. Finally, Section 5 draws conclusions and outlines further research.

2. Related Theory

2.1 ISTAR

In military and police operations, intelligence is the collection and analysis of information to guide commanders in making decisions. The US Department of Defense defines intelligence as "*1. The product resulting from the collection, processing, integration, evaluation, analysis, and interpretation of available information concerning foreign nations, hostile or potentially hostile forces or elements, or areas of actual or potential operations*" (US DoD, 2019, p.107). The term can also mean "*2. The activities that result in the product*" or "*3. The organizations engaged in such activities.*"

Military doctrine regards the intelligence process as a cycle of sub-processes (UK MoD, 2011):

- *Direction*. The commander determines the information required to achieve his/her mission.
- *Collection*. Using sensors, the intelligence organisation collects raw data from their sources.
- *Processing*. Collected data is progressively enriched by analysis to become intelligence that the commander can act on.
- *Dissemination*. The intelligence is reported to the commander.

This paper centres on collection and processing. This is divided into disciplines according to the nature of the resource used. Examples include human intelligence (HUMINT), imagery intelligence, measurement and signal intelligence, open-source intelligence (OSINT), signals intelligence, cyber intelligence (CYBINT), and technical intelligence. Collection may be wide-ranging (*surveillance*), focused (often over an extended period of time) on a potential adversary or on an area of interest (*reconnaissance*), or directed at a particular target (*target acquisition*). In military jargon, intelligence activities may be abbreviated as "intelligence, surveillance, and reconnaissance" (ISR) or as "intelligence, surveillance, target acquisition, and reconnaissance" (ISTAR). Because cyber intelligence for OCO at the operational level is directed mainly on acquiring and characterising target systems, organisations, and people, we term it cyber ISTAR.

2.2 OCO process

Military doctrine defines cyberspace as an "*operating environment consisting of the interdependent network of digital technology infrastructures (including platforms, the Internet, telecommunications networks, computer systems, as well as embedded processors and controllers) and the data therein spanning the physical, virtual and cognitive domains*" (UK MoD, 2016, p.1). Cyber operations are the "*planning and synchronisation of activities in and through cyberspace to enable freedom of manoeuvre and to achieve military objectives*" (ibid., p.51). They are categorised into four roles:

- *Defensive cyber operations (DCO)*, i.e. active and passive measures to preserve the ability to use cyberspace.
- *Offensive cyber operations*, defined as "*activities that project power to achieve military objectives in, or through, cyberspace*" (ibid., p.54).
- *Cyber intelligence, surveillance and reconnaissance*, i.e. ISTAR activities in, and through, friendly, neutral and adversary cyberspace to build understanding of the environment.
- *Cyber operational preparation of the battlefield*, i.e. activities conducted to prepare and enable cyber ISTAR and defensive and offensive operations.

In US DoD cyber operations doctrine (US DoD, 2018), cyber ISTAR activities are termed "cyberspace exploitation". They support both OCO and active DCO (ibid., Figure II-1, p.II-3), and take the form of the traditional intelligence cycle (ibid., p.II-10 and -11). By contrast, experienced hackers organise their information-gathering activities differently. They distinguish footprinting and reconnaissance (Mitnick & Simon, 2002; 2005). Footprinting is "about scoping out your target of interest, understanding everything there is to know about that target and how it interrelates with everything around it, often without sending a single packet to your target" (McClure, Scambray & Kurtz, 2012). By not sending packets to the intended target, the attacker avoids alerting the target's defenders to an impending attack.

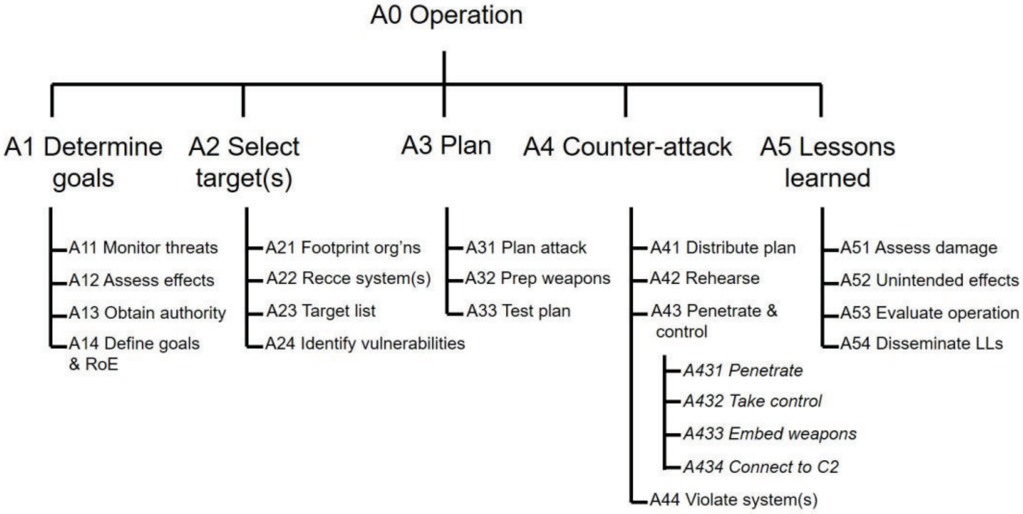


Figure 1: Canonical attack process model (Grant, 2013)

Military kill-chain models of cyber attack fail to make a distinction between footprinting and reconnaissance and between internal and external intelligence (Greene, 2016). Grant's (2013) canonical attack process model – Figure 1 – combines military and hacker thinking. In this model, an attack consists of five phases, with each phase being decomposed into processes. Cyber ISTAR activities appear in Phase 2. After (passive) footprinting (A21), the attacker turns to reconnaissance (A22). From the results, the attacker prepares a prioritised target list (A23). High-value targets can then be actively investigated more thoroughly to identify vulnerabilities (A24).

2.3 Ontology development

Ontology development draws heavily on the software engineering discipline and, in particular, object-oriented analysis and design. Noy and McGuinness (2001) present their aptly-named “Ontology Development 101” methodology, consisting of the following steps:

- Step 1: Determine domain and scope of the ontology.
- Step 2: Consider reusing existing ontologies.
- Step 3: Enumerate important terms.
- Step 4: Define object-classes and the class hierarchy.
- Step 5: Define properties of classes. Properties may be attributes (e.g. the system administrator’s telephone number) and/or inter-class relationships (e.g. the firewall is-part-of the target system).
- Step 6: Define the facets of properties, including cardinality, value-type, domain, and range.
- Step 7: Create instances of the object-classes, together with their property and facet values.

Ontologies can be developed to various levels of abstraction. A *conceptual* ontology consists only of the high-level classes obtained from performing steps 1 to 4. A *logical* ontology consists of classes and relations obtained from steps 1 to 5. A *physical* ontology represents all the detail needed to implement software, obtained from all seven steps. Given that the aim of this research is to develop a checklist for tactical intelligence-gathering, the result is a logical ontology. Filling in the checklist for a specific operation is equivalent to performing steps 6 and 7.

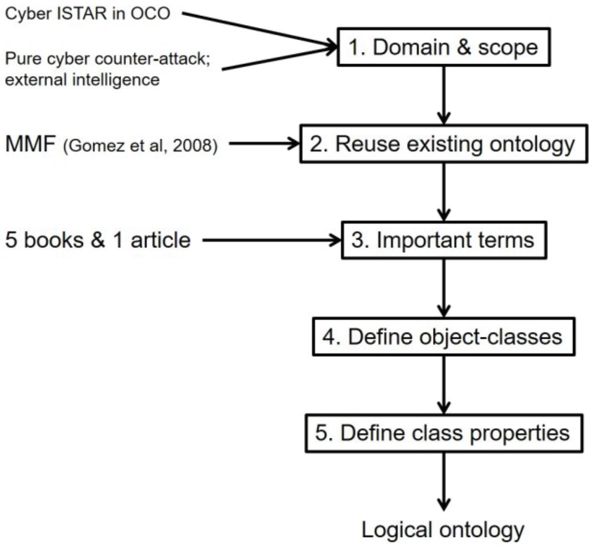


Figure 2: How methodology was applied to developing this ontology

Figure 2 shows how the Noy and McGuinness (2001) methodology was applied in developing the cyber ISTAR ontology. Regarding step 2, Grant (2019) conducted a literature survey to find ontologies for cyber intelligence gathering, but found none. Widening the survey to (kinetic) intelligence gathering resulted in a handful of hits, of which half described Gomez et al’s (2008) Mission & Means Framework (MMF). MMF is an existing (non-cyber) ontology, but covers only a small part of intelligence collection: matching sensors to tasks.

3. Developing the Ontology: steps 1 to 3

3.1 Steps 1 and 2: domain, scope, and ontology reuse

The domain and scope (step 1) have been defined in the preceding sections, as follows:

- *Domain*: a logical ontology for cyber ISTAR activities in OCO at the tactical level.
- *Scope*: intelligence collection for a pure-cyber counter-attack performed by a professional, state-level actor, such as the police, the military, or a national intelligence agency.

As regards reusing existing ontologies (step 2), Grant (2019) suggests combining MMF with van Heerden Chan, Leenan & Theron's (2016) network attack planning ontology (NAPO). Unfortunately, the information published to date on NAPO is incomplete, limiting reuse to MMF.

3.2 Step 3: Extracting terms from source texts

The extraction of important terms (step 3) was divided into four sub-steps:

1. Identify books on (white- or black-hat) hacking and cyber warfare.
2. Select the books that describe intelligence gathering in sufficient detail.
3. From the Table of Contents, identify which chapters must be studied.
4. Study these chapters, extracting important terms relating to the information to be gathered, together with the associated entities.

Candidate books were identified by keyword search of Amazon and Bol.com. Eight books were rejected as lacking a sufficiently-detailed description of cyber intelligence gathering. Another six books described cyber intelligence gathering in detail, but in an anecdotal form, i.e. describing the events in a particular cyber attack. Although these books were unsuitable for (exhaustively) extracting important terms, they may be used at some future date to validate the ontology. Finally, five books – (Winterfeld & Andress, 2013), (Andress & Winterfeld, 2014), (Sood & Enbody, 2014), (Conti & Raymond, 2017) and (Beaver, 2018) - were selected as being suitable sources of important terms. From previous research, an online article (Ruiu, 1999) was added.

Some observations on the hacker literature are worth recording here. Publications authored by black-hat hackers are overwhelmingly confined to online blogs. Moreover, they fail to describe intelligence gathering; hackers prefer to brag about penetrating and violating their targets. By contrast, Mitnick and Simon's books (Mitnick & Simon, 2002; 2005) emphasize intelligence gathering, often using social engineering techniques.

Other authors are coy about describing offensive activities. Both ISTAR and OCO are subject to tight security in professional organisations, and this will deter employees from publishing their knowledge in the open literature. Relevant books often present their knowledge as ethical hacking. For example, Shimonski (2015, p.xiv) writes: "[o]riginally, I wanted to write a book on 'how to spy' and it was deemed to be too dangerous to put into the hands of the public". Instead, Shimonski "attempt[s] to alert you of the dangers of using technology, … and how to protect yourself against the … attacks that now impact our lives". Likewise, Beaver (2018, p.1) details hacking tools and techniques, but emphasizes that this is for "professional, aboveboard [sic], and legal" purposes. A danger from relying on ethical hacking literature could be that the resulting ontology reflects a weakened hacker mind-set stemming from ethical hackers necessarily refraining from violating the systems they penetrate.

An increasing number of authors are heeding Lin's (2009) and Denning & Denning's (2010) calls for the open discussion of OCO in the scientific literature. A majority of the books selected for extracting important terms are authored by employees of or consultants for military forces. While this does not guarantee that they are experienced OCO professionals, it does indicate that the ontology should be relevant to professional OCO. Nevertheless, follow-on work should include a reality check, if permitted by security regulations.

More details on the publications selected for text extraction are as follows:

- Ruiu (1999). The author is a system administrator, writing for other system administrators. His aim is to persuade them to improve security on their systems, based on traffic analysis of hostile attacks on his systems. His article is written from an attacker's viewpoint, covering the phases of Reconnaissance, Vulnerability identification, Penetration, Control, Embedding, Data extraction/modification, and Attack relay. The objective of Reconnaissance is to identify attack destinations (i.e. targets), to identify potential attack relay bases (i.e. other systems that can be used to relay attacks to the target), and to identify the target's and relay bases' defences. Key terms identified include (roughly in order of discovery): attacker, sysadmin, target, network topology in target's region, target's conversation peers, tools, target's hosts, target's neighbours, target's operating system (OS), target's open ports, application, devices, target's ISP, target's traffic patterns, passwords, equipment (i.e. hardware, such as routers), servers, vulnerabilities, security system (e.g. firewalls, IDSs), website, employee, account, and protocol. It is noteworthy that this list omits footprinting terms, because attackers' footprinting activities will not have left traces on Ruiu's systems.

- Winterfeld & Andress (2013) and Andress & Winterfeld (2014). Andress is a security professional in the academic and business worlds, while Winterfeld also has a background in military intelligence. Although the titles and author order differ, the two books are largely identical. Since the 2014 book is marked as the second edition, our study is based on this version. The book's aim is to "*cover the strategic, operational, and tactical aspects of the conflicts in cyberspace today*" (Andress & Winterfeld, 2014, p.xiii). We divide it into four parts. The first part is explicitly based on US joint military publications and doctrine, and covers cyber threats, the cyber battlefield, cyber doctrine, and cyber warriors. The second part covers the logical, physical, and psychological weapons used. The third part covers computer network exploitation (CNE), attack (CNA), and defence (CND). The fourth part covers legal and ethical issues. ISTAR topics are to be found on pages 104 to 117 (recce tools, DNS, and scanning), on pages 138-139 (how logical and physical realms are connected), on pages 156 to 164 (social engineering and how it is used in the military), and in chapter 9 (CNE). Andress and Winterfeld identify three forms of reconnaissance: OSINT, passive recce, and APT surveillance. The first two of these are footprinting activities. Key terms identified include: tool, information source, attacker, vulnerability, exploit, human, target, network, (computer) system, attached device, endpoint and intermediate nodes, sponsor, information, attack vector (hardware, software, insider, software engineering), and tactics, techniques, and procedures (TTPs).
- Sood & Enbody (2014). Sood and Enbody are security researchers and consultants. Their book describes the mechanisms for targeted cyber attack, covering each attack phase in a separate chapter. Chapter 2 details intelligence gathering via the OSINT, CYBINT, and HUMINT disciplines. Online resources, information sources, and tools are tabulated. They divide the intelligence gathering process into four steps for transforming raw data into actionable information. The first step is the discovery and selection of publicly-available resources, such as the Internet, traditional mass media, publications, corporate documents, and exposed networks. The target may be an individual, a group of people, or an organisation, The second step is the extraction of data from the selected resources. In the third step, the data is correlated to find associations and to gain a deep insight into the target's behaviour and environment. Finally, the attack is outlined in the fourth step (which is outside the scope of this paper). Key terms identified include: attacker, target, individual, group, organisation, attack path, information resources, social networks, websites, mass media, documents, geographical locations, employer/employee data, relationships and contacts, achievements, operations, systems, IP addresses, domain names, devices, servers, wifi networks, applications, email, products, meetings, clients and partners, financial status, work culture, disclosures, and social security numbers. Explicit lists of the information of interest to attackers about individuals and organisations can be found on pages 14 to 15.
- Beaver (2018). Beaver is an information security consultant. His book is its sixth edition, indicating both that it is influential (with 80 citations in Google Scholar) and that it has been regularly updated as technology changes. It covers Windows, Linux, and macOS. It outlines the "computer hacking tricks and techniques that you can use to assess the security of your information systems" (ibid., p.1). While written for ethical hackers, it will undoubtedly also be used by black-hat hackers, despite Beaver's disclaimer. Chapter 2 characterises the hacker's mind-set, and Chapter 4 outlines the hacker's methodology. It describes each attack phase in detail, listing the tools used, and showing how to use each tool with screen-dumps. Footprinting is covered by Chapters 5 to 7, detailing open-source information gathering, social engineering, and physical security. Chapters 8 to 16 cover active reconnaissance as one part of the penetration of network hosts, operating systems, and applications. Information obtained from scanning systems is listed on page 54, from company websites on page 62, from WHOIS on page 65, from social engineering on page 71, from dumpster diving on page 76, from physical insecurities on page 84, from network infrastructure on pages 126 to 127, and from network analysis on pages 142 to 144. Key terms identified include: hacker, insider, organisation, information system, vulnerability, domain name, IP address, host and host name, protocol, open port, available share, service, application, tool, website, firewall, router, device, wireless networks and Bluetooth, permissions, authentication requirements, traffic, email and email addresses, information source (social media, websites, search), employee, contact information, patents and trademarks, publications, presentations and articles, vendors, clients and partners, directories, files, source code, developers, servers, organisational charts, phone lists, network diagrams, password lists, and meeting notes.

4. Presenting the Ontology

4.1 Steps 4 and 5: Object-classes and properties

Steps 4 and 5 in Noy & McGuinness' (2001) methodology involve identifying object-classes and their properties from the key terms enumerated in step 3. This necessarily requires design judgement. The resulting ontology is diagrammed in Figure 3. This shows only the upper levels of some of the class hierarchies. Lower-level classes have been omitted to fit the whole ontology into a single diagram.

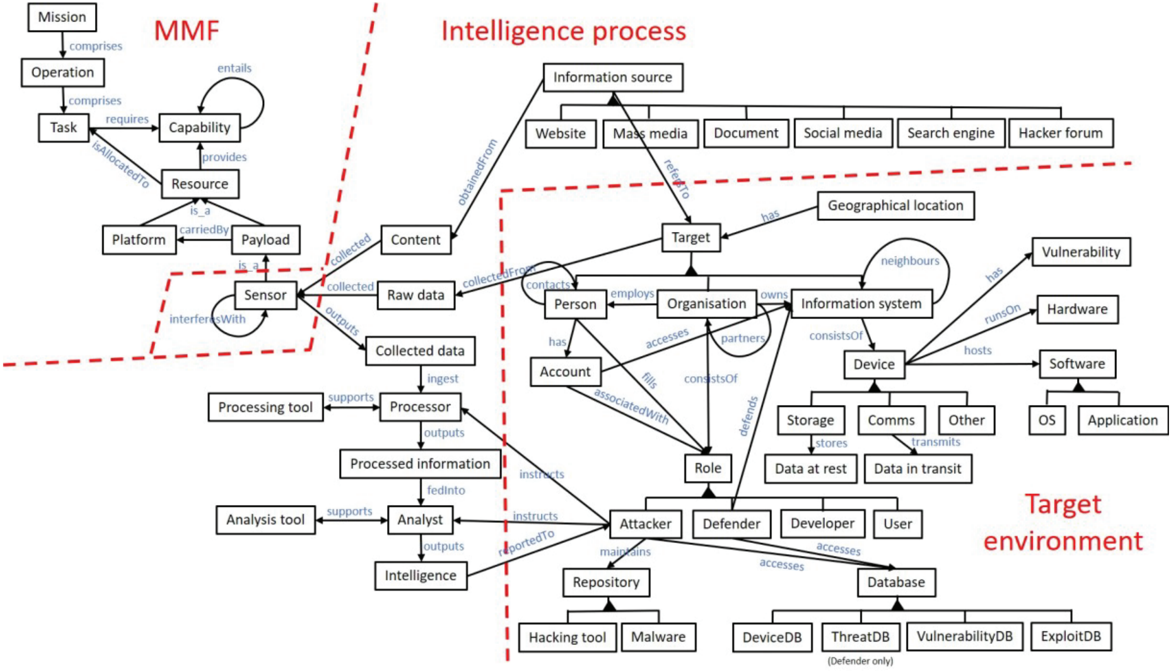


Figure 3: Cyber ISTAR ontology (upper levels).

The ontology was constructed by considering the key terms in turn. Those terms denoting object-classes are depicted as boxes, and, where possible, placed into a hierarchy with similar classes. In Figure 3, small filled triangles denote class hierarchy relationships. For example, there are three types of target: individual persons, organisations, and information systems. Likewise, hierarchies of roles, devices, software, databases, and repositories were identified. Other relationships between object-classes are shown as lines linking the relevant classes. Arrows show whether the relationships are directed. Most relationships have been identified from the authors' domain knowledge[1], rather than from the key terms. Some key terms were considered to be attributes, and are not shown in the figure. For example, user-names and passwords were considered to be attributes of Accounts.

The ontology is divided into three "slices". The MMF slice appears in the upper left-hand corner of Figure 3. The target environment appears in the lower right-hand quarter. Between them is a slice representing the intelligence process. The Sensor class links the MMF and intelligence process slices. The Target class and its Geographical location could appear in both the intelligence process and target environment slices.

4.2 MMF slice

The MMF slice is largely unchanged from Gomez et al (2008). Two object-classes were renamed to avoid confusion in the ISTAR context. Payload was originally named System, but in systems thinking this means an all-encompassing collection of parts with a boundary with its environment. Instead, we borrowed the term "payload" from aviation, where it denotes the useful load (e.g. passengers, bombs, but also sensors) that the platform carries. Resource was originally named Asset, but in the intelligence world the term "asset" refers to an insider agent. Resource is the combination of Payload and Platform, and can be assigned to a Task. Translating this into cyberspace, a virus or worm can be regarded as a platform for carrying malware to a target system. The malware payload does not necessarily need to be an exploit, but could be sensing functionality to observe the target from a neighbouring system, thus obscuring where it was sent from.

[1] A total of 34 years' experience in OCO, military intelligence, information warfare, and cyber security.

4.3 Intelligence process slice

The slice of object-classes between MMF and the target environment represent the intelligence process, together with the information sources from which intelligence is collected. The representation of the intelligence process is directly inspired by military doctrine. Data collected by the sensors is processed to become information, and then analysed to become intelligence. The information sources, their content, and the raw data obtained from the target are in the operational environment. Note that Sensor is included in both the MMF and the intelligence process slices.

4.4 Target environment slice

The Target hierarchy is at the centre of the target environment slice. Targets have a geographic location. The subclasses each have a self-relationship: Persons have contacts; Organisations have partners, clients, and vendors; and Information systems have neighbouring systems.

An Organisation consists of one or more Roles (or jobs or functions), each with a specialisation. The specialisations relevant to cyber attack are shown as subclasses. Similarly, an Information system consists of one or more Devices running on Hardware and hosting Software. A Device may also have one or more Vulnerabilities. Some of these Vulnerabilities may be widely known (e.g. being listed in a vulnerability database), others may be known to a small number of people (i.e. zero-days), and yet others are unknown. Devices may store data “at rest”, others may communicate data “in transit”. Other devices include power supplies, peripherals, etc. Software is subdivided into operating system (OS) and applications. Security services (e.g. firewall, IDS, anti-virus, etc) are not shown, but considered a part of the OS. Applications may include office applications, email, websites, and many others.

When a Person fills a Role relating to an Organisation, he/she may be provided with one or more Accounts giving access to the Information system(s) owned by that Organisation. Attackers and Defenders can access Databases, often online, of devices, vulnerabilities, and exploits. Defenders may also be given access to threats databases provided by cyber security consultants or sister organisations. Attackers maintain a Repository of hacking tools and malware.

4.5 Observations arising

Some observations arose during ontology development. While the books from which the key terms were extracted were deliberately selected for describing the attacker’s viewpoint, it was straightforward to identify some object-classes also found in defensive ISTAR, e.g. Hacker forum and Threat database. Some kinds of object, such as tools and data, appear in several places in Figure 3 because they could not be easily grouped into hierarchies. The Role hierarchy needs to be refactored because attackers, defenders, and developers are themselves users. Moreover, an attacker fills two different roles: one as an attacker of the target, and the other as a user of his/her own system.

The greatest difficulty in designing the ontology was to find a way of representing multiple information systems in networks. The explicit distinction between data “at rest” and data “in transit” within and between systems is a start, because different types of tool are used to exploit these two kinds of data. By adding the “neighbours” self-relationship from the Information system class to itself we gained the advantage of representing systems that are neighbours either in cyberspace or in physical space. Neighbouring systems of either type can be used as attack relays, although physical neighbours require the capability to jump air-gaps, as in Stuxnet.

5. Conclusions and Further Research

This paper proposes a logical ontology for cyber ISTAR activities in OCO at the tactical level. It focuses on intelligence collected from outside the target for a pure-cyber counter-attack performed by a professional, state-level actor, such as the police, the military, or a national intelligence agency. The ontology has been developed following steps 1 to 5 of Noy and McGuinness’ (2001) methodology, reusing Gomez et al’s (2008) Missions & Means Framework for matching sensors to tasks. Five books and an article on cyber warfare and other forms of black-hat hacking have been used as the source material for identifying key terms in offensive cyber ISTAR. The object-classes, relationships, and attributes in the ontology were identified from these key terms. The resulting ontology links MMF through the intelligence cycle to the target environment.

The main contribution of this paper is to fill a gap in the literature on cyber attack ontologies by proposing an ontology for cyber ISTAR activities in OCO. Key limitations are that the ontology has not yet been formalised or validated. Since it is based on literature in the open domain, it does not draw on the knowledge of experienced OCO professionals. Moreover, it does not cover footprinting and reconnaissance by insiders or by an attacker who has penetrated the target system.

The first step in further research should be to formalise the ontology using a widely accepted ontology language, such as OWL2. Then the ontology could be extended to include lower-level object-classes and relationships that have been identified from the source material, but not be shown in Figure 3. Attributes and other properties could also be added. The next step should then be to validate the ontology, either by evaluating it by experienced OCO professionals and/or by hand-simulating the published descriptions of actual targeted attacks. Work should also be done on extending the ontology to footprinting and reconnaissance by attackers inside the target.

References

Andress, J. & Winterfeld, S. (2014). *Cyber Warfare – Techniques, tactics, and tools for security practitioners*. 2nd edition, Syngress, Waltham, MA.

Beaver, K. (2018). *Hacking for Dummies*. 6th edition, Wiley Publishing Inc., Hoboken, NJ.

Conti, G. & Raymond, D. (2017). *On Cyber: Towards an operational art for cyber conflict*. Kopidion Press.

Denning, P.J. & Denning, D.E. (2010). Discussing cyber attack. *Communications of the ACM*, 53, 9, 29-31.

Gomez, M., Preece, A., Johnson, M.P., de Mel, G., Vasconcelos, W., Gibson, C., Bar-Noy, A., Borowiecki, K., La Porta, T., Pizzocaro, D., Rowaihy, H., Pearson, G. & Pham, T. (2008). An Ontology-Centric Approach to Sensor-Mission Assignment. In: Gangemi A., Euzenat J. (eds) Knowledge Engineering: Practice and Patterns (EKAW 2008). Lecture Notes in Computer Science 5268. Springer, Berlin & Heidelberg.

Greene, T. (2016). Why the Cyber Kill Chain Needs an Upgrade. *Network World*, 5 August 2016.

Grant, T.J. (2013). Tools and Technologies for Professional Offensive Cyber Operations. *International Journal of Cyber Warfare and Terrorism*, 3, 3, 49-71 (July-September), ISSN 1947-3435. DOI: 10.4018/ijcwt.2013070104.

Grant, T.J. (2018). Speeding up Planning of Cyber Attacks Using AI Techniques: State of the art. In Chen, J.Q. & Hurley, J.S. (eds.), Proceedings, 13th International Conference on Cyber Warfare & Security (ICCWS), National Defense University, Washington DC, 235-244.

Grant, T.J. (2019). Building an Ontology for Planning Attacks that Minimize Collateral Damage: Literature survey. In Van der Waag-Cowling, N. & Leenen, L. (eds.), Proceedings, 14th International Conference on Cyber Warfare & Security (ICCWS 2019), Stellenbosch, South Africa, ISBN 978-1-912764-11-2 / ISSN 2048-9870, 78-86.

Gruber, T. (1993). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43, 5-6, 907-928.

Lin, H. (2009). Lifting the veil on cyber offense. *IEEE Security & Privacy*, 7, 4, 15-21.

McClure, S., Scambray, J, & Kurtz, G. (2012). *Hacking Exposed 7: Network security secrets and solutions*. 7th edition, McGraw-Hill, New York.

Mitnick, K.D. & Simon, W.L. (2002). *The Art of Deception: Controlling the human element of surprise*. Wiley Publishing Inc., Indianapolis, IN.

Mitnick, K.D. & Simon, W.L. (2005). *The Art of Intrusion: The real stories behind the exploits of hackers, intruders, and deceivers*. Wiley Publishing Inc., Indianapolis, IN.

Noy, N.F., & McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Ruiu, D. (1999). Cautionary Tales: Stealth coordinated attack HOWTO. Digital Mogul, 2-7, July 1999. http://www.nswc.navy.mil/ISSEC/CID/Stealth_Coordinated_Attack.html (20 September 2006).

Shimonski, R. (2015). *Cyber Reconnaissance, Surveillance, and Defense*. Syngress, Waltham, MA.

Sood, A.K. & Enbody, R. (2014). *Targeted Cyber Attacks: Multi-staged attacks driven by exploits and malware*. Syngress, Waltham, MA.

UK MoD. (2011). *Understanding and Intelligence Support to Joint Operations*. 3rd edition, Chiefs of Staff, MoD London, UK, August 2011.

UK MoD. (2016). *Cyber Primer*. 2nd edition, Development, Concepts & Doctrine Centre, MoD Shrivenham, UK, July 2016.

US DoD. (2018). *Cyberspace Operations*. US DoD Joint Publication 3-12, US Department of Defense, Washington DC, 8 June 2018.

US DoD. (2019). *DoD Dictionary of Military and Associated Terms*. Joint Staff, US Department of Defense, Washington DC, USA.

Van Heerden, R.P., Chan, K.F.P., Leenen, L. & Theron, J. (2016). Using an Ontology for Network Attack Planning, *International Journal of Cyber Warfare and Terrorism*, 6, 3, 65-78.

Winterfeld, S. & Andress, J. (2013). *The Basics of Cyber Warfare*. Syngress, Waltham, MA.

The Weapons of Choice: Terrorist Armament Culture and the Use of Firearms in Online Propaganda and Identity-Building through Cyberspace

Ferdinand J. Haberl
University of Vienna, Austria
contact@fhaberl.com

DOI: 10.34190/EWS.20.019

Abstract: Weapons are an integral part of the identity and narratives of various paramilitary and extremist movements – especially jihadi groups and extremist right-wing terrorists. They have been tied to propaganda and psychological operations (PsyOps) and furthermore foster related narratives online. However, although weapons play a critical role with regards to identity-building and the paramilitary perception of statehood (also: *"rebel governance"),* they have largely remained absent from the academic discourse. At the same time, the presence of various armaments proliferates cyberspace as well as the real world. The Islamic State (IS) has even featured its own weapons factories in its propaganda videos in order to display the group's innovative capabilities, legitimisation and notion of nation-building. These improvised armaments have been used as a vehicle to transport IS' ideology aside of providing an ideological justification as to why certain weapons are used in combat. Also, when looking towards far-right terrorism one can recognise the central role of weapons; i.e. AR-15 assault rifles. It is thus relevant to assess the presence, utilisation and glorification of specific weaponry and how certain movements have attached their "identity-building" efforts to them, thus linking ideology and technological innovation. The weapon itself can indeed become part of a group's narrative and identity (especially with regards to masculinities), using cyberspace as a vehicle and platform to showcase its innovative capabilities and to distribute narratives. Thus, this paper aims to assess the relevance certain weapons have to a terrorist armament culture as well as on the respective movements bearing them, the impact improvisation and innovation can have on narratives and to what extend different paramilitary, extremist and terrorist movements have defined themselves through their weapons when engaging in cyber propaganda and psychological operations in cyberspace as well as offline.

Keywords: armament culture, cyber propaganda, Islamic State, far-right online culture, terrorism, social media

1. Introduction

> *"Grab my Gun and my Ammo,*
> *Strap my kamarband onto my chest,*
> *Get dressed up in my camo,*
> *Martyrdom is what I wanted best."*
> - Jihadi Poem (ansar1 forum, 2013)

Firearms play an integral role within the subcultural framework of various extremist movements. Yet, the impact weapons can have on a terrorist or jihadi group's identity, self-perception and online narratives has not been subjected to academic research and discussion (Haberl & Huemer, 2019). This is rather paradoxical, since groups like the Islamic State (hereinafter: IS) or various other militant groups or individuals have used their weapons as symbols in an on-going war of ideas, which they have extended from the real world in to cyberspace. This paper consequently aims to illustrate this phenomenon and thus enable a broader understanding of online propaganda and the related terrorist armament culture, as well as a wider academic discussion on this very topic.

Although debates about so-called 'gun culture' in wartimes and peacetimes are not necessarily uncommon, but academic research into this field is rather sporadic and oftentimes focused on the USA (Heinze, 2014), on mass-shootings or on governmental gun controll efforts in the wakes of such incidents. However, much like in the USA – with respect to incidents like the 2014 Bundy Stadoff (Urquhart, 2016) i.e. - studying 'terrorist armament culture' also significantly relates to concepts of statehood and nationhood and to the role guns can play or have played in nation-building (Heinze, 2014) and identity-building (Hegghammer, 2017). These concepts are also intrinsic to the Islamic State's propaganda (Reuter, 2015), they are commonly debated online (Haberl, 2019) and they are in fact intertwined, which has become evident either on the jihadi end or the far-right end of the radicalisation spectrum. This is especially relevant because social practices involving weapons have the power to relate people, social collectives, ideologies and ideas (Springwood, 2007), using cyberspace as a vehicle to achieve this goal. According to F. Springwood, a weapon can "becomes a tangible object of value inscribed with historical knowledge, knowledge of gendered social relations, violence, and power" (Springwood, 2007; p. 23), a framework that is indeed relevant with respect to identity-building and nation-building. Hence, for those of us trying to decipher the phenomena of terrorism, gaining an understanding of how guns are utilized by different terrorist groups or those affiliated with rebel governance (Krausharr, 2018), the study of armaments opens an additional window into the realms of terrorist cultural practices (Hegghammer, 2017). Indeed, especially since terrorist are all too eager to showcase their various weapons online, studying their armament culture has the potential to add an additional layer to the research on terrorist propaganda in cyberspace.

Researchers must look at the social and political context that surrounds certain armaments and how this context might add to our understanding of identity or biases (Metzl, 2019). We must aim to uncover what certain weapons mean to terrorist subcultures online and refrain from only assessing what they are capable of doing technologically. Hence, recognizing that within the realms of terrorist armament culture, weapons have also turned into powerful symbols that are shaped by online media, culture, rebel governance, as well as by those actors who manufacture and distribute them (Metzl, 2019). Since they have become integral elements of jihadi statehood, ideology and masculinity and since they have served similar narratives with regards to far-right terrorism, their presence in the cyber propaganda and online publications of terrorists must be studied.

2. Identity-building, Ideology and Culture

Armament culture is closely related to the ideology and identity of respective terrorist groups. Particularly religious – jihadi in this case – ideologies are a key-driver of operational and strategic decisions terrorist groups make. This is especially relevant with respect to religiously oriented movements, which are motivated by what they perceive to be the afterlife (Hoffman, 1995). The expected rewards in the afterlife lead groups to accept different risks associated with the weapons they choose. In other words, groups like IS, which is certainly more willing to accept the risk of death, are consequently more likely to select weapons that require a direct confrontation with their adversaries (Koehler-Derrick & Milton, 2017). Hence, this may offer an explanation as to why jihadi groups oftentimes feature knifes and cutting weapons in their propaganda pieces. Additionally, and perhaps even more importantly, it must be mentioned that jihadis ideologically and within their interpretation of Islam, aim to walk in the footsteps of the Prophet Muhammad. This is highly relevant since the Prophet owned 9 different swords, all of which had different names, and are frequently represented in jihadi online discourses, in their online forums and cullture (Lohlker, 2013).

Weapon systems – as a highly sophisticated and advanced cultural artefact – are the result of a process of cultural production (Luckham, 1984). Whilst constituting a remarkable achievement of technology and science, they also hold an intrinsic - nay, archaic - symbolic value. Guns have truly penetrated human consciousness through a whole range of cultural commodities, ranging from toy guns to videogames to motion pictures and in some cases also religion (Lohlker, 2013). As such, "armament culture'', has shaped mankind's consciousness (Luckham, 1984) as well as terrorist culture. We must therefore concentrate some research effort upon the literature, videoclips, publications, and other forms of expression generally recognized as cultural and relevant to the subgroup of terrorist/radical groups. Within this context, terrorist armament culture, can be seen as its own ideology, frameworks of ideas, practices and symbolism, allowing individuals to identify themselves with a certain identity and authority. Glock pistols, for instance, have been labeled a defender "of freedom and democracy" (MCNab, 2005) or the AK-47 type family is commonly assocaited with far-left guerilla movements like the Colombian FARC-EP, which featured two of them on its flag (Fischer, 2005). Despite the fact that the FARC-EP also utilized M16 assault rifles, the AK-47 was apparently deemed to be more symbolic of this far-left guerilla organisation. The civil version of the M16 is the AR-15, which has by now become somewhat of the counterpart to the AK-47 due to its frequent association with far-right terrorism, militias or lone wolfes like Brenton Harrison Tarrant, who committed the Christchurch mosque shootings in 2019 using an AR-15 type rifle, which he had covered with graffiti texts highlighting historic figures and personalities assocaited with "victories" against Islam (Fig. 1). It is acts like these that particualry underline the cultural imporatnce of firearms, specially since he opted to publish the associated pictures of his weapons on Twitter:

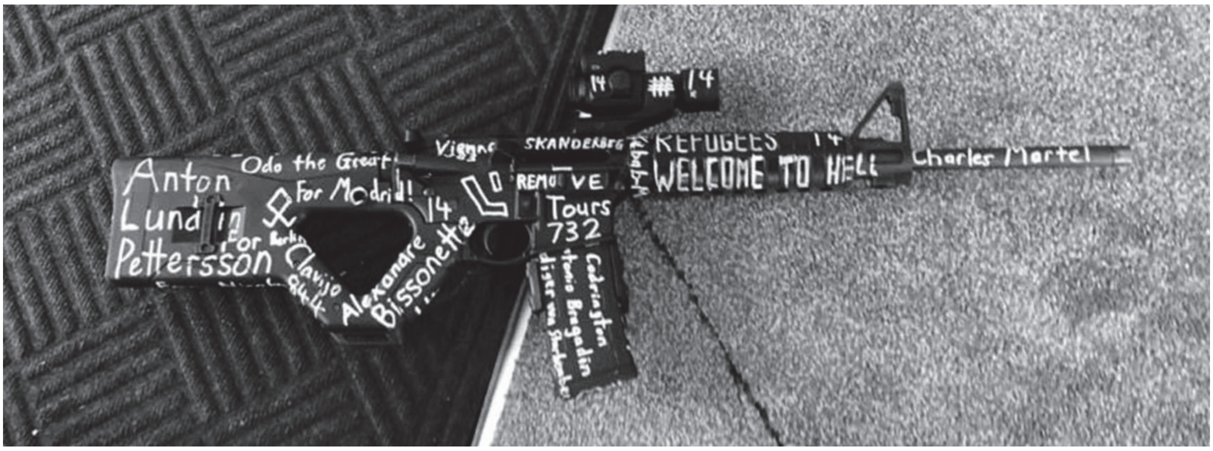


Figure 1: Picture published on the Twitter @BrentonTarrant (now suspended but accessible on archive.org)

Terrorist armament culture, according to R. Luckham, establishes a "constant dissolution and reconstitution of identities" (Luckham, 1984; pp. 1-44). As such, weapons themselves have turned into respective ideologies as they have become symbols and tools for the authority over man by man (Luckham, 1984). As such they have been used as vehicles of propaganda not only by the far-right terrorist Brenton Tarrant, but also in an IS reponse to the attack published on the pro-ISIS Telegram channel "Rahel". The post featured an AK-47 type rifle, an IS flag and a suicide vest (Memri.org, 2019). Much like the AR-15 rifle of Brentaon Tarrant, also the IS reponse featured an AK-type rifle covered with various messages: "Retaliation is coming soon" and "You have opened the gates of hell on your island" (Fig. 2):

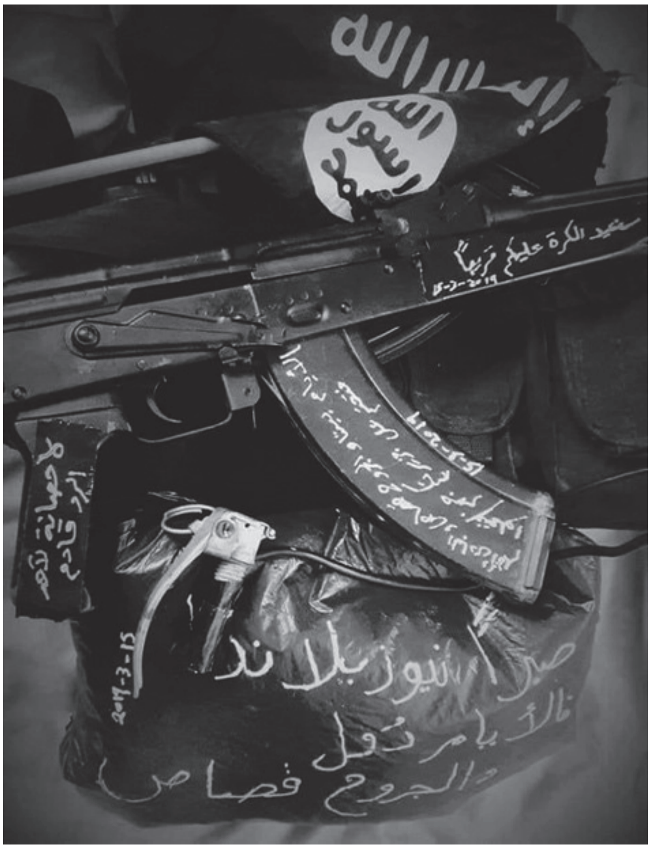


Figure 2: Picture published online on the pro-ISIS Telegram channel Rahel (@Rae_ll) on the March 15th 2019 (Memri.org, 2019)

Another example would be the Norwegian far-right terrorist Anders Breivik, whose Glock pistol had also been engraved with the word "Mjolnir" (the name of the hammer of Thor; Fig. 3) whilst his Ruger rifle featured the word "Gungnir" (Fig. 4), which is the name Odin's spear:

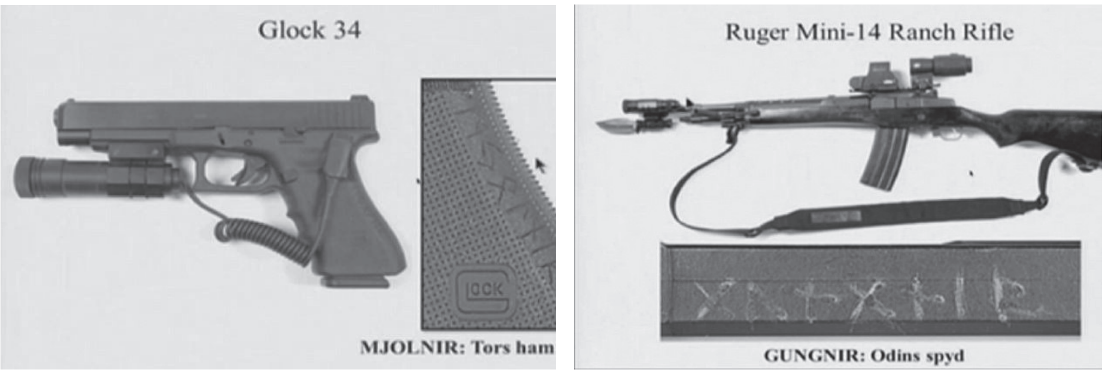


Figure 3 and **Figure 4**: The pictures of Anders Breivik's weapons have been published by the Norwegian government in 2014 (Edmondson, 2014)

It furthermore becomes relevant to investigate those who *de facto* manufacture the terrorist armament culture on which such identity-building efforts can take place. Here, those individuals, branches or institutions who design, produce as well as deploy weapons need to be regarded the creators of this very culture, that is so frequently depicted in cyberspace. Terrorist scientists, manufacturers and engineers, the propagandists, as well as the actual operators of a weapon need to be considered as such (Luckham, 1984), if we desire to fully understand terrorism, its narratives and perhaps if we attempt to improve counternarratives. Improving such counternarratives also requires an in-depth understanding of the complex role weapons play when they are not used for military purposes but rather as a part of the daily lives of people, as part of an ideological narrative, simply for pleasure or as objects that tell a story by themselves (Metzl, 2019).

Another significant part of the jihadi armament culture in this context is the "Reading Mujahideen" – commonly depicted reading the Quran, a weapon close by or in hand. Since jihadism is a transnational subculture, partially born during the Afghanistan wars, and inspired by other national and international conflict and grievances, the related imagery has also been spread globally – frequently featuring AK-47 type rifles. With this spread of jihadism around the world came the visual element and iconography featuring weapons that support this armament culture. The iconography of the" Reading Mujahideen" (Fig. 5) – effectively linking a spiritual struggle to weaponry - dates back to the Taliban and is a frequent element of the current jihadi armament culture too – especially with regards to IS' online activities (Lohlker, 2015):

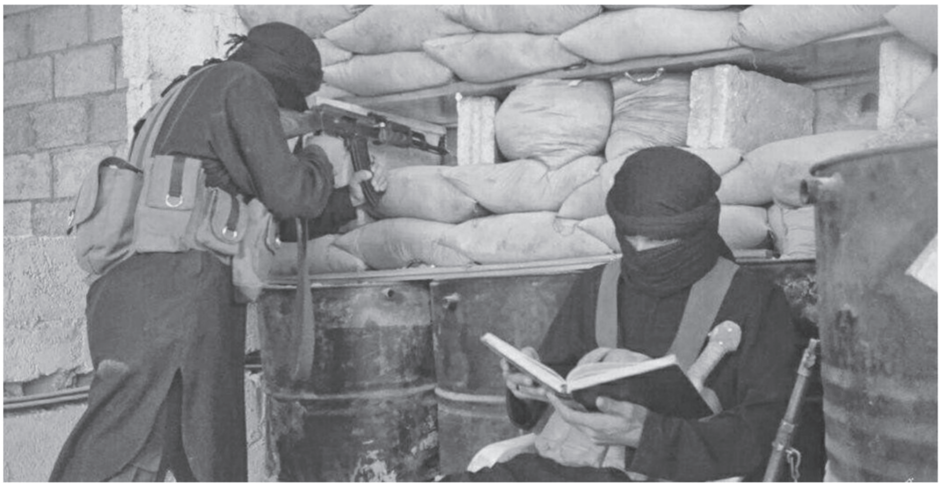


Figure 5: The image has been accessed on an IS-affiliated Telegram channel on Nov. the 29th 2015 (Lohlker, 2015)

Furthermore, upon closer analysis of these images one can actually observe a rather intimate relationship the fighters maintain with their firearms. There is, indeed, a reason for this intimacy, as they are perceived as tools for the transgression from the jihadi life to the jihadi afterlife. Hence, the weapon has become an indispensable tool for the fighter's "martyrdom" (Fig. 6) and thus also directly connecting his present identity to paradise, according to R. Lohlker (Lohlker, 2013). The Quran-reading jihadi fighters are a good example of the ideological importance of firearms, effectively linking religious identity to weapons – in this case a HS-9 by Hrvatski Samokres, which is sold by Springfield Armory under the name "Springfield XD" in the US (Fig. 6):

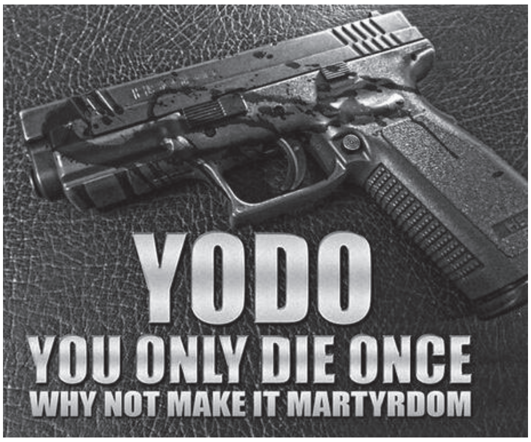


Figure 6: The image has been accessed online (The Guardian, 2014)

3. Hegemony and Statehood

The armament culture of terrorist groups can often also be connected to a hegemonic ideology, in which weapons dominate popular culture and terrorist ideas of nation building and statehood. With respect to IS in particular, this has become rather apparent as the group has frequently used armaments of various kinds in order to assort dominance, whilst expanding its Islamic caliphate – the backbone of the group's online propaganda. Indeed, the terrorist popular culture within the territories under IS' rebel governance has achieved this directly through an extended exposure to various weapons, military might, rifles and the resulting violence in order to convey a narrative of military force and power in the controlled area – the Islamic State truly being the best example, since the group opted to frequently display captured weapons highly effectively in its propaganda (Fig. 7). Although it also featured assault and battle rifles,[1] shotguns[2] and pistols[3] in its propaganda pieces, the manufacturing of armaments has been given a high priority. Much like conventional arms manufacturers, also terrorist arms producers are playing a key-role in this narrative and employ public-relations experts, publish propaganda, and hold what can be described as exhibitions or displays (Luckham, 1984). One example for weapons produced and presented by an IS-owned manufacturer and for somewhat of an IS public relations expert demonstrating the associated weapons (Fig. 7) has been published in an IS propaganda video online:

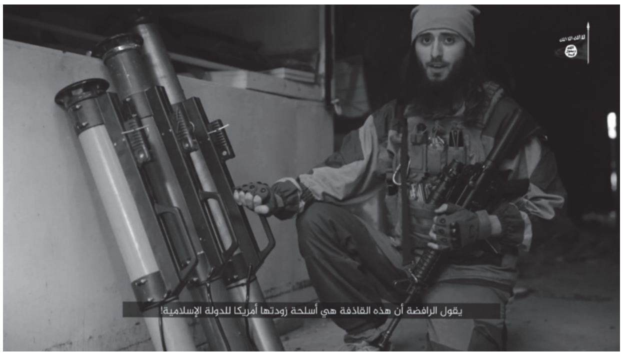

يقول الرافضة أن هذه القاذفة هي أسلحة زودتها أمريكا للدولة الإسلامية!

Figure 7: The image has been accessed online (Maps and Conflicts Database, 2017)

[1] Amongst others the French MAS Modèle 36, MAS-49, The former Soviet Soviet semi-automatic carbines, AK-47s, AKMs, AK-74M, AKS-74U, The Russian AK-103, the Hungarian AK-63 and Automata Módosított Deszant[fegyver] 1965, The Yugoslavian Zastava M70, the Chinese Type 56 and Type 81, the former East German MPi-KM, the American Bushmaster XM-15, M4A1 and M16 rifle, the former Nazi StG 44, the Belgian FN FAL, the Austrian Steyr-AUG (StG 77) and the former West German Heckler and Koch G3 could be identified

[2] Mainly the Italian Benelli M3 and Franchi SPAS-12 combat shotguns could be identified

[3] Various Austrian Glock version (Glock 17 and Glock 19 amongst others), Belgian Browning Hi-Power, an Italian Beretta M92, a Croatian HS2000, a German Walther P99, former Soviet Makarov pistol and TT-33, Ruger pistols chambered in different calibers (amongst others the Ruger P-Series), former. East German Empire P-08, Belgian FN Five-Seven and different ČZUB pistols (primarily but not exclusively the CZ75) could be identified

Importantly, the symbolic transformation of armaments into symbols of statehood legitimizes them within the governing systems and extends the hegemony of those who are in power in this very rebel state. In the ISIS propaganda videoclip, from which the screenshot above has been taken, the protagonist consequently - with respect to the M16 rifle he is carrying, almost apologetically - states:

"This weapon is a product of the Islamic State [the protagonist refers to the IS-built rocket launcher]. To Allah belongs all praise! And this weapon here is ghanima [spoils of war], which was taken from the *rāfiḍa* [lit: "those who reject] personnel, to whom America gives support and weapons [He refers to the US-made AR-15 he is carrying]." (Maps and Conflicts Database, 2017)

These politics built upon a terrorist armament culture provide a universal narrative in which leaders are enabled to establish claims to loyalty, whilst furthermore reflecting the tensions between the dominant and the dominated ideologies within the controlled society. The dominant ideology in such a society – like the one controlled by IS - can thus also control the cultural production through the direct ownership of the production means, which is why the use of a foreign weapon has to be justified ideologically. We have seen another extreme outgrowth of this again with regards to IS, which controlled the educational system, cultural values and heritage (i.e. the destruction of Palmyra in 2015), as well as the media sector, through which it aimed to manipulate public perceptions and its own legitimacy and claim to power – commonly through armament and security. Since IS' rebel governance has also controlled the content of children's textbooks (Fig. 8), the group's armament culture has consequently also affected this aspect of society and day to day life (Luckham, 1984). Weapons have hence become symbols of various concepts of nationhood – especially as a result of its symbolic transformation (Fig. 7 but also Fig. 8 below with regards to the IS education system):

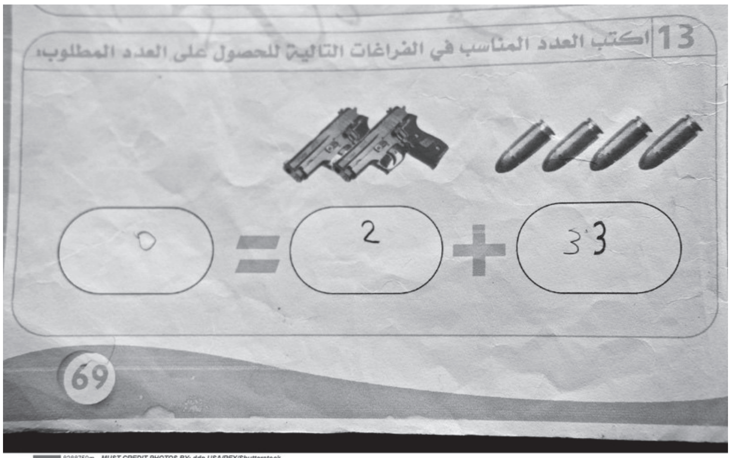

13
اكتب العدد المناسب في الفراغات التالية للحصول على العدد المطلوب:
٥
2
33
69

Figure 8: the image has been accessed online (Molloy, 2017).

In the USA gun producers have furthermore promoted their product through a narrative of frontiersmen, whose expansion and settlement in the Wild West was only possible due to their guns (Metzl, 2019). Intriguingly, the narrative of expansion through weapons and not too dissimilar imagery can be found when looking towards the propaganda of the Islamic State and its expansion towards Syria. Whilst the group portrayed its expansion to be motivated by anti-colonialism and by the goal of an Islamic caliphate, the visual elements of weaponry – without any real military reason indeed - is ever-present in the relevant IS videoclip, in which some of its fighters comment on the destruction of the Syrian-Iraqi border and the resulting expansion of the group (Mezzofiore, 2014).

The association of weapons with exercising control over a nation and frontiers – in the case if IS exercising control of frontlines: "We have no frontiers, only frontlines" (Pauls, 2016), appears to resonate within the many fighters who have joined the group for the purpose of nation-building, something that could be regarded as somewhat of a "frontiersmen characteristic", albeit self-identifying as pioneers on a quest towards building a nation in the name of god. In a similar vein, L. Gahman has observed the interaction of men and guns to produce a shared national identity, which - in his paper he researches rural Kansas in the USA - manifests itself in the "virtues of defending individualism, freedom, property, and religion, and has thus become labelled 'American'" (Gahman, 2015). Here one can indeed explore an increasing amount of research focussing on the connection of the images of guns to power, security, and independence (McIntosh, 2003 as well as Via, 2010).

Paradoxically, many terrorist organisations – especially IS in a twist of irony - proclaim to fight for rather similar goals, using their very own armament culture to support this struggle ideologically.

Intriguingly, the cultural conception of such ‘frontier masculinity’ (Gahman, 2015), according to Melzer (2009) as well as Via (2010), has created a narrative that reinforces the construction of nationalism[4] “by emphasizing the gun as a signifier of manhood” (Gahman, 2015). In a similar vein a BBC report by K. Morris notes that ‘carrying a gun is a passport to adult society for most men’ also in Afghanistan (Morris, 2003). Indeed, also in Iraq it is in fact rather unusual “to meet an adult Kurdish man who does not own at least one firearm” (Small Arms Survey, 2005), which is commonly referred to as the tribal ‘gun culture’ (thaqāfat as-silāh) throughout the Middle East and the Arabian Peninsula (Heinze, 2014). Previous research has found that gun ownership can also be connected to the narrative of a man “providing for his family, bonding with his children, and passing down technical expertise to future generations” (Cox, 2007 and Stroud, 2012). In this regard, handing over a gun or rifle to a child has become a “prominent rite of passage and nostalgic symbol of time spent with their father” (Gahman, 2015). The cultural elements L. Gahman describes in rural Kansas intriguingly appear to hold true for terrorist armament culture as well. The following screenshot from an IS video, which ultimately ends with an alleged Israeli intelligence officer being executed by a teenage boy, support a highly similar narrative of father and son bonding through the use of firearms (Fig. 9).[5]

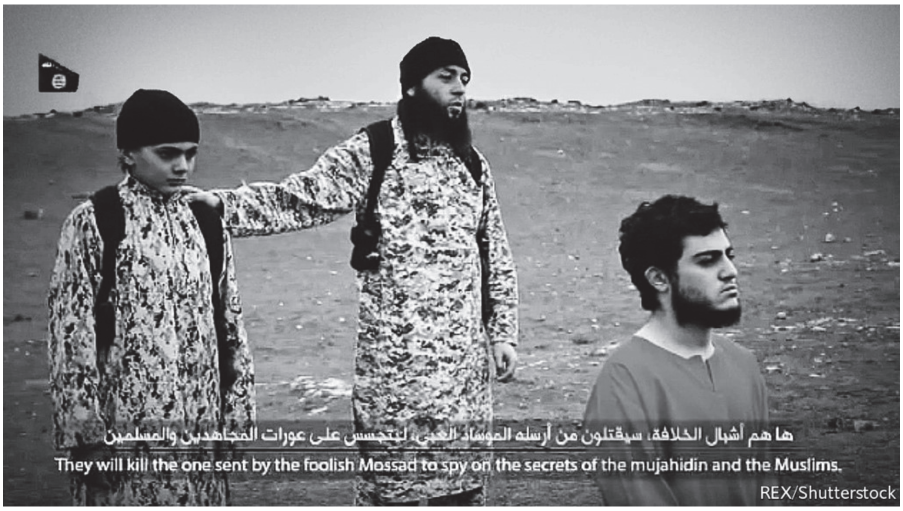

ها هم أشبال الخلافة، سيقتلون من أرسله الموساد الغبي ليتجسس على عورات المجاهدين والمسلمين
They will kill the one sent by the foolish Mossad to spy on the secrets of the mujahidin and the Muslims.
REX/Shutterstock

Figure 9: the image has been published online (Daily Sabah Mideast, 2015)

4. Masculinity

Armaments are oftentimes used as a symbol of masculinity, which is a common cultural aspect of many different societies. This is because the notion of masculinity has been assigned the role of “protector and defender”, which is frequently symbolized by men carrying weapons (Small Arms Survey, 2005). Indeed, in this regard jihadi groups use what R. Lohlker has termed “jihadi masculinities” for identity building (Lohlker, 2013). Still, the connection between firearms and the cultural formation of hegemonic masculinity (Connell, 2005), can be observed on each side of the radicalisation spectrum as well as throughout various cultures with no radical tendaencies at all (Small Arms Survey, 2005). Hence, the gun, as a symbol, has been gendered (Leonard, 2010), especially as a sign of power, control, dominance, authority, self-reliance and autonomy (Felson, R.B. & Pare, P.P., 2010). These themes consequently also relate to male honour protection motives and the maintenance of masculine ideals (Kristen, M. et al, 2019). Consequently, the image of the male fighter as role model of masculinity can be found easily and almost exclusively in connection to weapons (Fig. 10):

[4] IS’ ideology does not support any concepts of nationalism but only one Islamic caliphate.

[5] The video portrays the two protagonists as if they were “father and son”. Their true relationship could not be verified.

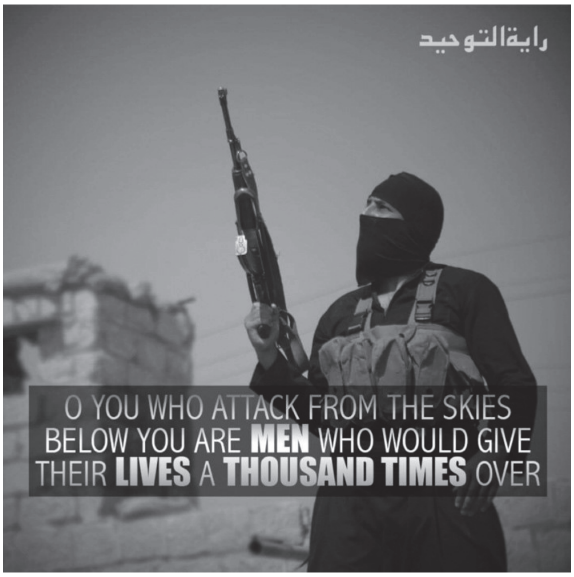

رايةالتوحيد
O YOU WHO ATTACK FROM THE SKIES
BELOW YOU ARE MEN WHO WOULD GIVE
THEIR LIVES A THOUSAND TIMES OVER

Figure 10: the image has been accessed online (Speckhard, 2015)

With respect to jihadi online propaganda, the featured armaments also symbolize virility and a way to resolve some of tensions dominating the jihadi identity construct (Lohlker, 2013) – some of which heavily relate to their perception of male and female roles. However, the role of women in this construct has become rather paradoxical. According to Lohlker as well as Devji, "the obsessive idea of autonomy is so important for men and for jihadis and the attached idea of dominance and absolute control is an illusion endangered especially by the existence or appearance of women. The (male) subject is dependent on something else – women, the "west" etc. – at this stage and will remain dependent in the future. The paradox of dependence and striving for autonomy cannot be solved than by use of violence … In the case of jihadis the solution for this paradox is transformed into a global conflict between religions and cultures, a conflict aiming at the salvation of humanity on a global level." (Lohlker, 2013 and Devji, 2005)

When analyzing the visual representation of how jihadi perceive women within their hegemonic masculinity construct, we will ultimately find three different types of featured and/or discussed women in their online publications: 1) the good-looking and seductive woman, who endangers the jihadis will and determination, taking him off his religious path; 2) the mother, being the woman who could theoretically stand between the jihadi and his divine duties; and 3) finally, the wife, who might entice the jihadi into marital and normal life in which there is no space for his struggle (Lohlker, 2013). However, whilst all three types of women are hardly acceptable options if a jihadi was to embark on his holy battle, the only type of female acceptable within the jihadi masculinity – paradoxically – is the wife, who will also turn towards the jihad (Lohlker, 2013). Indeed, upon the study of jihadi publications and discussion in cyberspace one will encounter almost bizarre imagery of *niqabi* women sporting AK-47 rifles (Fig. 11). Controlled by her *niqab*, whilst aggressively wielding a rifle, the caring mother, who has turned into a fury appears to be a significant manifestation of jihadi armament culture and its distribution online (Lohlker, 2013).

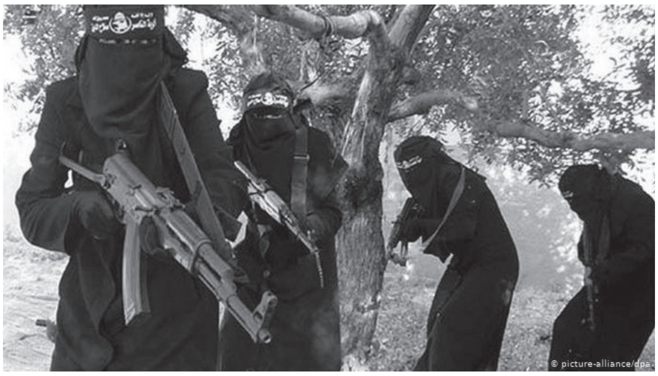


Figure 11: the image has been accessed online (Jansen, 2019)

5. Conclusion

Addressing guns on a cultural and symbolic level allows terrorism and jihadism researchers to investigate how armaments shape terrorist/jihadi identities, masculinity and their notion of statehood (Metzl, 2019), using cyberspace as a vehicle and carrier of their armament culture. Furthermore, armament culture has actually been identified as one key reason as to why many post-conflict recovery programs have failed (Farnam, 2003), which renders the study of it to potentially have real-life implications outside of academic debates and with respect to establishing security in conflict areas. Since it is the very same armament culture that has also enabled them to gain legitimacy, power, authority and momentum (Small Arms Survey, 2005), also governmental practitioners engaged in counter-messaging through counter-narratives may find this approach useful and could apply aspects of this research online. The terrorist armament culture can be countered by decoding the symbolism by which weapons have been incorporated into their hegemonic ideology (Luckham, 1984) and thus by targetting a key-element of their identity-building efforts, masculinities in their propaganda and self-perception – especially when engaging these groups on the virtual battlegrounds of the 21st century.

References

Al Arabiya News. (2015. , 03 12). Iraqi fighter gains social media following in fight against ISIS. *Al Arabiya News*.

ansar1. (2013, 12 05). *www.ansar1.info*. Retrieved from http://www.ansar1.info/showthread.php?t=29569

Connell, R. W. (2005). *Masculinities.* Berkeley: University of California Press.

Cox, A. A. (2007). Aiming for Manhood: The Transformation of Guns into Objects of American Masculinity. In C. Springwood, *In Open Fire: Understanding Global Gun Cultures* (pp. 141–152.). Oxford: Berg: Berg.

Daily Sabah Mideast. (2015, 03 11). New ISIS video appears to show child executing alleged Israeli spy. *Daily Sabah Mideast*.

Devji, F. (2005). *Landscapes of the Jihad: Militancy, Morality, Modernity.* Cornell Univ. Press.

Edmondson, N. (2014, 06 02). *Anders Behring Breivik Trial: Norse God Thor and Odin Engravings on Killer\'s Guns*. Retrieved from International Busienss Times: https://www.ibtimes.co.uk/anders-behring-breivik-thor-odin-guns-police-336683

Farnam, A. (2003). *Gun culture stymies the UN in Kosovo.* The Christian Science Monitor.

Felson, R. B., and P. P. Pare. (2010). Gun Cultures or Honor Cultures? Explaining Regional and Race Differences in Weapon Carrying. *Social forces 88 (3)*, 1357–1378.

Fischer, T. (2005). 40 Jahre FARC in Kolumbien. Von der bäuerlichen Selbstverteidigung zum Terror. *Sozial. Geschichte. Zeitschrift für historische Analyse des 20. und 21. Jahrhunderts*, 77-97.

Gahman, L. (2015). Gun rites: hegemonic masculinity and neoliberal ideology in rural Kansas. *Gender, Place & Culture, 22:9*, 1203-1219.

Haberl, F. (2019). Retreating Into the Desert - The New Paradigms of Jihadi Covert Action, Subversion and Intelligence Operations . In R. Lohlker, *Jihadism Revisited: Rethinking a Well-Known Phenomenon (Jihadism and Terrorism).* Vienna: Logos.

Haberl, F., & Huemer, F. (2019). The Terrorist/Jihadi Use of 3D-Printing Technologies - Operational Realities, Technical Capabilities, Intentions and the Risk of Psychological Operations. *ICCWS 2019 Proceedings*.

Hegghammer, T. (2017). *Jihadi Culture.* Cambridge: Cambridge University Press.

Heinze, M.-C. (2014). On ‘Gun Culture’ and ‘Civil Statehood’ in Yemen. *Journal of Arabian Studies*, 70-95.

Heinze, M.-C. (2014). On ‘Gun Culture’ and ‘Civil Statehood’ in Yemen. *Journal of Arabian Studies*, 70-95.

Hoffman, B. (1995). Holy Terror: The Implications of Terrorism Motivated by a Religious Imperative. *Studies in Conflict & Terrorism 18, no. 4* .

Jansen, F. (2019, 04 08). *Der Tagesspiegel*. Retrieved from Versklavtes Kind dem Hitzetod ausgesetzt: https://www.tagesspiegel.de/politik/prozess-gegen-deutsche-dschihadistin-versklavtes-kind-dem-hitzetod-ausgesetzt/24194706.html

Koehler-Derrick, D., & Milton, D. J. (2017). Choose Your Weapon: The Impact of Strategic Considerations and Resource Constraints on Terrorist Group Weapon Selection. *Terrorism and Political Violence*, 909-928.

Krausharr, W. (2018). Rebellenherrschaft. *Mittelweg 36; Zeitschrift des Hamburger Instituts für Sozialforschung.*, 3-86.

Kristen, M. et al. (2019). Gun Enthusiasm, Hypermasculinity, Manhood Honor, and Lifetime Aggression. *Journal of Aggression, Maltreatment & Trauma, 28:3*, 369-383.

Leonard, D. J. (2010). Jumping the Gun: Sporting Cultures and the Criminalization of Black Masculinity. *Journal of Sport and Social Issues 34 (2)*, 252–262.

Lohlker, R. (2013). Jihadi Masculinities. *Studying Jihadism Vol. 3*.

Lohlker, R. (2013). *Jihadism: Online Discourses and Representations.* Vienna: Vienna University Press.

Lohlker, R. (2016). The "I" of ISIS: Why Theology matters. *Internvetionen*.

Luckham, R. (1984). Armament Culture. *Alternatives X*, 1-44.

Maps and Conflicts Database. (2017, 05 18). *A DETAILED LOOK AT ISIS WEAPONS PRODUCTION IN IRAQI CITY OF MOSUL – MANY PHOTOS*. Retrieved from Maps and Conflicts Database: https://maps.southfront.org/a-detailed-look-at-isis-weapons-production-in-iraqi-city-of-mosul-many-photos/

McIntosh, P. (2003). White Privilege and Male Privilege. In M. &. Kimmel, *Privilege: A Reader* (pp. 147–160). Boulder: Westview Press.

MCNab, C. (2005). *Glock: The World's Handgun.* Amber Books.

Melzer, S. (2009). Gun Crusaders: The NRA’s Culture War. . New York: New York University Press.

Memri.org. (2019, 03 15). *Memri.org*. Retrieved from On Social Media, Jihadis React To New Zealand Mosque Attacks, Expressing Sympathy And Calling For Retaliation: https://www.memri.org/reports/social-media-jihadis-react-new-zealand-mosque-attacks-expressing-sympathy-and-calling

Metzl, J. (2019). What guns mean: the symbolic lives of firearms. *Palgrave Communications volume 5, Article number: 35*.

Mezzofiore, G. (2014, 06 30). Iraq Isis Crisis: Is This the End of Sykes-Picot? *International Business Times*.

Molloy, M. (2017, 02 16). Islamic State textbooks featuring guns and tanks 'used to teach children maths' in school. *The Telegraph*.

Morris, K. (2003, 02 22). *BBC*. Retrieved from Afghanistan's gun culture challenge: http://news.bbc.co.uk/2/hi/south_asia/2788167.stm

Pauls, P. (2016, 06 07). Jürgen Todenhöfer mit „Inside IS“ Der Terrorist aus dem Bergischen Land. *Berliner Zeitung*.

Small Arms Survey. (2005). *Gun Culture in Kosovo*. Retrieved from http://www.smallarmssurvey.org/fileadmin/docs/A-Yearbook/2005/en/Small-Arms-Survey-2005-Chapter-08-EN.pdf

Speckhard, A. (2015, 10 10). *ICSVE*. Retrieved from THE HYPNOTIC POWER OF ISIS IMAGERY IN RECRUITING WESTERN YOUTH: https://www.icsve.org/the-hypnotic-power-of-isis-imagery-in-recruiting-western-youth/

Springwood, C. (2007). Gunscapes: Toward a Global Geography of the Firearm . In *Open Fire. Understanding Global Gun Cultures* (p. 3).

Stroud, A. (2012). Good Guys With Guns: Hegemonic Masculinity and Concealed Handguns. *Gender and Society 26 (2)*, 216–238.

The Guardian. (2014, 06 23). Who is behind Isis's terrifying online propaganda operation? *The Guardian*.

Twitter. (2015, 03 18). *Twitter*. Retrieved from https://twitter.com/raqqa_mcr/status/578185792767574016

Urquhart, J. (2016). Bundy's Federal Feud: Timeline of Events. *Las Vegas Review-Journal*.

Via, S. (2010). Gender, Militarism, and Globalization: Soldiers for Hire and Hegemonic Masculinity. In L. &. Sjoberg, *Gender, War, and Militarism: Feminist Perspectives* (pp. 42–56). Santa Barbara: CA: Praguer

What LEGAD needs to know? Analysis of AARs from Locked Shields (2012-2018)

Jakub Harašta
Masaryk University, Faculty of Law, Institute of Law and Technology, Czech Republic
Jakub.harasta@law.muni.cz

DOI: 10.34190/EWS.20.075

Abstract: The role of lawyers in cybersecurity grows. However, it is often hard to pinpoint what knowledge should be required from legal professionals in this area. This study aims to rectify this knowledge deficiency by directly what is expected in terms of knowledge and skills from government and/or military legal advisor engages in cyber operations. To identify this skillset, the author analyse texts of After Action Reports (AAR) of Locked Shields exercises conducted between 2012 and 2018. Participation in Locked Shields is considered beneficial for the development of professional skills and faring well in Locked Shields is prestigious for all organisations involved. At the same time, Locked Shields exercise contains legal play, which involves legal professionals in solving legal issues arising within a simulated cyber crisis. Therefore, it is only logical to use the Locked Shields as the standard for identification of required knowledge and skillset for legal advisors. The study is structured as follows. After the Introduction, the second part focuses on related work in terms of the role of AARs and the role of lawyers in exercises generally. This part specifies the method used to obtain information from collected AARs (coding of relevant information and subsequent content analysis). The fourth part frames the analysis by providing specific research questions. The fifth part presents the results of this study mainly in terms of the overall goal of the legal play and tasks that legal advisers were supposed to answer. The sixth part briefly discusses the limitation of this study. The final part, Conclusion, provides answers to research questions.

Keywords: cybersecurity, exercises, Locked Shields, legal advisors, after action report, document review

1. Introduction

Real crises, such as large-scale fires in densely populated areas or large-scale terrorist attacks, occur scarcely. However, their potential impact is high and proper response is of the essence. Suboptimal reaction time or decision-making face to face crisis has serious consequences in terms of loss of assets, monetary damage or (human) casualties. Therefore, when possible, these events need to be prepared for as the wrong answer can be worse than no answer or reaction whatsoever. It is without a doubt that practice and training of any kind serve as preparation for the actual performance. Running drills and exercises allows us to evaluate readiness, identify knowledge gaps or skill gaps or to better accommodate new tactics and new equipment. Besides drilling standardised response to ongoing crises, exercises can serve as an educational tool. It is documented that experiencing a situation in a controlled environment is more beneficial for later recollection compared to passively observed lectures (Freeman, 2014). Exercises in this regard serve as an active learning tool (Dewar, 2018:3).

But what is to be taught using this tool and how exactly? Cyber exercises often engage lawyers. There seems to be some sort of understanding that cybersecurity is not in a legal vacuum and that involvement of legal advisors is somehow important. This study aims to provide a systematic overview of the type of tasks that lawyers face when engaged in large-scale cyber exercise Locked Shields.

Locked Shields is organised annually since 2012 by the CCD CoE in Tallinn, Estonia. It is the biggest and most complex exercise of its kind and every year attracts significant media attention. Moreover, organisers prepare after-action reports (AARs) from individual exercises to ensure communication of outcomes to the target audience. By analyzing AARs from Locked Shields exercises, this paper intends to study how are lawyers engaged in the exercise and possibly generalize what is the scope of both knowledge and skills required from lawyers in the area of cybersecurity.

2. Background

2.1 Institutional Learning: the Role of After Action Reports

Analysing and dissecting past mistakes is an inherent part of any learning process. However, facing a real-life crisis, mistakes tend to get expensive fast. It is necessary to ensure that mistakes are not repeated on the institutional scale even if individuals leave their jobs and move elsewhere. As stated by Levitt and March, organisations need to be able to encode *“inferences from history into routines that guide behaviour.”* (Levitt and March, 1988:320) These routines are formed on an interpretation of the past, and not so much on anticipation of the future (Levitt and March, 1988:320). There are many mechanisms to extract knowledge from past events, being real-life crisis or exercises. These include in-progress reviews, after-action reviews, hot wash-ups or different methods of debriefing (Donahue and Tuohy, 2006:3). These methods differ in the level of detail and the time delay from the exercise itself.

Especially in exercises, participants can be evaluated during the exercise itself. They can be granted points for their activity, which serves as a feedback mechanism. This is considered a feedback method with minimal time-delay. It is, in a way, continuous evaluation of success during an exercise. After the exercise, there is often a hot wash-up. Hot wash-up is conducted immediately after the exercise in the form of a lecture and/or discussion. Its goal is to summarize impressions from the exercise from both organisers and participants. Once the exercise is concluded, organisers often prepare AAR to evaluate specific events, provide a more detailed insight into mechanisms within the exercise and to conclude the exercise as a whole.

There is no generally accepted way of preparing AARs (Donahue and Tuohy, 2006:12). Documents often differ in scope, structure and level of detail not only between different organisations but also throughout different years within the same organisation. AARs are often available, but the issue is how to create systematic lessons learned from AARs. It is often mentioned that the process of learning is highly problematic and often not followed through even when specific issues are identified (Donahue and Tuohy, 2006:3, Savoia et al, 2012:2961). This leads to repeated identification of the same issue without it actually being addressed. This is the first aim of this study - to provide a systematic overview of the specific issue throughout Locked Shields exercises between 2012 and 2018, to recurrent topics or specific development is identified.

2.2 Lawyers in Exercises

Legal advisors or general counsels (LEGAD) of armed forces or public agencies and bodies play an essential role. Even if not directly tasked with decision-making in terms of policy or political decisions, they help navigate the legal environment, which serves as a constraint to decision-making. Participation of LEGADs in exercises is one of the prominent obligations of any lawyer. In armed forces, their main role is to provide education in international humanitarian law to others (Sandoz et al, 1987:953). We can extend this into a more general notion of ensuring that personnel they provide legal advice to is prepared not to violate the law. Outside of these obligations, LEGADs should focus on their own study and on systematic preparation of personnel for proper application of legal framework and respect to it.

LEGADs' role in military exercises was well described by Goździewicz (2015) who served as LEGAD in NATO Joint Force Training Centre in Poland. LEGAD focuses on the application of the law in specific operations, which includes working knowledge of rules of armed conflict, knowledge of both NATO and national rules of engagement, targeting with regard to obligations arising from international humanitarian law, identification and application of legal framework relevant for conducting counterterrorism operations, special operations and intelligence gathering (Goździewicz, 2015:30). Anderson concludes similarly, only adding the requirement for working and actionable knowledge of human rights law and the EU law (Anderson, 2016:17-18).

Furthermore, LEGADs need to be aware of what is going to be asked from them. Goździewicz lists a mix of requirements focusing on soft-skills and general awareness of LEGADs roles, such as Goździewicz (2015:40):

- Trust between LEGAD and her commanding officer;
- LEGAD's awareness that a commanding officer looks for solutions, not prohibitions (unless absolutely necessary);
- LEGAD needs to be available to the commanding officer, but also to other personnel within contingent (or, more broadly, organisation);
- LEGAD needs to provide clear and brief advice;
- LEGAD should not be afraid to admit uncertainty and to ask for more time for analysis.

It is absolutely imperative to understand that military commander and LEGAD do not approach the law of armed conflict as an esoteric intellectual exercise (Newton, 2007:880). To them, it is the issue of utmost practical importance, as it allows them to fulfil their goals while respecting the shackles of applicable law. This applies even outside of military organisations and stems to public agencies and private corporations. Ability to provide actionable legal advice that can serve as a relevant basis for further decision-making is a mix of knowledge and soft-skills.

While this is strictly speaking obligatory only in armed forces, government cyber response teams responding to a cyber incident should follow this reasoning as well. Being aware of legal constraints and being prepared to act based on knowledge of these constraints is important and it is LEGADs role to ensure this. That being said, it is unclear what is expected of LEGADs in terms of knowledge about cyber operations and legal framework governing those operations. This is the second aim of the study – to find out what is the skill-set and knowledge set for LEGADs advising on cyber operations (military or civilian).

3. Research Question

As outlined in the previous part, there are two main issues. First, preparation, dissemination and ex-post availability of AARs to training audience and eventually to the public is not sufficient to establish a functional mechanism of influencing institutional behaviour. Lessons learned need to be synthesised and organised in a larger overview, preferably with individual recommendations and actionable findings for further development. Second, there is a lack of details in terms of cyber-specific LEGAD knowledge and skill-set. Some of what is to be expected from cyber-LEGAD could be inferred from general requirements, but specifics are lacking.

To address these two issues simultaneously, this study aims to answer the following research question: *What is the required knowledge and skill-set of LEGADs engaged in advising on cyber activities in the military, law enforcement agencies or government?*

To answer this general research question, I suggest breaking it down to following specific research questions:

1. What is the declared goal of LEGADs' participation in exercise?
2. How were lawyers engaged in an exercise in terms of organisation?
3. What were the tasks that LEGAD had to respond to?
4. What was expected of LEGADs in terms of their answers?

These questions are mainly motivated by the availability of AARs from exercises, and unavailability of data on real-life cyber incidents.

4. Methods and Data

This paper employs a qualitative analysis of AARs from Locked Shields. Qualitative research is primarily exploratory research – it is interpretive because it is used to gain an in-depth understanding of a situation or phenomenon. It uses content analysis of written or recorded materials drawn from participants' responses as well as observations or document reviews. The central part of data analysis is the coding - naming and labelling information, categories and properties occurring in materials. Additionally, qualitative analysis is often related to small-scale studies.

All of the above applies to this study. Locked Shields is organised every year starting 2012. It is currently the biggest cyber exercise in the world. However, with AAR from Locked Shields 2019 not yet available, the number of cases analysed in this study is rather small. I intend to provide a detailed account of the LEGADs' engagement in Locked Shields and generalize from these findings answers to the research questions above.

With regard to data, organisers of Locked Shields prepare AARs to summarise the exercise, describe its technical background, map the tasks that participants had to solve and to eventually provide correct answers and scoring details. AAR 2012 and AAR 2013 are publicly available online. Other AARs were provided per request for research purposes, as these AARs are currently not available online. Therefore in this study, AAR 2012, AAR 2013, AAR 2014, AAR 2015, AAR 2016, AAR 2017 and AAR 2018 were analysed.

5. Results

5.1 Goals

The declared goals of the legal play (as is the participation of lawyers in Locked Shields often labelled) developed over time. As already mentioned, Locked Shields was organised for the first time in 2012. The declared goal for Locked Shields 2012 (LS12) was to train legal experts. Selected method for this training was the exposure of LEGADs to the exercise which should have allowed them to analyse and to observe the events.

As this was the first year, organisers also declared the goal for LEGADs to come up with a plan to make the legal team's involvement more valuable for future exercises (CCD CoE, 2012:5).

This goal probably yielded some results, as LS13 was more articulate in terms of the overall goal of the legal play. Lawyers were invited mainly for analysis of the complex legal issues arising in the context of armed conflict. Additionally, the goal of their participation was the facilitation of communication between the legal and technical experts, education of legal experts about information technology and, to an extent, education of technical experts about the law (CCD CoE, 2013:41). The shift from LS12 is apparent as we see the three-pronged approach to LEGADs engagement, which is very typical for next years as well: (1) LEGADs observe and provide legal risk management, (2) LEGADs learn about technical background of cyber operations, and (3) LEGADs teach relevant legal framework to other participants.

In 2014, declared goals changed only a little. Lawyers were again expected to learn more about information technology with particular attention to the technical execution of cyber operations. Actions undertaken by technical experts need to be understood by lawyers because only then are lawyers able to fulfil their role and allow satisfaction of legal constraints. Additionally, lawyers were required to extract relevant facts from technical experts and to practice providing legal advice on cyber legal issues. I believe this signifies a shift from merely providing opinions to actually providing answers. Overall, a shift from purely knowledge-based approach to approach containing both the knowledge and skills of providing expert legal advice. Another combination of knowledge and soft-skills is the requirement of lawyers being required to be able to explain relevant issues to technical experts (CCD CoE, 2014:7). Briefing complex legal matters in a way understandable to non-legal audiences serve to develop LEGADs explanatory skills, provides technical experts insight into a legal framework, but also aids LEGADs to identify gaps in their own knowledge.[1]

Essentially similar word-per-word goals were declared for LS15 (CCD CoE 2015:9), LS16 (CCD CoE 2016:9) and LS17 (CCD CoE 2017:11). This can probably be interpreted as some sort of general agreement on goals of lawyer's involvement in Locked Shields. AAR2018 contains no specification of the LEGADs involvement whatsoever.

5.2 Organisation

The lower level compared to the overall declared goals is the organisation – how were lawyers actually involved in the exercise.

LS12, as a pilot year, contained multinational legal team. This team was tasked with developing fictional legislation for the whole exercise and also with analysing and observing the whole exercise from a legal perspective. The multinational team advised all players organised in nation-based teams (so-called Blue Teams). AAR12 provides us with feedback to these organisational patterns and it seems to be negative. According to participants, the high-level background scenario was simple and the fictional legislation did not provide a legal team with sufficient opportunity for discussion of possible legal caveats (CCD CoE, 2012:43). Within the feedback, the legal team also suggested that legal experts should be assigned to individual teams, and these individual lawyers should serve as points of contact for separate legal team focusing mainly on international law (CCD CoE, 2012:51).

As with the declared goals, this pilot year and feedback obtained during it was crucial for the further development of the exercise. Starting LS13, LEGADs were assigned to individual teams (CCD CoE, 2013:6) and they were supposed to brief their teams on teams' legal status, applicable law, rights and obligations. LEGADs were also tasked with individual queries coming from respective HQs - these scripted events (so-called injects) served to develop the scenario and test LEGADs in their ability to provide concise advice on legal risks associated with decision-making. There were also lawyers among the organisers (so-called white team) and these were tasked with the evaluation of LEGADs performance and communication with them regarding possible questions arising from legal play (CCD CoE, 2013:6). This model became a gold standard to LEGADs organisational involvement in the Locked Shields, as it was repeated in LS14 (CCD CoE, 2014:31), LS15 (CCD CoE, 2015:32), LS16 (CCD CoE, 2016:28) and LS17 (CCD CoE, 2017:35). There is no mention about the organisation of LEGADs in LS18.

[1] This is an application of Feinmaynian 'If you cannot explain something in simple terms, you do not understand it enough'.

5.3 Tasks

It is important to make a clear distinction between the two broadly defined tasks that LEGADs undertake in exercise. The first, LEGAD is advising his team whenever necessary - however, this information does not appear in AARs, as it is the individual decision of every team and not concerning organizers of the whole exercise. The second, LEGAD is replying to scripted events (injects) presented to him by organisers. In 2012, the legal team was advising only when asked. Since 2013, injects are a regular part of the exercise. AAR13, AAR14, AAR15 and AAR16 contain detailed annexe with a description of all the injects in the exercise. AAR15 contains only a general description, as the detailed annexe for legal play is part of AAR. Overall, LS13 and LS14 contained 8 injects each, LS15 contained 11 injects, LS16 10 and LS17 12 injects. AAR18 contains no useful data in this regard.

Injects generally fall into three categories.

First is the legal briefing for the team. LEGAD is tasked with explaining its team the legal framework in which the team operates, as well as specific rights and obligations arising from it. As the exercise spans over two days and the fictional scenario escalates, briefing usually takes part at the beginning of each day. Briefings were formally introduced as injects in LS13.

The second category of injects are technical quizzes where LEGADs obtained points for answering relatively technical questions (eg. 'Why do domain names sometimes have more than one IP address' in LS14 (CCD CoE, 2014:Annex IX)). These appeared in LS13 (CCD CoE, 2013:8) and LS14 but were later abandoned.

The third category of injects were specific tasks related to the fictional scenario requiring LEGADs to analyse a specific situation and provide its legal assessment. These injects show a trend in growing complexity and knowledge requirements of LEGADs. In 2013, there were four injects addressing specific queries. Two came from the upper echelon and required lawyers to exhaustively explain existing rules of engagement covering the fictional operations, and explore possible responses with regard to coordination of cyber-attacks through social networks. Other two injects covered communication with journalist collecting data for his work and the draft of official government response to allegations of planning a cyber-offensive. Both required explanation of complex legal issues to non-lawyers.

There were four injects coming from the upper echelon in LS14. These were concerned with collecting specific information from IT personnel and providing accompanying legal analysis to it, aiding to assign responsibility to a non-state actor, evaluation of a legal framework for requesting cooperation within NATO structures, and legal assessment of hack-back in response to cyber espionage.

Detailed information on LS15 is not available as the AAR15 does not contain detailed annexe. However, from the general description contained within the AAR15, it is evident that LEGADs were tasked with advising the forensic team on data protection and privacy safeguards, the legality of hack-back, the legal response against attack on SCADA systems and legal assessment of responses on both cyber and kinetic operations against the power grid. In LS16, the number of injects from upper echelon raised to eight. Injects were concerned with communication of appropriate privacy and data protection safeguards to forensic team, legal evaluation of possibilities to remove content from websites and blogs, hack-back to regain control of hijacked drone, legal response to publication of misinformation, preparation of written legal brief for an official response prepared by the government and to explore legal options for response to overall heightened hacking activity within ongoing and escalating conflict. Interestingly enough, alternative injects appeared to allow legal play to respond to what happened over the course of the technical play. LS17 was increasingly complex - the fictional background scenario was related to an island nation. Injects were concerned with other regimes than cyber (maritime law, environmental protection law, human rights law etc.) in the previously unseen scope. LEGADs were forced to consider the close interplay of different regimes when facing situation concerning election hacks, surveillance of underwater cables, hacking activity with physical consequences (oil spills). Injects were, compared to previous years, more complex and LEGADs were encouraged to accompany their answers with references. Similarly to LS16, alternative wording for certain injects were introduced to force LEGADs to communicate with technical experts to establish facts.

5.4 Answers

Answer keys or sample answers are available in Annexes to AAR13, AAR14, AAR16 and AAR17. Even in case of expected answers to injects, increased professionalization of organizers and subsequently increased requirements on LEGADs are apparent. Selected answers to one of the injects in LS13 are listed in AAR13 and it is evident that LEGADs often refer to literature or case law in their answers even when not specifically encouraged to do so. Answers are highly complex and aim to provide as detailed advice as possible taking into account possible interplay between national and international law. This seems to be the standard throughout all analysed AARs.

Answer keys listed in AARs demonstrate that were evaluated not only based on the quality of the provided advice, but also on other aspects related to it. For example, LEGADs in LS14 were tasked with providing legal assessment in the issue of the request for cooperation within NATO structures. Answer key breaks down the evaluation into different elements. LEGADs were granted points from organisers based on their knowledge (eg. whether their analysis discussed separate options for assistance, such as art. 4 of NATO Treaty, art. 5 of NATO Treaty, deployment of Rapid Response Teams), but also on their skill or overall awareness of their role (eg. whether LEGADs' analysis laid down options but left the final decision upon the decision-maker).

6. Limits of Analysis of After Action Reports

As in every research, this study cannot provide a full picture. It is important to note that lack of information about the role of lawyers in LS18 limits results of this analysis. I have approached several people who participated in LS18. During non-formal discussions, it became apparent that there was a change in the organisation of LEGADs' participation in the exercise. LEGADs were detached from technical teams and elevated to strategic play that concerns higher-level decision-making processes. Strategic play participation is voluntary, which led to fewer lawyers involved in the exercise, as well as to different modus operandi for those taking part. As this study is concerned with the analysis of AARs from Locked Shields exercises, and this information is missing in AAR18, it is not part of this study and will not be further analysed. However, I believe that it must be mentioned for the sake of completeness. It also serves to demonstrate a serious limitation of this study - what did not appear in AARs is not reflected in the conclusion, even if it played a major part in the actual conduct of exercise.

7. Conclusion

To conclude this study, it is necessary to answer research questions laid down in the text above.

The first question was concerned with the declared goals of LEGADs' participation in exercise. Throughout different years of Locked Shields exercises, declared goals stabilised as requiring LEGADs to (1) develop legal knowledge in other participants, (2) develop their own knowledge with regard to both the law and IT, (3) develop their own skills with regard to advisory activity. This is in direct agreement with identified background literature, mainly with Goździewicz's (2015:40) remarks to what is requested from a legal advisor in military exercises outside of cybersecurity.

The second question was concerned with the way that lawyers are involved in the exercise. Not taking into account LS12 (pilot year) and LS18 (no information provided in AAR, although there was a shift in the organisation not covered in this analysis), LEGADs are assigned to individual teams.

The third question was concerned with tasks that LEGADs have to respond to. Surprisingly enough, in spite of Locked Shields being focused on cybersecurity, development in complexity witnessed in LS17 and LS18 requires LEGADs to demonstrate working knowledge of different areas of both international and domestic law. Even when, in some regard, their analysis does not have to be exhaustive, LEGADs were tasked with a wide range of tasks including anything from rules of engagement and *ius ad bellum* all the way to the intellectual property and data protection law.

The fourth question was concerned with what was expected of LEGADs while responding to injects. Scoring in Locked Shields is concerned not only with what is contained in the answer (eg. identification of relevant legal framework) but also with how the answer is framed. LEGADs are required to be knowledgeable of their area, but also to know their role in the respective organisation and incident response mechanisms. LEGADs are not supposed to make decisions, but support decision-making. Not overstepping own role is a trait of great LEGAD and Locked Shields organisers took this into account when evaluating the answers. Not only knowledge but also skills and overall awareness is important.

Therefore, it is necessary to conclude this study that based on evidence from AARs of Locked Shields 2012 to 2018, LEGADs are facing difficult challenges, but their overall role in cyber operations is very similar to LEGADs' role in military exercises unrelated to cybersecurity. Required knowledge of both the law and technical details related to cyber operations is extensive.

Acknowledgment

This research was supported by ERDF 'CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence' (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

References

Anderson, D. "The Functions of the Legal Adviser: Advising, Negotiating, Litigation", in Zidar, A. and Gauci, J.-P. eds. (2016). *The Role of Legal Advisers in International Law*, Brill Nijhoff, Leiden.

Dewar, R. (2018). "Cybersecurity and Cyberdefence Exercises", [online], CSS Risk and Resilience Reports, Center for Security Studies, Zurich, https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/Cyber-Reports-2018-10-Cyber_Exercises.pdf.

Donahue, A. and Tuohy, R. (2006). "Lessons We Don't Learn: A Study of the Lessons of Disasters, Why We Repeat Them and How We can Learn Them", *Homeland Security Affairs*, vol 2, pp 1-28.

Freeman, S. , Eddy, S., McDonough, M., Smith, M., Okoroafor, N., Jordt, H. and Wenderoth, M. (2014). "Active Learning Increases Student Performance in Science, Engineering, and Mathematics", *Proceedings of the National Academy of Sciences of the United States of America*, vol 111, no 23, pp 8410-8415.

Goździewicz, W. (2015). "Training a Combat Legal Advisor: Tactical Level Observations and Lessons Identified from Trainings and Exercises", *NATO Legal Gazette*, no 36, pp 29-40.

Levitt, B. and March, J. (1988). "Organizational Learning", *Annual Review of Sociology*, vol 14, pp 319-338.

NATO CCD CoE (2012). "Cyber Defence Exercise Locked Shields 2012: After Action Report", [online], https://ccdcoe.org/uploads/2018/10/LockedShields12_AAR.pdf.

NATO CCD CoE (2013). "Cyber Defence Exercise Locked Shields 2013: After Action Report", [online], https://ccdcoe.org/uploads/2018/10/LockedShields13_AAR.pdf.

NATO CCD CoE (2014). "Locked Shields 2014: After Action Report".

NATO CCD CoE (2015). "Locked Shields 2015: After Action Report".

NATO CCD CoE (2016). "Locked Shields 2016: After Action Report".

NATO CCD CoE (2017). "Locked Shields 2017: After Action Report".

NATO CCD CoE (2018). "Locked Shields 2018: After Action Report".

Newton, M. (2007). "Modern Military Necessity: The Role & Relevance of Military Lawyers", *Roger Williams University Law Review*, vol 12, no 3, pp 877-903.

Sandoz, Y., Swinarski, C., and Zimmermann, B. eds. (1987) *Commentary on the Additional Protocol of 8 June 1977 to the Geneva Convention of 12 August 1949,* Martinus Nijhoff Publishers, Geneva.

Savoia, E., Agboola, F. and Biddinger, P. (2012). "Use of After Action Reports (AARs) to Promote Organizational and Systems Learning in Emergency Preparedness", *International Journal of Environmental Research and Public Health*, vol 9, no 8, pp 2949-2963.

Strengthening the Cybersecurity of Smart Grids: The Role of Artificial Intelligence in Resiliency of Substation Intelligent Electronic Devices

Ion A. Iftimie[1] and Gazmend Huskaj[2]

[1]Eisenhower Defence Fellow, NATO Defense College, Rome, Italy; European Union Research Center, George Washington School of Business, Washington, D.C.; Central European University, Vienna, Austria

[2]Swedish Defence University, Stockholm, Sweden; University of Skövde, Sweden

iftimie@gwu.edu
gazmend.huskaj@fhs.se

DOI: 10.34190/EWS.20.043

Abstract: The Executive Order 13800—Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure—issued by the President of the United States, calls for an evaluation of the "readiness and gaps in the United States' ability to manage and mitigate consequences of a cyber incident against the electricity subsector." In May of 2018, the Office of Management and Budget finished evaluating the 96 risk assessments conducted by various agencies and published Federal Cybersecurity Risk Determination Report and Action Plan (Risk Report). While the report embraced a broad defending forward strategy, it made no reference to smart grids or their vulnerable intelligent substations and did not address how federal agencies plan to respond to emerging threats to these systems. While the paper does not discuss how to attack the smart grids in the cyber domain, the contribution to the academic debate lies in validating some of the vulnerabilities of the grid's substations in order for government, private industry, academia, and civil society to better collaborate in disrupting or halting malicious cyber activities before they disrupt the power supply of the United States and its Transatlantic allies. We also discuss how Artificial Intelligence and related techniques can mitigate security risks to cyber-physical systems. Until this technology becomes available, however, standardization of cyber security efforts must be enforced through regulatory means, such as the enforcement of security-by-design Intelligent Electronic Devices and protocols for the smart grid.

Keywords: Cyber Security; Artificial Intelligence; Intelligent Electronic Devices; Substations; Smart Grids

1. Introduction

The development of smart grids has not been achieved with cybersecurity in mind, despite the fact that connectivity of their unsecure components to the internet made them prime "targets for hacker groups to attack the infrastructure" (Nartmann, Brandstetter, and Knorr 2009). In their quest for corporate efficiency and out of responsibility to shareholders, transmission system operators invested heavily in intelligent services "that save time and effort" (Benzarti, Triki, and Korbaa 2017) in order to streamline processes and reduce costs. This is best illustrated by the Department of Energy (DoE) initial definition of smart grids as a "class of technology people are using to bring utility electricity delivery systems into the 21st century, using computer-based remote control and automation. These systems are made possible by two-way communication technology and computer processing that has been used for decades in other industries. They are beginning to be used on electricity networks, from the power plants and wind farms all the way to the consumers of electricity in homes and businesses. They offer many benefits to utilities and consumers mostly seen in big improvements in energy efficiency on the electricity grid and in the energy users' homes and offices" (Muratori, Schuelke-Leech, and Rizzoni 2014). The accelerated adoption of these automated technologies "has led to an increased probability of security deficiencies inherent in these systems" (Nartmann, Brandstetter, and Knorr 2009), affecting the overall security of the smart grids. At least initially, safety considerations (an additional, underestimated cost of the smart grids) were not the intended purpose of this technology transformation (meant to save costs), but rather the consequence of a plethora of unexpected vulnerabilities that threatened to take the smart out of smart grids.

This paper looks at Baldwin's definition of security as "low probability of damage to acquired values" and suggests that smart grid operators should consider Baldwin's three questions when assessing smart grid vulnerabilities: security for whom, for which values, and from what threats? (Baldwin 1997). In other words, operators must first determine 1) what critical systems to protect, 2) from which risks, and 3) by which means? (Cherp and Jewell 2014). This paper is organized as follows. Section 2 (following this introduction) discusses what vital systems of the smart grid to protect, where intelligent relay systems are defined. Section 3 discusses from what threats to protect them, where the cyber threat is defined. Section 4 discusses by what means to defend forward, with an emphasis on Artificial Intelligence. Finally, section 5 concludes the paper.

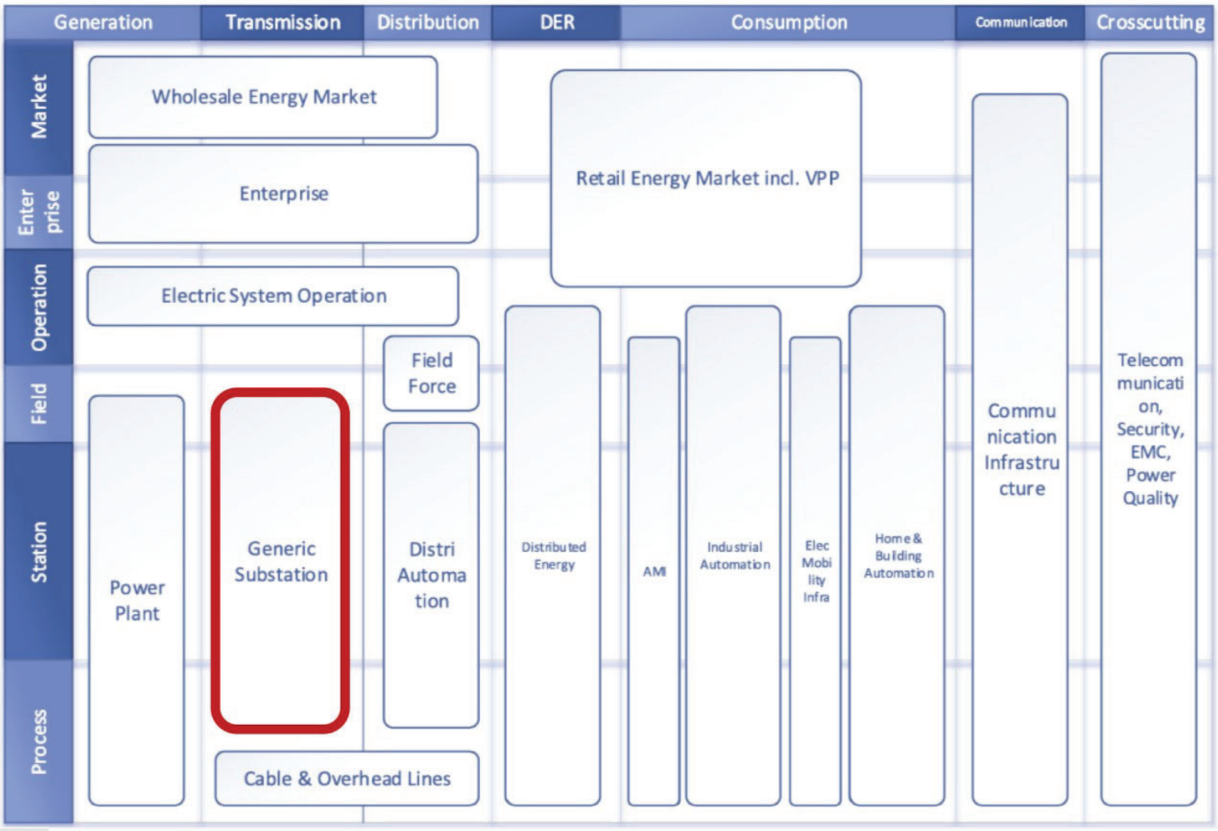


Figure 1: Smart grid components based on the International Electrotechnical Commission (IEC) Smart Grid Standards Map

Source: Rafał Leszczyna in the International Journal of Critical Infrastructure Protection (Leszczyna 2018)

We will be utilizing the International Electrotechnical Commission (IEC) Smart Grid Standards Map (see figure 1) to filter through “512 standards published by standardisation bodies including IEC, ISO, CISPR, EN, ENTSO, ETSI, ITU-T, W3C, IEEE, and the IETF” (Leszczyna 2018, 74) and identify relevant guidelines for the development of defending forward strategies. For the purpose of assisting with this assessment, we decided to limit our focus on the protection of intelligent substation protection and control systems. We took this decision after a Con Edison preliminary report showed that “primary and backup relay systems did not isolate a faulted 13,000-volt distribution cable at West 64th Street and West End Avenue” (Con Edison 2019), causing a power outage that affected midtown Manhattan and parts of the Upper West Side (and 72,000 customers) of New York for one night in July of 2019. While the failure of the substation was caused by excess demand, we argue that the same effect could have been easily achieved through a surgical cyber attack by an adversary with both the capability and intent to hurt the United States, and propose that government entities (such as the Department of Homeland Security’s Cybersecurity and Infrastructure Security Agency) must do more to protect substation protection and control systems from potential cyber attacks.

2. What to Protect? Understanding the Intelligent Substation Protection and Control Systems

In the contemporary security environment of the 21st century, a surgical cyber attack against an intelligent component of the substation protection and control system could cause smart grids to experience “operational failures and loss of synchronization. This operational failure may damage critical power system components which may interrupt the power supply and make the system unstable" (Anwar and Mahmood 2014). From the government perspective, this is a threat to critical infrastructure protection. The resulting blackout could then lead to second and third level effects such as high financial losses for businesses, high financial penalties for transmission system operators, high customer dissatisfaction, and even the loss of life if the power supply of a hospital's emergency room is being impacted (contrary to the purpose of the smart grid).

In Intelligent Substation Protection and Control Systems the "fiber optic cable as the transmission carrier, and the optical signal based on the coding technology replaces the previous electrical signal, which guarantees the real-time validity of the sampling and the timeliness of the information feedback; [meaning that] the entire substation data acquisition and processing rely on Internet monitoring" (Hai-bo et al. 2017). Other than that, however, primarily due to the high costs associated with developing and changing entire substations, the protection system structure remained similar in both intelligent and conventional substations. The typical system has multiple data processing points that "perform similar algorithms on voltage and current data. There is also only limited communication between protection devices, and each device is only aware of the bay in which it is installed" (Ljungberg 2018). Duplication of functions among data processing points that monitor current and voltage data (with limited communication among each other, see figure 2) means that exploitation of one data processing point through a surgical cyber attack could lead to cascading effects and even to a full-blown regional blackout.

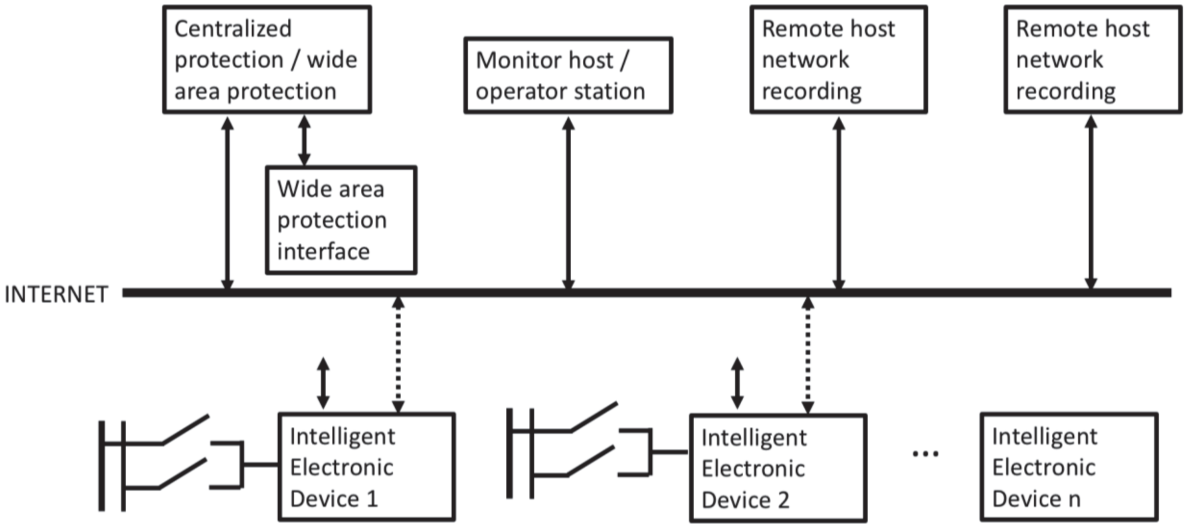

Centralized protection / wide area protection
Monitor host / operator station
Remote host network recording
Remote host network recording
Wide area protection interface
INTERNET
Intelligent Electronic Device 1
Intelligent Electronic Device 2
...
Intelligent Electronic Device n

Figure 2: Simplified representation of intelligent substation relay protection configuration

Source: Diagram adapted from Hai-bo et al. in Journal of Electronic Research and Application (Hai-bo et al. 2017)

These vulnerabilities are empirically highlighted by the 2013 attack on the Metcalf substation in California, and most recently the midtown Manhattan power outage of July 2019. The attack on the Metcalf substation, where attackers managed to cut the connected telephone cables before snipers disabled 17 large power transformers by shooting at the substations, highlighted just how vulnerable substations are to physical attacks. The chairman of the Federal Energy Regulatory Commission (FERC) was quoted saying that "if [the attack] was widely replicated across the country, it could take down the U.S. electric grid and black out much of the country" (Congressional Research Service 2018). It took 27 days for the local utility to restore the substation. Two years after the Metcalf attack, and on the 42nd anniversary of the power outage of July 13, 1977, parts of the city of New York was left in darkness. People were trapped in subways and elevators, Broadway shows had to be canceled, and police had to direct traffic as traffic lights were not working (Barron and Zaveri 2019). The utility provider noted a failure at the "13,000-volt distribution cable at West 64th Street and West End Avenue [and] the relay protection system at West 65th Street substation did not operate as designed" (Con Edison 2019). Both primary and backup systems did not operate as expected. If the relay systems would have failed in the case of the Metcalf substation attack, much of the Silicon Valley could have been left in darkness.

Protective relays are defined by the IEEE as "relays whose function is to detect defective lines or apparatus or other power system conditions of an abnormal or dangerous nature and to initiate appropriate control circuit action" (IEEE 1992, 59). Their purpose is to first protect humans and second equipment (Leelaruji and Vanfretti 2011). Thus, a key criterion is the "immediate availability criterion [which] is necessary to avoid serious outages and damages to parts of or the entire power network" (Leelaruji and Vanfretti 2011, 2). This availability criterion means that a protective relay should be reliable and available only when it receives a signal stating that the protected unit is under stress, and send a signal to the circuit breakers to seal off faulty components (Leelaruji and Vanfretti 2011, 2). Reliability, a protection characteristic of protective relays, consists of dependability and security (Rifaa 2016). Dependability is "The facet of reliability that relates to the degree of certainty that a relay or relay system will operate correctly" (IEEE 1992, 23). In other words, the device "must operate when required" (Rifaa 2016, 28), backups have to exist, and testing of the device hardware and software should be done. Security, on the other hand, is defined as "that facet of reliability that relates to the degree of certainty that a relay or relay system will not operate incorrectly" (IEEE 1992, 65). Therefore, security concerns are related to "incorrect relay/device operation" (Rifaa 2016, 29), "including operating faster than intended" (NERC 2010, 2).

3. From What Threats to Protect? Defining the Cyber Threat

In 2007, Project AURORA, a staged cyberattack demonstrated by the Idaho National Laboratory (INL), "demonstrated how a cyberattack could be used to destroy power generation equipment" (Congressional Research Service 2018, 17). We pose that the INL demonstration risks becoming a national security nightmare for the United States and its allies due to the inherent vulnerabilities in protective relay system and related software and protocols of the substations they belong to. Cyber attacks against SCADA systems (such as Distributed Denial of Service Attacks, zero-days, and Advanced Persistent Threats) have already grown exponentially over the past decade (IBM 2015), increasing the possibility of cascading catastrophes, crucial outages to include regional blackouts, and potentially irreparable damage done to critical infrastructure and the stakeholders that it supports. Attacks against cyber-physical systems of the substations are posed to grow in a similar exponential fashion over the next decade. This is not a theoretical threat, but one that has already happened in 2015, when a cyber attack against Ukraine's distribution utility substations caused a grid power outage that affected over 225,000 customers for several hours. In 2016, another cyber attack managed to shut down several substations in Kiev, leading the United States government to assess the potential for a similar attack against the U.S. grid "as a possibility" (Congressional Research Service 2018, 1). Examples of advanced persistent threats that could be used against smart grids include variations of Stuxnet, Duqu, or Black Energy. For example, Disakil, an upgraded version of Black Energy, "is being linked to the Ukrainian power outages in December, 2015" (Leszczyna 2018, 70).

Smart grids, their intelligent substations, and respective protective relay systems were never designed with confidentiality and integrity in mind. For example, protective relay systems were designed only to cut off faulty equipment. The European Smart Grid Task Force changed this view when it redefined smart grids as "electricity networks that can efficiently integrate the behaviour and actions of all users connected to it: generators, consumers and those that do both in order to ensure an economically efficient, sustainable power system with low losses and high quality and security of supply and safety" (Sanz et al. 2013). However, the added safety component was added in the smart grid definitions by U.S. and the EU entities only after substations were already enhanced with vulnerable protective relay systems. Intelligent substations remained particularly vulnerable to cyber attacks due to their dependence on smart devices that are connected to communication networks (including connectivity to the Internet). "When enhanced communication is used in the distribution network, cyber-security becomes an essential part of the overall security. A secure product is not in itself sufficient, as potential vulnerabilities may arise from insecure integration into existing infrastructures. While a substation can form a separate, secured island for energy distribution, it must also provide an information firewall for parties communicating with the substation and the associated distribution network" (Valtari 2013, 22). Each poorly protected network connection of the substation "opens a potential entry for an attacker [exposing] every network layer and the technology used [to] a nearly 24/7 potential attacker activity" (Leszczyna 2018, 71). This is especially valid in the case of protective relay systems.

The majority of today's protective relay systems, also known as Intelligent Electronic Devices (IEDs), have a microprocessor, allow remote access (communications), and require software for configuration and monitoring (Kulkami and Mannazhi 2006; Leelaruji and Vanfretti 2011). Remote access to IEDs requires communication means through e.g. telecommunications and fiber optic cables (as described in figure 2). Achieving communication requires various protocols, divided into physical-based protocols and layer-based protocols (Kulkami and Mannazhi 2006; Leelaruji and Vanfretti 2011). Physical-based protocols are the most basic of protocols, however, they are also included into the layer-based protocols, which the Internet also uses (Kulkami and Mannazhi 2006; Leelaruji and Vanfretti 2011). Some of the protocols used in communications with substations include the 1) Distributed Network Protocol (DNP) 3.0, 2) ModBus, 3) IEC 61850, and 4) the Local Operating Network (LON) protocol. Apart from enabling communication, they have an additional trait in common: all have vulnerabilities that could be leveraged through a cyber attack. These cyber attacks can happen in two ways: 1) by taking advantage of "poorly configured and maintained" IEDs and 2) by taking advantage of "vulnerabilities built into the software or hardware" of the IEDs (Winkler and Gomes 2016):

1. Papa & Shenoi analysed the DNP 3.0 protocol and "identified 28 attacks and 91 attack instances" where the attacks could interrupt, modify and fabricate control system assets, but also lead to "obtain network or device configuration data to corrupting outstation devices and seizing control of the master unit" (East et al. 2009, 80).
2. An attacker exploiting the vulnerabilities in the ModBus protocol could "obtain information about field devices, interrupt the master unit, modify Modbus assets [and] lockout the master unit and seize control of the slave devices" (Huitsing et al. 2008, 42). Some of these attacks could lead to denial of service on the target (Benbenishti 2017).
3. The IEC 61850 vulnerabilities could lead an attacker to steal login credentials and conduct a denial of service attack. Stealing login credentials would give an attacker the ability "cause physical damage to the smart grid, for example they could trip circuit breakers and cause undue stress on the distribution network" (Wright and Wolthusen 2017, 249)
4. The LON protocol, based on ISO/IEC 14908, has the same inherent vulnerability: it is "broken by design" (Jovanovic and Neves 2015, 314), because it uses weak cryptography. An attacker can get access to smart meters and exploit them for malicious use.

A cyber attack exploiting the multiple vulnerabilities of protective relay systems could have devastating consequences, and unfortunately, there is no cheap and immediate solution in sight (other than redesigning substation configurations from the ground up). Even newer substation protection systems separated into "data processing, communication, voltage measurements, current measurements and breaker control" (Ljungberg 2018) components are not built with cyber security or cyber resiliency in mind, putting the entire grids at risk of blackouts caused by cyber attacks.

4. By What Means to Defend Forward? Protecting the Smart Grids in the Age of Artificial Intelligence

Countries like the United States of America, Norway, Holland (among others) are looking to the possibility of using Artificial Intelligence (AI) to enhance their grids resilience in their countries, to include cyber resiliency of their respective substations. Most of their grids, however, are composed of sections of updated (smart) technologies, with components that are over 30 years old, limiting the efficiency of AI systems implementation. In fact, most of the world's energy grids need modernizing, to include the US, where "the average age of power plants is over 30 years and of power transformers is over 40 years" (Wolfe 2017). Upgrading of the grid is, however, very challenging, as it is a long term, transformation process, in which incompatibilities between smart technologies and commercial off the shelf components are bound to occur (Kruger, Abu-Mahfouz, and Hancke 2015). These incompatibilities have made it even more difficult to implement comprehensive protection against cyber threats.

Nevertheless, AI can assist with this modernization of energy grids and we propose that AI capabilities should be integrated into both operations (the business side) and any cyber-physical components (the Operational Technology side) moving forward. The operations side is particularly important in predicting cyber vulnerabilities because AI technologies can assess the bulk of information coming to and from the energy grids, find those weak spots (or massively stressed spots), enabling grid operators to redistribute or change the flow of power to allow for those spots to be strengthened and updated over time. This also allows for a greater resilience to be built into the grid from the bottom up, with AI slowly becoming the backbone of all smart grids, assisting in rendering future cyber attacks against substations inefficient in causing a power outage regionally or even locally. AI can enable smart grid managers to have more processed data to make the best decisions for protecting the smart grid from cyber attacks. Instead of ignoring bulks of raw data, AI would be able to process it, highlight trouble spots, pinpoint weak places within the grid, and show micro trends (that a human may not otherwise pick up). With this more processed data and the ability to pick up on microtrends (through advanced analytics processes), AI would enable energy grid managers to spot potential disasters before they happen and stop them, or even slow them down (Zhang, Wu, and Chang 2018). This includes stopping a cyber attack before substations are shut down and blackouts occur. The more information about smart grids is processed and studied, the easier it becomes to prioritize cyber threats to their substations and relay systems. The study of these microtrends would create the ability for energy grid managers to understand and know how to better prepare for cyber attacks and create cyber resilience across their grids, to account for causes in fluctuations and issues in the grid that may not have been understood or known before (Cooke and Porter 2012, 206). This is critical for the early warning of cyber attacks against multiple substations (as it happened in Ukraine in 2015).

At the local, substation level, successful AI integration into the smart grid requires, however Centralized Protection and Control (CPC) in its intelligent substations. This is particularly because the protection and control functionality of substations "needs real-time process data and directly affects network safety. This functionality has strict requirements for reliability and cyber security" (Valtari 2013, 49). Ellevio, a major electricity distribution company in Sweden, has already considered implementing CPCs in its substations in order "to centralize as many of the station's processes as possible into a central unit, as opposed to having the station made up of several autonomous systems that work and communicate independently of each other" (Ljungberg 2018, 8). From a cyber security standpoint, the argument for CPC implementation is that protecting one central system is easier than protecting several autonomous systems that work and communicate independently of each other. While that may be the case, however, a successful cyber attack against a centralized system could have bigger negative consequences than one against a decentralised system. It could thus be a higher risk "to centralize too many functions as offensive techniques tend to evolve faster than defensive ones. As substations already are possible targets for malign intrusions it would be unfavourable to make them more attractive for potential cyber attacks" (Ljungberg 2018, 69).

While at the local levels AI and Centralized Protection and Control (CPC) may eventually become the ideal solution for protecting intelligent substations from cyber attacks, decentralisation of the overall grid will eventually become the ultimate solution for defending against big, regional blackouts caused by cyber attacks. Here, a major issue that governments must account for in cybersecurity planning and defending forward is a rising trend of private users (prosumers) selling excess power back to companies due to more green energy generation (wind, solar, water) being produced globally (Rathnayaka, Potdar, and Kuruppu 2011). This adds additional vulnerabilities that cyber adversaries may be able to leverage to create relay failures. AI can help assimilate all of these issues and show where the vulnerabilities to cyber attacks are, and provide the mitigation options to potential cascading catastrophes and outages resulting from a cyber attack. As more and more individuals create their own energy, consume less than they have, and sell to power companies the excess, AI will be needed to manage the added power intakes to the grid as well as the new prosumer connections to the grid to avoid creation of vulnerabilities. Especially because these connections will not just be taking energy but will also be giving energy, thus creating two way energy flow, which also means two way data flow, doubling the number of potential vulnerabilities. AI will also be needed to monitor the security and resiliency of all of these new connections to ensure that the security and resiliency of the smart energy grid is not compromised because of the additional connections. New connections do not always have the same level of security or technology as others and AI would be able to best handle all of the new data, maintain energy flow and keep security to allow for a more overall resiliency across the smart grid.

Previous research shows that in order to accommodate for a growing number of prosumers, local energy needs can be better supported through the use of microgrids where energy flow is continually monitored and adjusted by AI. This means that the use of AI to enhance cyber security will also move in the direction of smart grids. New battery and other storage technologies can be used to permit energy to continually flow within these microgrids despite major cyber attacks against local substations and potentially even assist with rerouting power during outages to better stabilize and ensure the smart grid is able to adjust for fluctuations and changes (Wolfe 2017). While decentralization of the grid may be necessary, substations' IEDs will still need to be managed centrally, in real-time in order to successfully implement AI systems and be able to account for and mitigate changes of fluctuation or address unexplained micro-trends in cyber attack mitigation strategies.

On top of IEDs, the Internet of Everything (IoE) introduces even more types of equipment and power draws throughout the smart grid. As more and more of these items come online, they produce not only unique power draws from the system, but also introduce greater variability to the security and resilience of the smart grid critical systems. As discussed earlier, many of these smart items do not have security in place for when they are connected to the internet, and can create an opening to allow viruses, worms, cyber attacks, and other hacks to infiltrate the smart grid's substations. With a sharp increase in vulnerable smart devices comes a massive risk to the security and resiliency of the grid. However, with AI, this security risk can be mitigated, as AI can monitor all the items connected to the grid for fluctuations, attacks, or even minor fluctuations and changes in the connections, security and anything else that would affect the resilience and security of the network. AI can thus be used to better manage the data flow, micro-trends, power fluctuations, and also for maintaining cyber security and cyber resiliency throughout the systems. These AI enabled smart microgrids can be used to better supply and maintain local power grids, while also to better maintain the cyber security and resilience of the smart grid and the overall grids themselves.

Until AI capabilities are implemented in every cyber-physical component of smart grids, standardization of cyber security efforts must be enforced through regulatory means. These regulations are necessary in an environment where "the number of documents concerning the standardization and regimentation of cyber security in the energy sector has increased tremendously" over the past decade (Nartmann, Brandstetter, and Knorr 2009). Without proper control of standards there is no standardization, which is needed for smart grids resiliency, given their cross-border interdependencies. For example, in the United States, the North American Electric Reliability Corporation Critical Infrastructure Protection (NERC CIP) requires operators to implement measures and controls "to protect the bulk power system from cyber threats. [The] choice of assessment method is left to the operator [and NERC CIP] doesn't indicate any particular methods either" (Leszczyna 2018, 76). Good standards for substations cyber resiliency include ISO/IEC 27000, IEC 62351, and IEEE 1686; as well as IEC 62443 (ISA 99), GB/T 22239, ISO/IEC 15408, IEC 61850, IEC 62056-5-3, ISO 15118, ISO/IEC 27019. With operators using different standards to implement measures and controls, the resiliency of the smart grid varies from city to city and/or state to state.

5. Conclusions

Throughout this paper we have discussed how smart grid operators can consider Baldwin's definition of security when assessing smart grid vulnerabilities. The smart grid is an extremely complex distributed system, with many operators, legal demands and with no proper control of standards. In Section 2 (what to protect), we discussed intelligent substation protection and control systems. These have been increasingly interconnected thanks to new information- and communications technology (meant to enable more efficient monitoring). However, the mix of old and new components, has increased the complexity of an already complex system. The role of telecommunications is extremely important, and a loss of that is highlighted in the Metcalf-case. The protective relay, which failed to function properly in the Manhattan July 2019-case, is also an extremely important IED. In Section 3 we mentioned some of the cyber threat actors, their tools and the impact their attacks had on the Ukrainian smart grid. In addition, we also showcased that all communications protocols used between IEDs and Master SCADA-systems have inherent vulnerabilities. In Section 4 (Defending Forward), we discussed how AI and related techniques can mitigate security risks to cyber-physical systems. Until then however, standardization of cyber security efforts must be enforced through regulatory means. One example of a regulatory means is to demand security-by-design IEDs and protocols for the smart grid. The security-by-design of a power grid today is focused on protecting humans and equipment, while there is a complete lack of security-by-design on protocols and related software and hardware to protect the power grid from humans and equipment.

References

Anwar, Adnan, and Abdun Naser Mahmood. 2014. "Cyber Security of Smart Grid Infrastructure." *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/1401.3936.

Baldwin, David A. 1997. "The Concept of Security." *Kokusaigaku Revyu = Obirin Review of International Studies* 23 (1): 5–26.

Barron, J., and M. Zaveri. 2019. "Power Restored to Manhattan's West Side After Major Blackout." *New York Times*, July 13, 2019. https://www.nytimes.com/2019/07/13/nyregion/nyc-power-outage.html.

Benbenishti, Liron. 2017. "SCADA MODBUS Protocol Vulnerabilities." https://www.cyberbit.com/blog/ot-security/scada-modbus-protocol-vulnerabilities/.

Benzarti, S., B. Triki, and O. Korbaa. 2017. "A Survey on Attacks in Internet of Things Based Networks." In *2017 International Conference on Engineering MIS (ICEMIS)*, 1–7. ieeexplore.ieee.org.

Cherp, Aleh, and Jessica Jewell. 2014. "The Concept of Energy Security: Beyond the Four As." *Energy Policy* 75 (December): 415–21.

Con Edison. 2019. "Con Edison Statement." Con Edison. https://www.coned.com/en/about-us/media-center/news/20190715/con-edison-statement-preliminary-findings-re-west-side-power-outage.

Congressional Research Service. 2018. "Electric Grid Cybersecurity." R45312. Congressional Research Service.

Cooke, Philip, and Julie Porter. 2012. "13 Smart RDAs in Europe: Advanced Practices." *Regional Development Agencies: The Next Generation?: Networking, Knowledge and Regional Policies*, 206.

East, Samuel, Jonathan Butts, Mauricio Papa, and Sujeet Shenoi. 2009. "A Taxonomy of Attacks on the DNP3 Protocol." In *Critical Infrastructure Protection III*, 67–81. Springer Berlin Heidelberg.

Hai-bo, Kong, X. I. E. Xiao-dong, Wang Ling-fei, S. U. N. Peng, and X. U. Bing. 2017. "Study on Relay Protection of Process Layer and Substation in Relay Protection of Intelligent Substation." *Journal of Electronic Research and Application* 1 (1): 19–26.

Huitsing, Peter, Rodrigo Chandia, Mauricio Papa, and Sujeet Shenoi. 2008. "Attack Taxonomies for the Modbus Protocols." *International Journal of Critical Infrastructure Protection* 1 (December): 37–44.

IBM. 2015. "Security Attacks on Industrial Control Systems." IBM Security. https://www.ibm.com/downloads/cas/4PE7DQ3G.

IEEE. 1992. "Standard Definitions for Power Switchgear." C37.100-1992. IEEE. https://ieeexplore.ieee.org/servlet/opac?punumber=2898.

Jovanovic, Philipp, and Samuel Neves. 2015. "Practical Cryptanalysis of the Open Smart Grid Protocol." In *Fast Software Encryption*, 297–316. Springer Berlin Heidelberg.

Kruger, C. P., A. M. Abu-Mahfouz, and G. P. Hancke. 2015. "Rapid Prototyping of a Wireless Sensor Network Gateway for the Internet of Things Using off-the-Shelf Components." In *2015 IEEE International Conference on Industrial Technology (ICIT)*, 1926–31. ieeexplore.ieee.org.

Kulkami, C. S., and N. Mannazhi. 2006. "Substation Automation System for 33/66 kV S/S at North Delhi Power Limited." In *2006 IEEE/PES Transmission Distribution Conference and Exposition: Latin America*, 1–6. ieeexplore.ieee.org.

Leelaruji, Rujiroj, and Luigi Vanfretti. 2011. "Power System Protective Relaying: Basic Concepts, Industrial-Grade Devices, and Communication Mechanisms." KTH Royal Institute of Technology. http://www.diva-portal.org/smash/get/diva2:464427/FULLTEXT01.pdf.

Leszczyna, Rafał. 2018. "Standards on Cyber Security Assessment of Smart Grid." *International Journal of Critical Infrastructure Protection* 22 (September): 70–89.

Ljungberg, Jens. 2018. "Evaluation of a Centralized Substation Protection and Control System for HV/MV Substation." UPTEC F 18026. Uppsala Universitet. https://pdfs.semanticscholar.org/8f33/4154dd6b2eeb10262085e0eb30f67fc0c9d4.pdf.

Muratori, Matteo, Beth-Anne Schuelke-Leech, and Giorgio Rizzoni. 2014. "Role of Residential Demand Response in Modern Electricity Markets." *Renewable and Sustainable Energy Reviews* 33 (May): 546–53.

Nartmann, Bernd, Thomas Brandstetter, and Konstantin Knorr. 2009. "Cyber Security for Energy Automation Systems--New Challenges for Vendors." In *The 20th International Conference and Exhibition on Electricity Distribution (CIRED)*. IET. https://digital-library.theiet.org/content/conferences/10.1049/cp.2009.0639.

NERC. 2010. "Reliability Fundamentals of System Protection." NERC. https://www.nerc.com/comm/PC/System%20Protection%20and%20Control%20Subcommittee%20SPCS%20DL/Protection%20System%20Reliability%20Fundamentals_Approved_20101208.pdf.

Rathnayaka, A. J. D., V. M. Potdar, and S. J. Kuruppu. 2011. "An Innovative Approach to Manage Prosumers in Smart Grid." In *2011 World Congress on Sustainable Technologies (WCST)*, 141–46. ieeexplore.ieee.org.

Rifaa, R. 2016. "Power System Protective Relays: Principles & Practices." In *PES/IAS Joint Chapter Technical Seminar*.

Sanz, A., P. J. Piñero, S. Miguel, and J. I. Garcia. 2013. "Real Problems Solving in PRIME Networks by Means of Simulation." In *2013 IEEE 17th International Symposium on Power Line Communications and Its Applications*, 285–90. ieeexplore.ieee.org.

Valtari, Jani. 2013. "Centralized Architecture of the Electricity Distribution Substation Automation-Benefits and Possibilities." Edited by Pekka Verho. PhD, Tampere University of Technology. https://trepo.tuni.fi/handle/10024/114194.

Winkler, Ira, and Araceli Treu Gomes. 2016. *Advanced Persistent Security: A Cyberwarfare Approach to Implementing Adaptive Enterprise Protection, Detection, and Reaction Strategies*. Syngress.

Wolfe, Franklin. 2017. "How Artificial Intelligence Will Revolutionize the Energy Industry." Harvard University Blog, Special Edition on Artificial Intelligence. August 28, 2017. http://sitn.hms.harvard.edu/flash/2017/artificial-intelligence-will-revolutionize-energy-industry/.

Wright, James G., and Stephen D. Wolthusen. 2017. "Access Control and Availability Vulnerabilities in the ISO/IEC 61850 Substation Automation Protocol." In *Critical Information Infrastructures Security*, 239–51. Springer International Publishing.

Zhang, Bing, Jhen-Long Wu, and Pei-Chann Chang. 2018. "A Multiple Time Series-Based Recurrent Neural Network for Short-Term Load Forecasting." *Soft Computing* 22 (12): 4099–4112.

Online Polarization, Radicalization, and Conspiracy

Petri Jääskeläinen[1] and Aki-Mauri Huhtinen[2]
[1]University of Tampere, Finland
[2]Finnish National Defence University, Finland
petri.jaaskelainen@tuni.fi
aki_huhtinen@hotmail.com

DOI: 10.34190/EWS.20.008

Abstract: During the last decade, digital technologies have promised equality and democracy, but social platforms such as Amazon, Facebook, Google, and Twitter are being (mis)used for political, economic and security control. Despite social and economic progress, hate and fear are being increasingly promoted online all over the world.. The gap between the use of propaganda by authoritarian and democratic governments is closing. Classically, authoritarian governments target propaganda at both their own populations and at the populations of other countries. Now, also democracies are using individuals to design and operate fake and highly automated social media accounts. The public conversations on social media have become more polluted and polarized. Conspiracy is one of the key elements of political propaganda. Viewing reality and the environment more as a rhizome without a single starting point or a rational goal has become a relevant framework to help build a better understanding of world online events. The idea of the Internet as a rhizome or the rhizomatic nature of networks is based on the idea of the non-linear, non-hierarchical, ever changing aspects of a biological organism (Deleuze and Guttari 1983). For example, information hacking and cybersecurity accelerate each other's development. There is no single, universal Internet, but rather a multitude of Internets. Social automation, machine learning, and artificial intelligence mean that propaganda is becoming more sophisticated and harder to track. If something extremely bad such as a terror attack happens somewhere in the world, often early warning signs can be found in the deep network. In our paper, we will analyse, how imageboards (an imageboard is a type of Internet forum that revolves around the posting of images, often alongside related text and discussion) such as 4chan and 8chan are linked to modern conspiracy theories such as QAnon, an imageboard-born conspiracy theory about the deep state targeting Donald Trump, and how these rhizomes can radicalize people. Imageboards are a certain type of online platform which consist mainly of text and images. The users are anonymous, and they do not use an avatar or other form of identification on their posts, which consist of memes, imageboard-exclusive language and other forms of communication. The focus of this paper will be on how the language, jokes, memes, tweets and interaction on these platforms help users to slide into more radical theories and thoughts through the game-like nature of the interaction. Imageboards rely on the anonymity of individual actors, and this allows more freedom to express one's own views without being judged. These platforms consist of anonymous posters and a combination of text and images; nothing is actually personal there. If individual users want to rise up from the sea of anonymous users, they must refer to their crowd to gain recognition by using certain symbols, signs, words and jokes. Using these symbols may give the radicalized mind more acceptance and a chance to spread more radicalization. In summary, we aim to show that online polarization, radicalization, and conspiracy are rhizomatic in nature and constantly come into existence without any specific place, time or reason.

Keywords: conspiracy, polarization, radicalization, rhizomatic, social media

1. Struggle between rhizomatic and hierarchical networks

Online social networks are increasingly becoming areas of criminality, misuses, hate speech, harassment, loss of reputation or money, and loss of trust in the system (Jutterström 2019, 207; López-Fuentes 2018). The problem lies not only with lone acting hackers, leakers or whistle-blowers, but also increasingly involves corporations and governments using the lack of trust and confidence as a strategy for managing risk and to make a profit (Buchanan 2009; Haggart et al. 2019). For example, information hacking and cybersecurity accelerate each other's development. The news of a fragile point in a network risks public trust in the security process and online products (Burkart and McCourt 2019; Harris 2014). During 2018, Facebook had to testimony in the U.S. Congress and this episode has spelled the end of the optimism of social media (Napoli 2019, 1). Facebook has become a symbol of a phenomenon that never imagined or intended to be misused. It has also become a rhizome without centralized control. The flagships of information ecosystems, such as YouTube, Twitter, Instagram, Snapchat, and WhatsApp have also been linked to the dark side of the information network. Social media never intended to be a news media outlet, but they are operating as such (Napoli 2019, 4). This is exactly the nature of rhizomatic actions.

States and their governments can intrude into the strategically important networks of other states to test their own network security, but at the same time they risk political stability and open the gate to possible escalation. Information intrusions can collect valuable information on what tools of influence or weapons a potential adversary has deployed (Buchanan 2017, 3-6). Overall, the distinction between weapons and tools has never been quite clear, but nowadays most weapons systems require computer connections and the same kind of software that is used in the Internet of Things.

Decentralized technology does not guarantee decentralized outcomes (Schneider 2019, 280). Journalistic, editorial, and media consumption decisions are increasingly being influenced by complex, data-driven algorithms developed by large teams of engineers and computer scientists. Political and social decision-making is based more on algorithmic tools than traditional human actors or organizations (Napoli 2019, 6). However, the so-called dark network accessed with special browsers such as Tor (an acronym for The Onion Router), is not only an extreme, or criminal communication platform, but is also useful and provides access to truthful information. For example, the BBC "has made its international news website available via the Tor network, in a bid to thwart censorship attempts from China, Iran and Vietnam in order to try to block access to the BBC News website or programmes," (BBC News 2019).

On an international level, also the gap between the use of disinformation by authoritarian and democratic governments is vanishing. Classically, authoritarian governments have targeted disinformation at both their own populations and at populations in other countries. Now, democracies too are using individuals to design and operate fake and highly automated social media accounts. Social automation, machine learning, and artificial intelligence mean that propaganda is becoming more sophisticated and harder to track (Woolley and Howard 2018, 12; Conley 2009). The public discourse ranging from the high international diplomatic level to ordinary people chatting on social media has become more polluted and polarized (Woolley and Howard 2018, 2). For example, in the U.S.A. there is the evidence that the more aggressive and personalized political targeting of potential voters has influenced voters to vote more extreme representatives to Congress (Benkler et al. 2018, 301). President Trump is not totalitarian, but he presents one version of reality to his followers and another version to the outside world. In addition, his messages are directed towards "the uprooted, isolated and alienated individuals of modern mass society who wish nothing more than to escape from reality," (Sanders 2019, 757).

2. The rhizome metaphor as new way to understand information influence

The concept of normal and the concept of extreme cannot exist without each other. Paradoxes, tension, dialectics, duality, and contradictions can be seen as a normal part of organizational existence (Cunha and Clegg 2018, 14, Hernes 2017). An increasingly rhizomatic and fragmented world emphasises adversaries or enemies who define 'us'. This pressure contrasts cultural and religious values and gives rise to a nationalist brand of populism (van Beek 2019, viii). However, extremism as a belief does not directly mean that it will conclude in an act of violence (Scaife 2017, 30). Anyone can select another person's offensive memes, texts, or pictures anonymously in order and use them in a violating manner in a rhizomatic network. It is true that popular online social networks today are centralized and based on a client-server paradigm. However, there is no single, universal Internet, but rather, a multitude of Internets. In addition, the freedom of expression, which has been seen as the 'lifeblood of democracy', has become increasingly extremized and 'deadly to democracy' (Scaife 2017, 46-47). Deleuce and Guttari (1986, 4) described this as the 'two-headed' phenomenon of rhizomatic processes.

The rhizomatic way of seeing organizations is not about celebrating chaos over order, rather it is about analytically exploring organizations in their extremization. The transformation from a classical organization to rhizomatic organization is not only a structural issue but also a performative one (Kornberger 2006, 58-70; Hess 2008; Munro and Thanem 2018; Robinson 2009). The metaphor of the *rhizome* opens the possibility to progress in multiple directions and not just one specific direction. Unlike something that 'is' or exists, a rhizome expresses more the concept of 'and' (Robinson 2017, 62). For example, some of the key political leaders at an international level charm their audiences with an ideological schizophrenia, using opposing values and ideologies in their arguments and presenting them a rhizomatic manner (Benkler et al. 2018, 306). To become organized things, issues or actions need some kind of representation, and this representation is rooted in the practice of a particular community with concrete people in a concrete world. The gap between the concrete world and its representation needs improvisation, which sometimes becomes a figure of extremization (Tsoukas 2019). For example, the power in Nazi Germany was based on the constantly changing coalitions and competition between the leaders within an inner circle without any specific institutions. The structurelessness of this regime was the source of its strength and "it allows the leader to play rivals off one against another, thereby preventing the emergence of any opposition or rival centres of power," (Sanders 2019, 757). We can see the same idea of this sort of structurelessness in Putin's and Trump's leadership and management styles. "Totalitarian leaders cannot survive if the regime stabilizes," (Sanders 2019, 758).

In the military 'world', the rhizomatic phenomenon can be seen in swarming, where opposing forces are not of the same size, or capacity as the defender's own units. In such situations, swarming involves the use of a decentralized force against an opponent (Edwards 2000). In a terrorist swarming attack, the general objectives of operational cells are agreed in a coordinated manner, but not continuously controlled by the core organization. Swarming operations have included multiple and near-simultaneous attacks. A typical civilian application of a rhizome is the Uber taxi sharing company that it is a technology company rather than a transportation company. Their slogan "We don't transport goods or people – We just facilitate it," is typical example of the benefits which can arise from rhizomatic networks (Napoli 2019, 6).

A rhizomatic platform typically does not offer any original content, but merely facilitates, distributes or synchronizes someone else's created content (Napoli 2019, 7). The borderline between classical media and journalistic companies and Google, Facebook or other new 'technology companies' has become blurred. Previously, the traditional media companies created new content, while the 'technology' companies distributed someone else's content to the audience. However, there is no need for human editorial intervention in technology-driven facilitators of content creation and dissemination (Napoli 2019, 11).

Before we discuss data collection, it is important to explain that the metaphor of the rhizome has five principles. The first principle of a rhizome is the connection. Any point on a rhizome can be connected with any other and must be. A rhizome makes it possible that two beings with nothing to do with one another can be connected (Deleuze and Guttari 1983, 11). The second principle of the rhizome is heterogeneity. A rhizome is always open to connections. The third principle of the rhizome is multiplicity. Multiplicity refers to the size or dimension, which cannot increase or decrease without changing its nature (Deleuze and Guttari 1983, 14-15). The fourth principle of a rhizome is rupture. A rhizome can be cracked and broken at any point, but it can start off again following one or another of its lines (Deleuze and Guttari 1983, 18-20). The fifth principle of a rhizome is mapping. Contrary to a tracing, which always returns to the same point, a map has multiple entrances. A map is a matter of performance (Deleuze and Guttari 1983, 25-26).

3. Imageboards as rhizomes

The rhizomatic nature of the Internet allows everyone from single individuals to major corporations and governments to engage in persuasion and to have influence. One phenomenon that shows the influence of multiple users combined is trolling (making deliberately provocative or offensive online posts). Trolls (those who troll others) are identified fully with the collective. The trolls' motto is that 'none of us are as cruel as all of us'. Trolling has no moral limitations and no criterion other than to ridicule someone or something. Because nothing is sacred, this makes trolls antisocial. In hierarchical societies, the fear of ridicule tends to make people clam up, but trolls do not care; their community is an anonymous and networked swarming rhizome (Seymour 2019, 113). The term *troll* means 'wandering aimlessly' or 'searching for a game', but at the same time is represents a challenge to those who are the authorities or set themselves above others (Merrin 2019, 202).

Trolling is present on rhizomatic platforms such as 4chan and 8chan imageboards. These imageboards rely on the anonymity of individual actors, and this gives more freedom for the users to express their personal views without being judged. While the individuals attend discussions on these boards they become a collective of anonymous people. While these individuals are connected by an online network, they do not have anything else in common most of the time, at least visibly since they are all just anonymous posters and a combination of text and images.

For example, in the book Norwegian Tragedy (about the largest ever terrorist attack in Norway), the author Aage Storm Borchgrevink shows that the perpetrator of the attack, Anders Behring Breivik, was an outcast who wanted to be a part of a community. At first, he wanted to be a hip hop gangster, then became involved in the far-right. Borchgrevink (2014, 106) mentions that far-right gangs take care of outcasts who were not welcome elsewhere. 4chan, 8chan and similar image forums have served a similar purpose by collecting people and opinions who are not wanted elsewhere. The far right tries to push the range of topics tolerated in public discourse (the Overton window) (Daniels, 2018, 62). For this reason, the memes created on image boards aim to become mainstream.

One example meme is the claim that Jewish people are behind every bad thing that happens. This is an old antisemitic rhetoric based back on *The Protocols of the Elders of Zion*, an antisemitic classic. On image boards, there are often campaigns targeting Jewish people. Sometimes these campaigns spread out of the boards and onto mainstream platforms such as Twitter (The Next Web, 2019). Using the Jewish people as well-known symbols makes the imagined community of antisemites stronger and that way it pushes the Overton window even farther. It is argued mainly by the people voicing these views that restricting these ideas is an attack against the freedom of speech.

While people consume what is on these platforms, their view of the world may shift towards the extreme as the Overton window moves. The normal on these platforms can become quite radical. The content of 4chan is eclectic and varying (Daniels 2018, 64). Since the posters are anonymous, nothing is personal there compared to platforms such as Facebook. Individuals who do something different may stand out from the crowd of anonymous users. The Christchurch shooter Brenton Tarrant in his manifesto referred to multiple memes which originated on image boards. His manifesto states that he is one of them and this gives him more credibility in the eyes of the users.

If individuals want to rise up from the sea of anonymous users, they must refer to their crowd to gain recognition. The use of the OK-sign as a symbol of the far right, referring to conspiracy theories such as QAnon, and spreading them using memes and other symbols are all part spreading the radicalization. They are used in the same way as ISIS used hand gestures (Pri.org, 9/2014) or created their own music. These elements are meant to create a sense of community which the users are a part of. Using these memes even ironically, which is one thing that dominates the 'alt-right' loosely connected online-based white nationalist movement, may lead to more radicalization since that social world is built upon certain symbols, signs, words and jokes. As Balca Arda wrote " (...) an Internet meme becomes a performative show that is collaboratively created on social media through the desire for participation," (2015, 80). Using such symbols may give radicalized thoughts more acceptance and a chance to spread more radicalization.

Sometimes the symbols may not be originally from the alt-right. Miller-Idriss researched three cases to demonstrate this point. These were brands created by or for far-right consumers; they included brands, logos and such without an original connection to the far-right but which has been adopted by them and products which deliberately or accidentally deployed their symbols (2019, 124). Miller-Idriss argues that both social media and the Internet have accelerated the spread of old symbols and online communities have helped to create completely new symbols (2019, 131). The imageboard platform itself makes it possible to make these symbols known and from there they can leak into the web outside. The ironic and unironic sharing of these symbols and memes for fun makes them even better known. Some of the symbols are created on the platforms mentioned, and they are a part of the way people use media: they want to create things.

Writer Walter Kirn wrote about QAnon in Harper's Magazine (6/2018). He points out the core essence of QAnon and other Internet conspiracies by saying, "The audience for Internet narratives doesn't want to read, it wants to write. It doesn't want answers provided; it wants to search for them." This is one major reason that QAnon is successful. It combines many theories into one and gives the "players" clues to follow. They can interpret them according to their beliefs and search to find more information. QAnon has led to a few violent acts in the US so it is safe to assume it radicalizes people.

Yet, it is hard to follow the process of rhizomatic extremization. Anonymous online boards are full of pretty much everything you can imagine. The only way to follow the process is to find some examples to follow. QAnon, with its cryptic messages, which has caused people to carry out violent acts, may be one of them. Even as a theory focusing on the US, it has spread all around the world. While mainstream media follows the movement critically (for example Vice 7/2019) it still gathers a large following.

Another example of the radicalization of imageboard culture can be found in the conversations on the neo-Nazi forum Iron March. The forum has been a birthplace for violent neo-Nazi movements such as the Atomwaffe Division. The forum's contents were leaked last year, and activists created a database of it (ironmarch.exposed). People on many imageboards call themselves newfags and oldfags, based on how long they have been using the platform. The same terms were used on Iron March. One can also find many posts by searching for memes and other terms used by imageboard writers. This indicates that these users are active on both kinds of platforms and other places such as Reddit. The difference between a neo-Nazi forum such as Iron March and an imageboard like 4chan lies in the anonymity. Many of the users from the leak were identified, since the platform required them to add personal information. Imageboards, instead, are anonymous and they do not need any kind of registration, usernames or so on, so the users can be more heterogeneous than on a specified forum. Their connection may simply be the platform itself.

As previously mentioned, the people participating in discussions on imageboards may become a single collective of anonymous users. They are heterogenous, but the shifting of the Overton window is present. "Without negative emotional reactions to offensive language, people are more willing to believe in its content and to treat its contents as guiding principles," (Bilewicz, Soral & Winiewski, 2017, 10). If a platform's discussion shifts in a direction where hate-speech is not punished but encouraged, it can slide into a more radical direction.

Since 4chan is anonymous, the participants of a conversation must demonstrate some degree of subcultural literacy (Tuters 2019, 40). That literacy is filled with ironic memes and "sensationalist behavior that helps one to be noticed" (Tuters) and it feeds certain behavioral patterns. This makes the alt-right's discussion on 4chan's /pol/ board very effective.

Tuters points out that the users of imageboards use trolling to explain their actions. They say they are sort of role-playing (2019, 38). To them, violent threats or other forms of harassment can be thought of as an ironic game and it should not be taken seriously. Since Bilewicz & co argued that hate speech without negative reactions can make people believe the content, it is likely that it has already happened on many imageboards. The ironic content blends with real extremist thoughts creating a place, from where they can move to other platforms, such as the Iron March-forum. The imageboards instead work as a rhizome, and are always open to new connections if they play by the platform's rules.

Tuters (2019, 46) also writes "In spite of all the ironic posturing, what we should not overlook is the extent to which these communities also represent the concerns of those who perceive their identities as under threat." This is one of the key thoughts about why and how the users can and do organize themselves on imageboards. They create a mutual enemy, such as an elite, or the Jewish people or some other entity, and fight against it. For example, the QAnon conspiracy theory is based on fighting against an enemy. There is not as much ironic posturing present and the participants take the threat seriously. This is another way to play a role-playing game.

4. Conclusion

Automated algorithms classify, filter, and prioritize content created by others based on values which are internal to the system and the preferences and actions of users. The decisions and relationships are encoded into rhizomatic systems. To engage in editorial and gatekeeping related to the flow of information, social media applications are increasingly key players in the dissemination of political news and values (Napoli 2019, 12-13). The algorithmically controlled media is not exactly the 'black box' of a controlled system, but it is more akin to a rhizome without a clear understanding of the information production, dissemination or consumption processes (Napoli 2019, 53). A key issue concerning the lack of understanding in these processes is the exponential speed of the algorithms. Human beings have great difficulty understanding how these rhizomatic algorithms operate without computing assistance and how these operations affect the flow of information and news.

Sometimes social media's algorithms appear to prefer shocking content. For example, on YouTube's recommendations list there were a clear bias towards recommending conspiracist videos (Guardian 2/2018). The algorithms can also have racial biases, for example, in hate speech detection (Sap & co. 2019, 1672). These kinds of examples are part of the reason that the content leaking out of imageboards can be successful on other platforms too. However, it has been proven that there is no empirical evidence proving that concerns about filter bubbles are correct (Zuiderveen Borgesius & co 2016, 10). Based on this, the filter bubbles caused by algorithms should not be viewed as the only reason for online radicalization. However, extreme content can lure users from other services to join image boards, where the speech seems to be freer. This makes rhizomatic platforms ever-changing as new members leave and join the platform. When those who join bring their radicalized worldview with them, it can eventually shift the whole tone of discussion within the platform.

The rhizomatic nature of the Internet gives possibilities for individuals and major corporations and governments to engage in persuasion and have an influence. The act of trolling is present for example on rhizomatic web platforms such as 4chan and 8chan imageboards. On image boards, there are often campaigns targeting individuals, groups or organizations in order to polarize the discussion. These anonymous boards are full of pretty much everything one can imagine, especially highly racist, fundamental, pornographic or violent materials. The only way to monitor the process is to find some examples to follow. Therefore, induction is the key means to track the dark side of Internet. This requires a lot of human resources until automatized and artificial intelligence tools are advanced enough to help with this sort of monitoring.

References

Arda, B. (2015) The Construction of a New Sociality through Social Media: The Case of the Gezi Uprising in Turkey, *Conjunctions. Transdisciplinary Journal of Cultural Participation*, 2(1), pp. 72-99. Available at doi: 10.7146/tjcp.v2i1.22271.

Borchgrenvik, A. S. (2012). *Norjalainen tragedia*. Finnish ed. Kuopio. Oy Scanria Ab 2014.

Benkler, Y., Faris, R., and Roberts, H. (2018) Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics, [ebook], New York, Oxford University Press. Available at: https://www.oxfordscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624.

BBC News (2019) *BBC News launches 'dark web' Tor mirror*, [online], [viewed 23 October 2019]. Available from https://www.bbc.com/news/technology-50150981.

Buchanan, I. (2009) Deleuze and the Internet, in M. Poster, D. Savat, eds. *Deleuze and New Technology*, Edinburgh, Edinburgh University Press, pp 143-160.

Buchanan, B. (2017) The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations, [e-book], Oxford Scholarship Online. DOI:10.1093/acprof:oso/9780190665012.003.0001.

Burkart, P. & McCourt, T. (2019) *Why Hackers Win: Power and Disruption in the Network Society*, US, University of California Press.

Conley, V.A. (2009) Of Rhizomes, Smooth Space, War Machines and New Media, in Poster M. and Savat D. eds. *Deleuze and New Technology*, Edinburgh, Edinburgh University Press, pp 32-44.

Cunha, M.P. & Clegg, S. (2018) Persistence in Paradox, in Farjoun M., Smith W., Langley A., and Tsoukas H. eds. *Dualities, Dialectics, and Paradoxes in Organizational Life,* Oxford Scholarship Online, 14-34. DOI: 10.1093/oso/9780198827436.001.0001.

Daniels, J. (2018). The algorithmic rise of the far-right. Contexts Volume: 17 issue: 1, page(s): 60-65 Article first published online: April 3, 2018; Issue published: February 1, 2018. https://doi.org/10.1177/1536504218766547

Deleuze, G. & Guattari, F. (1986) *Nomadology: The War Machine*, Semiotext(e).

Deleuze, G. & Guattari, F. (1983) *On the Line*, New York, Semiotext(e).

Edwards, S.J.A. (2000) *Swarming on the Battlefield: Past, Present, and Future*, [online], RAND Corporation, Santa Monica, CA. Available at: https://www.rand.org/pubs/monograph_reports/MR1100.html.

Guardian (The) (2018) *How an ex-YouTube insider investigated its secret algorithm*, [online], [2 February 2019]. Available at: https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot.

Haggart, B., Henne, K. & Tusikov, N. eds. (2019) *Information, Technology and Control in a Changing World. Understanding Power Structures in the 21st Century*, [e-book], Springer International Political Economy Series. Available at: https://doi.org/10.1007/978-3-030-14540-8_1" https://doi.org/10.1007/978-3-030-14540-8_1.

Hernes, T. (2017) Process as the Becoming of Temporal Trajectory, in Langley A. and Tsoukas H. eds. *The SAGE Handbook of Process Organization Studies*. London, Sage.

Hess, A. (2008) Reconsidering the Rhizome: A Textual Analysis of Web Search Engines as Gatekeepers of the Internet, in Spink A. and Zimmer M eds. *Web Search, Multidisciplinary Perspectives*, [e-book], Berlin, Springer. pp 35-50. Available at: https://doi.org/10.1007/978-3-540-75829-7" https://doi.org/10.1007/978-3-540-75829-7.

Just Security (2019) *1. August 2019 Trump's Encouraging QAnon May Result in Violence—Just ask the FBI*, [online]. Available at: https://www.justsecurity.org/65659/trumps-encouraging-qanon-may-result-in-violence-just-ask-the-fbi/.

Jutterström, M. (2019) Problematic Outcomes of Organization Hybridity: The Case of Samhall, in Alexius S. and Furusten S. eds. *Managing Hybrid Organizations Governance, Professionalism and Regulation*, [e-book], Cham, Palgrave MacMillian, Available at: https://doi.org/10.1007/978-3-319-95486-8" https://doi.org/10.1007/978-3-319-95486-8.

Kornberger, M., Rhodes, C. and Bos, R. (2006) The Others of Hierarchy: Rhizomatics of Organising, in Fuglsang M. and Bent Meier Sorensen eds. *Deleuze and the Social*, Edinburgh, Edinburgh University Press, pp 58-74.

López-Fuentes, F. de A. (2018) Decentralized Online Social Network Architectures, in Özyer T., Bakshi S., Alhajj R. eds. *Social Networks and Surveillance for Society, Lecture Notes in Social Networks*, [e-book], Cham, Springer. Available at: https://doi.org/10.1007/978-3-319-78256-0_1.

Merrin, W. (2019). President Troll: Trump, 4Chan and Memetic Warfare, in Happer C., Hoskins A. and Merrin W. eds. *Trump's Media War*, [e-book], Cham, Palgrave Macmillan. *Available at:* https://doi.org/10.1007/978-3-319-94069-4_4" Available at: https://doi.org/10.1007/978-3-319-94069-4_4.

Miller-Idriss, C. (2018). What Makes a Symbol Far Right? Co-opted and Missed Meanings in Far-Right Iconography. In Maik Fielitz, Nick Thurston (Eds.), *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US* (pp. 123–136). Bielefeld: transcript Verlag. https://doi.org/10.14361/9783839446706-009

Munro, I. & Thanem, T. (2018) Deleuze and the Deterritorialization of Strategy, *Critical Perspectives in Accounting*, Vol. 53, pp 69-78. DOI**:** 10.1016/j.cpa.2017.03.012.

Robinson, K. (2017) Gilles Deleuze and Process Philosophy, in Langley A. and Tsoukas H. eds. *The SAGE Handbook of Process Organization Studies*, London, Sage, pp 56-70.

Public Radio International (PRI) (2014) *ISIS has a new hand sign — and it means far more than 'We're #1',* [online]. Available at: https://www.pri.org/stories/2014-09-04/isis-has-new-hand-sign-and-it-means-far-more-we-re-1.

Robinson, K. ed. (2009) *Deleuze, Whitehead, Bergson. Rhizomatic Connections*, New York, Palgrave Macmillan.

Sap, M.; Card, D.; Gabriel S.; Choi, Y.; Smith, N. A. (2019) The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Pages 1668–1678. DOI: 10.18653/v1/P19-1163.

Sanders, P. (2019) Leadership and populism: A parallel reading of Hannah Arendt and Franz Neumann, *Leadership* 2019, Vol. 15(6), pp 750–767. DOI: 10.1177/1742715019837807.

Scaife, L. (2017) *Social Networks as the New Frontier of Terrorism. #Terror*, New York, Routledge.

Schneider, N. (2019) Decentralization: an incomplete ambition, *Journal of Cultural Economy*, Vol 12(4), pp 265-285. DOI: 10.1080/17530350.2019.1589553.

Seymour, R. (2019) *The Twittering Machine*, London, Indigo Press.

Soral, W.; Bilewitcz M.; Winiewski, M. (2018) Exposure to hate speech increases prejudice through desensitization. Aggressive Behavior Volume 44, Issue 2 March/April 2018. Pages 136-146. https://doi.org/10.1002/ab.21737

The Next Web (2019) 4chan users pose as Jews to peddle anti-Semitic propaganda on Twitter, [online], (22 August 2019). Available at: https://thenextweb.com/socialmedia/2019/08/22/4chan-users-are-posing-as-jews-to-peddle-propaganda-on-twitter/?utm_source=copypaste&amp;utm_medium=referral&amp;utm_content=4chan%20users%20pose%20as%20Jews%20to%20peddle%20anti-Semitic%20propaganda%20on%20Twitter&amp;utm_campaign=share%2Bbutton

Tsoukas, H. (2019) *Philosophical Organization Theory*, Oxford Scholarship Online, DOI:10.1093/oso/9780198794547.001.0001.

Tuters, M. (2018). LARPing & Liberal Tears. Irony, Belief and Idiocy in the Deep Vernacular Web. In Maik Fielitz, Nick Thurston (Eds.), *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US* (pp. 37–48). Bielefeld: transcript Verlag. https://doi.org/10.14361/9783839446706-003

van Beek, U. ed. (2019) *Democracy under Threat. A Crisis of Legitimacy?* Cham, Palgrave Macmillan. https://doi.org/10.1007/978-3-319-89453-9.

Vice (2019) *People Tell Us How QAnon Destroyed Their Relationships*, [online], (11 July 2019). Available at: https://www.vice.com/en_us/article/xwnjx4/people-tell-us-how-qanon-destroyed-their-relationships

Woolley, S.C. & Howard, P.N. (2018) Introduction. Computational Propaganda Worldwide, in Woolley S.C. and Howard, P.N. eds. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*, [e-book] Published to Oxford Scholarship Online. DOI:10.1093/oso/9780190931407.001.0001.

Zuiderveen Borgesius, F.J., Trilling, D., Möller, J., Bodó, B., de Vreese, C.H. and Helberger, N. (2016) Should We Worry about Filter Bubbles? *Internet Policy Review*, Vol. 5 (1). Available at: https://doi.org/10.14763/2016.1.401

An Ontology for the South African Protection of Personal Information Act

Y. Jafta[1], L. Leenen[1,2], and P. Chan[3]
[1] University of the Western Cape, South Africa
[2] The Centre for Artificial Intelligence Research (CAIR), South Africa
[3]Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
2858132@myuwc.ac.za
lleenen@uwc.ac.za
kchan@csir.co.za

DOI: 10.34190/EWS.20.054

Abstract: The protection and management of data, and especially personal information, is becoming an issue of critical importance in both the business environment and in general society. Various institutions have justifiable reasons to gather the personal information of individuals but they are required to comply with any legislation involving the processing of such data. Organisations thus face legal and other repercussions should personal information be breached or treated negligently. Most countries have adopted privacy and data protection laws or are in the process of enacting such laws. In South Africa, the Protection of Privacy Information Act (POPIA) was formally adopted in 2013 but it is yet to be implemented. When the implementation of the Act is announced, role players (responsible parties and data subjects) affected by POPIA will have a grace period of a year to become compliant and/or understand how the Act will affect them. One example of a mandate that follows from POPIA is data breach notification. This paper presents the development of a prototype ontology on POPIA to promote transparency and education of affected data subjects and organisations including government departments. The ontology provides a semantic representation of a knowledge base for the regulations in the POPIA and how it affects these role players. The POPIA is closely aligned with the European Union's General Data Protection Regulation (GDPR), and the POPIA ontology is inspired by similar ontologies developed for the GDPR.

Keywords: Legal Ontology, Data Protection, Information Privacy, Protection of Personal Information Act, General Data Protection Regulation

1. Introduction

The rapid increase in connectivity and gathering of data have shifted the focus of legislators in various jurisdictions to the protection of personal information. This shift in focus has an impact on the privacy rights of individuals. Most countries have adopted, or are in the process of, adopting privacy laws. The objective of legislation with regards to data protection and privacy is to ensure the security of individuals' personal information in jurisdictions (Bartolini et al., 2015b). The European Union (EU) introduced the Data Protection Directive (DPD) as early as 1995 (Birnhack, 2008). The EU revised this directive into the General Data Protection Regulation (GDPR) in 2015 and updated it in 2018 on the reformed it (Srncova et al., 2019). The United Kingdom adopted the Data Protection Act in 1998 (United Kingdom Government Gazette, 1998) in addition to the EU DPD and this was implemented in 2000 (Botha et al., 2017). The United States of America does not have a single comprehensive law regulating the use and collection of personal information but has enacted several privacy laws since 2001 (Information Shield, nd). In South Africa, the Protection of Personal Information (POPI) Act was enacted on November 26th, 2013 (South African Government Gazette, 2013) although the full enforcement date is yet to be determined by the country's privacy regulator. South African businesses, government departments, and civil society is in the process of complying with the Act, but are facing implementation challenges and lack of awareness of the impact of the POPI Act. The Act aims to align the regulation of personal information in South Africa with international standards and the South African constitution. Embedded in section 14 of the Constitution of the Republic of South Africa, is the right to privacy of each South African citizen. The right to privacy includes a right to protection against the unlawful collection, retention, dissemination, and use of personal information" (Constitution of the Republic of South Africa, 1996).

While businesses have justifiable reasons to acquire personal data as information assets to reach their business goals, they are required to comply with any legislation involving the processing of such data (Bartolini & Mathuri, 2015a). Nevertheless, such information is susceptible to abuse. With the rise of social media, businesses have new means to gain traction in the information space. A prime example of this was the recent "Facebook Cambridge Analytica data scandal". Cambridge Analytica procured the personal information of countless individuals' Facebook profiles in the absence of their consent and used it for political purposes, incurring gross breaches of privacy (Common, 2018). In the wake of this, the EU reformed its data protection law, the General Data Protection Regulation (GDPR). These changes became effective on 25 May 2018 (Srncova et al., 2019).

Whilst enforcement of the South African POPI Act 4 of 2013 (POPIA) is yet to commence, the regulations to the Act were only published on 14 December 2018 (South African Government Gazette, 14 December 2018). Once implementation is confirmed, there will be a one-year grace period for entities to become compliant. The POPIA seeks to protect the right of privacy that applies to individuals and juristic entities (referred to as data subjects) by the establishment of strict guidelines on how to obtain and process information (South African Government Gazette, 2013). These guidelines affect organisations (referred to as the responsible party) and data subjects in various ways, and it is therefore important that these role players are well informed with regards to the implications. It is with this in mind that a knowledge base, through a semantic representation, for legislation of the POPIA will be valuable for assisting with the education of both the data subjects and organizations on the POPIA. The POPIA is closely aligned with the European Union's General Data Protection Regulation (GDPR), and the POPIA ontology is inspired by similar ontologies developed for the GDPR (Pandit et al., 2018, Miles et al, 2007).

This paper provides the analysis, design, and implementation of a basic ontology on the data protection domain with regards to the POPIA. The purpose of the POPIA ontology is to support understanding and awareness of the Act, and for future versions to be integrated into systems where automated compliance checking may be required. Section 2 provides a background on semantic technologies, their importance, and influences in the scope of knowledge bases and information formalisation. Section 3 describes related work on legal domain ontologies concerning the privacy and protection of data. Section 4 outlines the functional and non-functional requirements. The methodology for developing the ontology is discussed in section 5. Section 6 outlines the design, implementation, and evaluation of a prototype POPIA ontology and the conclusion and future development follow in the last section.

2. Semantic Technology

One of the biggest problems we face today is an overload of information. This is evident in various domains as the availability of large-scale information is more abundant than ever before. In the context of the World Wide Web (WWW), search engines have made rapid progression in handling and providing vast amounts of information. However, search engines are struggling to identify relevant results for search terms (Benjamins et al., 2002). Semantic technology is regarded to be the single most important factor for advancing the Web; it encodes meanings separately from data. It is able to deal with large data sets and link them together via self-describing interrelations, allowing it to be processed by machines. It is viewed as the leading framework to deal with the diverse and huge size of assets on the Web. The Semantic Web is an expansion of the present Web, wherein data is given unambiguous meaning. (Sheth et al., 2004). The ability to make good decisions efficiently is essential in allowing organisations to better manage and process large amounts of data.

An ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts (Man, 2013). It is the process of formalising knowledge and expressing the concepts and their relations in a given domain. An ontology defines a common vocabulary for researchers who need to share information in a domain (Noy & McGuinness, 2001). Ontologies can represent vocabularies to investigate the connections between various terms, enabling machines and people to decipher their "meaning" unambiguously (Horrocks et al., 2003). For example, "a pizza ontology might include the information that Mozzarella and Parmesan are variants of cheese, that cheese is not a kind of meat or seafood, and that a vegetarian pizza is one whose toppings do not include any meat or seafood. This information allows the term *pizza topped with only Mozzarella and Parmesan* to be unambiguously defined as a specialization of the term *vegetarian pizza*" (Horrocks et al., 2003). As a result, ontologies introduce a sharable and reusable knowledge base, enabling the extension of knowledge of a given domain.

The formal representation of the information in an ontology allows for the extension of existing knowledge via inferences made on the existing knowledge base. The reasoning capabilities associated with technologies also allow for the identification of data inconsistencies in the knowledge base. Ontologies are used in various domains as a type of information depiction about the world or some subset of it (Man, 2013). Examples of domains include Artificial Intelligence, Cybersecurity, the Semantic Web, Biomedical Informatics and the legal domain. Ontologies in the legal domain support the organisation of legal documents and provide support for legal reasoning (Bartolini & Mathuri, 2015a). Several successful ontologies have developed in the health sector domain, for example, the Gene Ontology (Consortium, 2001) and the Protein ontology (Sidhu et al., 2005).

The Web Ontology Language (OWL) is a "Semantic Web language intended to represent and capture rich and complex knowledge about things, and their relations between them" (Parsia et al., 2012). The knowledge expressed by OWL enables it to be reasoned about by using automated reasoners that verify the consistency of the domain knowledge within an ontology or revealing hidden knowledge. OWL ontologies promote reuse and modularity, as it can be published in the WWW and be referenced from other OWL ontologies. (Parsia et al., 2012). Knowledge concepts captured from data, in a given domain, are reasoned about in a rich hierarchical structure of concepts and their inter-relationships (Sidhu et al., 2004). These relationships help with the matching of concepts even if the data sources describing these concepts are not 100% uniform. Modelling knowledge in OWL has two focal points. "As a descriptive language, it can be used to formalize domain knowledge, and as a logical language, it can be used to make inferences from this knowledge" (Krötzsch, 2012). OWL 2 is a World Wide Web Consortium (W3C) standard. SPARQL, standardised by W3C (SeaBorne et al., 2008), is the standard query language for ontologies.

3. Related Work

The POPIA is very closely aligned with the GDPR, the latter being more expressive. One of the differences is that the POPIA applies the term "data subject" to natural and juristic persons, whereas the GDPR only applies this to natural persons (Da Veiga et al., 2018). There are various approaches related to expressing jurisdictional regulations such as the GDPR as an ontology (Pandit et al., 2018, Miles et al, 2007). The development of a POPIA-based ontology is inspired by these ontologies; no such ontology could be found in the literature at the time of writing. The objective of this paper is the development of an ontology to demonstrate concepts expressed within the POPIA and how it will influence the data subjects and responsible parties upon enactment. The goal is to create a prototype knowledge base for the POPIA concerning the lawful processing of personal information. The source for the descriptions of the POPIA concepts is derived from the Act's documentation (South African Government Gazette, 2013).

Pandit and Lewis (2017) developed an ontology, GDPR text extensions (GDPRtEXT), that provides an approach to allude to the ideas and terms conveyed inside the GDPR. It was developed using the "Simple Knowledge Organization System" (SKOS) (Miles et al., 2007) as a source for the descriptions of the GDPR concepts, and the developers described it as "… a Semantic Web language for representing formally structured vocabularies". The terms are linked to the appropriate focus points in the GDPR text using a URI pattern which links each term to a distinct resource within the GDPR. The GDPRtEXT ontology does not make use of inference to provide a better understanding of compliance obligations.

Bartolini et al. (2015b) developed an ontology for the GDPR for the data protection requirements. It shows the data protection prerequisites with regards to the GDPR changes and introduces a methodology for incorporating it into a workflow process to express these necessities inside a business procedure through the ontology. The goal of the ontology is to provide support for data controllers in accomplishing consistency with the GDPR enactment. This was done to create an ontological representation of the obligations of data controllers, and the comparing privileges of information subjects.

The GDPRov project (Pandit & Lewis, 2017) is an ontology concerned with the management of compliance through recognising provenance information identified with assent and individual personal information required for consistency documentation. It is an OWL 2 linked open data ontology that represents the provenance of assent and data lifecycle work processes for the GDPR. It outlines the provenance of exercises, for example, "data securing, usage, storage, deletion, and sharing of consent and the life cycles of data". The ontology uses SPARQL, an RDF query language, to query the provenance information described to find information relevant for compliance.

The PrOnto ontology (Palmirani et al., 2018) is a privacy ontology that captures the main concepts in the GDPR. These include "data types and documents, agents and roles, processing purposes, legal bases, processing operations, and deontic operations for modelling rights and duties". The objective of PrOnto is to assist with legal thinking and verification of compliance. It achieves this by applying defeasible logic reasoning as opposed to solely information retrieval. (Palmirani et al., 2018). Defeasible reasoning is a rule-based method for efficient reasoning with inconsistent data (Kontopoulos et al., 2013).

4. Requirements

4.1 Functional Requirements

The requirements for the development of a POPIA ontology are defined by a set of competency questions the ontology should be able to answer. These questions will serve as the litmus test in the evaluation phase of the future mature ontology and helped to define the scope of the ontology (Noy & McGuinness, 2001). A few of these questions are:

- What is the responsibility of the various role-players?
- What is considered to be personal information?
- How does a role player who collaborates with an international organization deal with the management of compliance with the Act?

The set of competency questions is likely to change as the POPIA ontology matures and is being evaluated and maintained.

4.2 Non-functional Requirements

Protégé (Musen, 2015), an open-source ontology editor created at the Stanford University Center for Biomedical Informatics Research, was used to develop the ontology. One of the main strengths of Protégé is its user interface and the flexible manner in which it can be extended to provide additional functionality in the form of plug-ins. One of the design goals of Protégé is to be compatible and adaptable with other systems for knowledge representation and knowledge extraction. (Lambrix et al., 2003). This compatibility enables possible integration with other knowledge bases within the same domain.

Since the ontology is based on an Act, it will always be susceptible to change, due to changes or amendments in regulations. An initial list of non-functional requirements that will need to be satisfied is Adaptability, Reusability, Configurability, Testability, Maintainability, and Quality.

5. Methodology

There are various methodologies to develop ontologies including the Cyc methodology, the methodology of Uschold and King, the methodology of Gruninger and Fox, METHONTOLOGY, the KACTUS approach, and the SENSUS-based methodology (Fernández-López et al., 2002). METHONTOLOGY enables the development of ontologies at the knowledge level and follows an iterative approach utilising evolving prototypes. It supports prototyping and comprises of the following processes: Specification, Knowledge Acquisition, Conceptualisation, Formalisation, Implementation, and Maintenance. Noy & McGuinness' Ontology Development Guide (2001) presents an iterative development process that repeats continuously to enhance the ontology and consists of the following sub-processes:

- Determine the domain and scope by defining a set of competency questions.
- Explore the reuse of existing ontologies.
- Listing key terms in the ontology.
- Create the classes and class hierarchy.
- Create the properties of classes.
- Create features for the defined properties.
- Create instances.

The authors used a combination of METHONTOLOGY and the Noy & McGuinness methods. The two methods have complementary processes. METHONTOLOGY provides high-level activities for the development life cycle and the *Development Guide* method provides granular steps for design and implementation which are intricate phases in the development life cycle.

METHONDOLOGY was implemented in the higher-level development processes such as the specification and knowledge acquisition processes. During the specification phases the purpose of the ontology, intended users and scope were determined. The output of the specialisation phase is a natural language description. The knowledge acquisition phase consisted of finding sources of information such as the results on GDPR ontology development, and occurs concurrently with the first phase. The other phases in METHONTOLOGY overlap to a great extend with those of the Noy and McGuinness methodology but the latter is more specific. The next section explains how the Noy & McGuinness methodology was followed to design the POPIA ontology.

6. Design, Implementation & Evaluation

The design follows the sub-processes of the *Ontology Guide* as specified in section 5. All the key terms considered important for the knowledge base are listed, an initial class hierarchy was defined, and properties for classes and their features were identified.

6.1 6.1. Important Terms

The identification of terms from the Act will assist in explaining concepts in the POPIA. The terms were chosen based on the competency questions the ontology should answer. To enable reasoning about the terms, the properties of the terms are also taken into consideration. The glossary of terms includes, but not limited to, *data subject*, *processing*, *accountability*, *right to access*, *operator*, *person*, *responsible party* and *personal information*.

For determining the classes, a list of key terms from the glossary was identified that describe the concepts having independent existence (Noy & McGuinness, 2001). The list of terms was then grouped according to the main concepts they represent within the Act. These groups form the basis of the ontology architecture. It is made up of the following main concepts:

- The conditions for lawful processing of personal information
- The rights of data subjects
- Data processing
- Person
- Action

These concepts are used to conceptualise and establish the class definitions of the ontology.

6.2 Classes

The Conditions for lawful processing of personal information is defined in Chapter 3 of the Act. It details eight conditions, with sub-conditions, that are to be complied with by the *Responsible Party*. “The responsible party must ensure that the conditions set out in this Chapter, and all the measures that give effect to such conditions, are complied with at the time of the determination of the purpose and means of the processing and during the processing itself " (South African Government Gazette, 2013).

The list of conditions is “Accountability, Processing limitation, Purpose specification, Further processing limitation, Information quality, Openness, Security Safeguards and Data subject participation". A subset of these concepts was used to define the classes for the lawful processing of personal information. From this list, the following subclasses of the *LawfullCondition* root class were created: *Accountability*, *InformationQuality*, *Lawfulness*, *Openness*, *PurposeSpecification*, and *Security*.

The rights of data subjects is defined Chapter 2 Section 5 of the Act. “A data subject has the right to have his, her or its personal information processed following the conditions for the lawful processing of personal information as referred to in Chapter 3". The concepts used for describing these rights as classes are derived from the nine rights described. These rights include, but not limited to, “the right to be notified when personal information is collected, the right to request correction of personal information and the right to object to the processing of personal information". (South African Government Gazette, 2013). From these rights, a *DataSubjectRight* class was created to with subclasses representing the individual rights: *RightToCorrection*, *RightToDeletion*, *RightToObject* and *RightToSubmitComplaint*.

Data processing (Chapter 1) provides the definitions and purpose of the Act. Personal information processing is defined as follows. “Processing means any operation or activity or any set of operations, whether or not by automatic means, concerning personal information" (South African Government Gazette, 2013). This includes amongst others the collection, storage, transfer or destruction of information. In addition, Chapter 8 Section 71 provides additional information on personal information processing. “A data subject may not be subject to a decision which results in legal consequences for him, her or it, or which affects him, her or it to a substantial degree, which is based solely on the basis of the automated processing of personal information intended to provide a profile of such person including his or her performance at work, or his, her or its credit worthiness, reliability, location, health, personal preferences or conduct". This section highlights automated processing and profiling which are used to further describe data processing.

These concepts are used to extend the class definitions for conceptualising data processing: A DataProcessing class with subclasses *ProcessingActivity*, *AutomatedProcessing*, *InformationTransfer*, *Profiling*, and *ManualProcessing.*

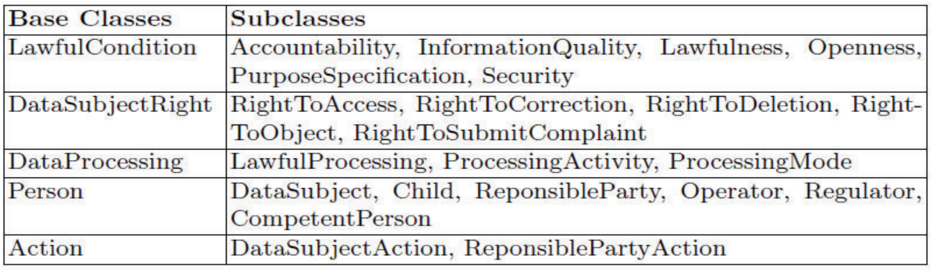
Person is described in Chapter 1 of the Act as "a natural person or a juristic person". It also describes the following concepts as a person. A child, who is "a natural person under the age of 18 years who is not legally competent". A competent person is any "person who is legally competent to consent to any action or decision being taken in respect of any matter concerning a child". A data subject refers to "the person to whom the personal information relates". An operator is "a person who processes personal information for a responsible party". The "responsible party is a public or private body or any other person which, determines the purpose of and means for processing personal information". Finally, the last concept identified for defining the "person" concept is the regulator. (South African Government Gazette, 2013). Chapter 5 Section 39 defines the regulator as a juristic person.

From these definitions, the following classes are created: *Person*, *Regulator*, *Child*, *DataSubject*, *CompetentPerson*, *Operator* and *ResponsibleParty*.

Action - There exists a link between the responsible party and a data subject due to the lawful conditions required for information processing and the rights that data subjects have. This link was used to describe two new concepts, a data subject action, and a responsible party action. A data subject has the right to exercise their rights such as the right to objection of personal information processing (Chapter 2, Section 5 of the Act). One of the conditions of lawful information processing is the responsibility of the responsible party to notify data subjects when their personal information is accessed and collected. These two concepts form the basis for concepts described as an Action, since it is an action that has to be performed by the responsible party, and in the case of the data subject, can be performed.

Table 1: describes the core classes.

Table 1. Core class Hierarchy



Base Classes	Subclasses
LawfulCondition	Accountability, InformationQuality, Lawfulness, Openness, PurposeSpecification, Security
DataSubjectRight	RightToAccess, RightToCorrection, RightToDeletion, Right-ToObject, RightToSubmitComplaint
DataProcessing	LawfulProcessing, ProcessingActivity, ProcessingMode
Person	DataSubject, Child, ReponsibleParty, Operator, Regulator, CompetentPerson
Action	DataSubjectAction, ReponsiblePartyAction

6.3 Properties

The properties are created by identifying the terms that provide the correlation between the classes defined and then expressing these correlations as relations within the ontology.

A subset of terms that was used to form the core properties of the ontology is listed below:

- A data subject can **exercise** his/her **rights**.
- A data subject **has personal information**.
- Data processing is **performed by** an operator.
- The responsible party to **ensure** conditions for lawful processing.

The highlighted terms are used to create properties. For example, "exercise rights" relates a data subject to an action they can perform, which in turn is related to a data subject right. An operator performs data processing for a responsible party, thus creating a relationship between the *Operator*, *DataProcessing* and *ResponsibleParty* classes.

Table 2: describes the core properties

Table 2. Core properties

Property	Domain	Range
exerciseRight	DataSubject	DataSubjectRight
hasInformation	DataSubject	PersonalInformation
mustEnsure	ResponsibleParty	LawfulCondition
performProcessing	Operator	DataProcessing

6.4 Concepts and Relations from related work

There is some overlap between the regulations of the GDPR and the POPIA. Thus some concepts from the ontology developed by Bartolini et al. (2015b) were used, and in some instances changed, to fit into the POPIA ontology. The GDPR ontology has a concept "Action" which represents the actions of data subjects. This concept was used in the POPIA ontology as a root class and represented the actions of data subjects and the responsible party as a *DataSubjectAction* and *ResponsiblePartyAction* class. The following properties were used: *accessData*, *exerciseRight, objectTo*, *performProcessing* and *processingPerformedBy* (Bartolini et al., 2015b).

7. Evaluation

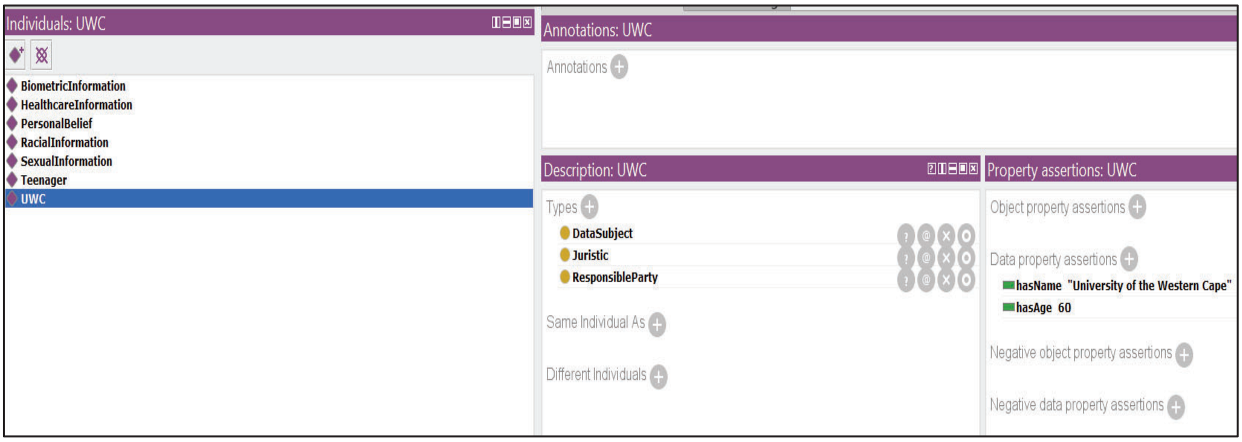
The ontology was evaluated by performing a set of queries using the DL (Description Logic) Query plugin in the Protégé editor. The evaluation consisted of answering an initial set of competency questions which includes querying the rights of data subjects, describing personal information, special personal information and the responsibility of the responsible party.

To provide an overview of the ontology model, in terms of numbers, the following statistics is provided:

Classes	:	60
Object properties	:	13
Data properties	:	3
Individuals	:	7
Nodes	:	65
Edges	:	119

The high number of edges provides an idea of how interconnected the classes are. Higher numbers, generally allows for further inferencing. This allows more general queries to return possible individuals without having to define sub-relations explicitly.

Figures 1 to 5 show two queries. Both the individuals, *UWC* and *Teenager* are members of the *Person* class; *UWC* is a member of the subclasses *ResponsibleParty* and *DataSubject* while *Teenager* is in the subclass *DataSubject* and the sub-subclass *ChildDataSubject*. Figures 1 and 2 show the representation of these individuals in the ontology respectively.



Individuals: UWC

BiometricInformation
HealthcareInformation
PersonalBelief
RacialInformation
SexualInformation
Teenager
UWC

Annotations: UWC

Annotations

Description: UWC

Types
DataSubject
Juristic
ResponsibleParty

Same Individual As

Different Individuals

Property assertions: UWC

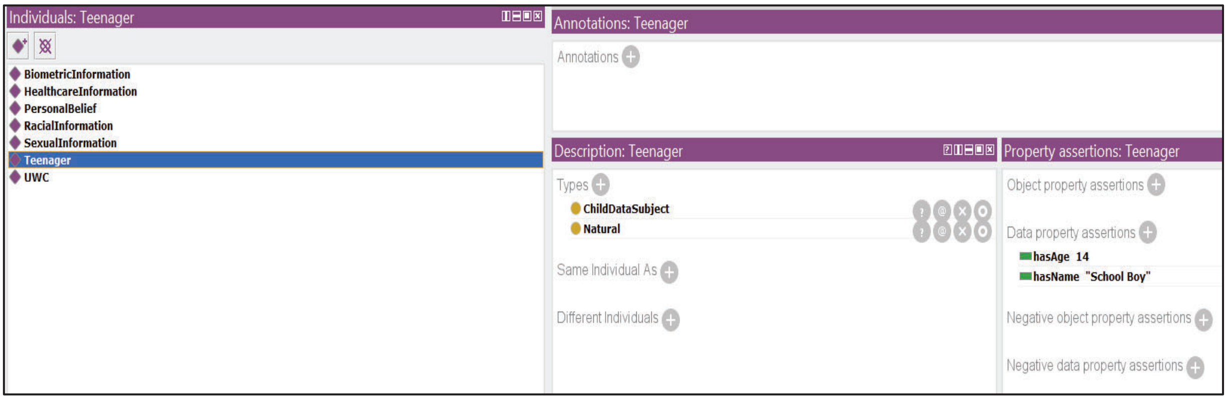
Object property assertions

Data property assertions
hasName "University of the Western Cape"
hasAge 60

Negative object property assertions

Negative data property assertions

Figure 1: Description of the Individual "UWC"



Individuals: Teenager

- BiometricInformation
- HealthcareInformation
- PersonalBelief
- RacialInformation
- SexualInformation
- Teenager
- UWC

Annotations: Teenager

Annotations

Description: Teenager

Types

- ChildDataSubject
- Natural

Same Individual As

Different Individuals

Property assertions: Teenager

Object property assertions

Data property assertions

- hasAge 14
- hasName "School Boy"

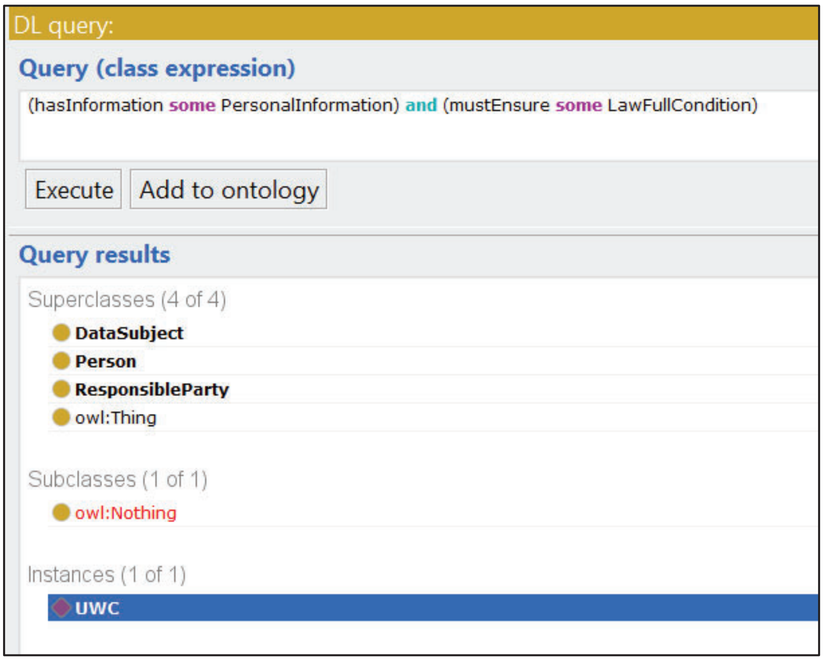
Negative object property assertions

Negative data property assertions

Figure 2: Description of the Individual "Teenager"

The first query is "List any entities who have to ensure that some condition required by the Act is satisfied AND has some PersonalInformation". The query and the reply of the reasoner (i.e. *UWC*) is shown in Figure 3. Figure 4 shows the explanation given by the reasoner for the answer of the first query.

Then we change the query to "List any entities who have to ensure that some condition required by the Act is satisfied OR has some PersonalInformation". The answer (i.e. *UWC* and *Teenager*) is shown in Figure 5.



DL query:

Query (class expression)

(hasInformation some PersonalInformation) and (mustEnsure some LawFullCondition)

Query results

Superclasses (4 of 4)

- DataSubject
- Person
- ResponsibleParty
- owl:Thing

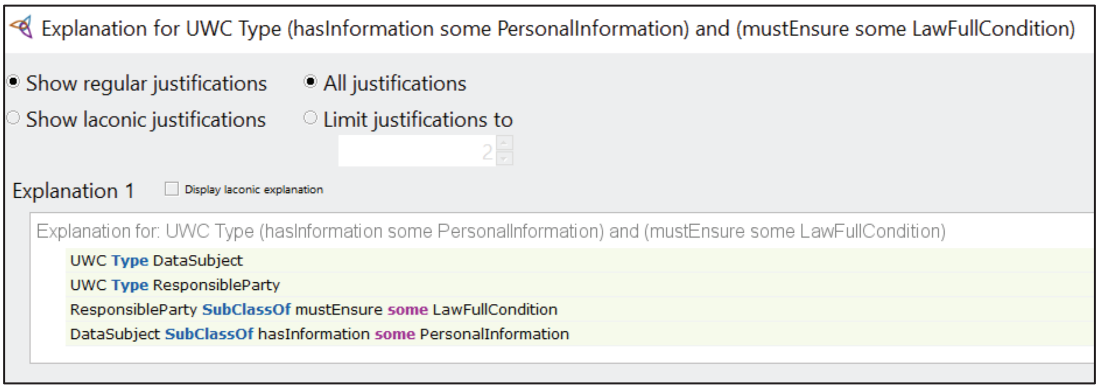
Subclasses (1 of 1)

- owl:Nothing

Instances (1 of 1)

- UWC

Figure 3: The first query and answer



Explanation for UWC Type (hasInformation some PersonalInformation) and (mustEnsure some LawFullCondition)

● Show regular justifications ● All justifications

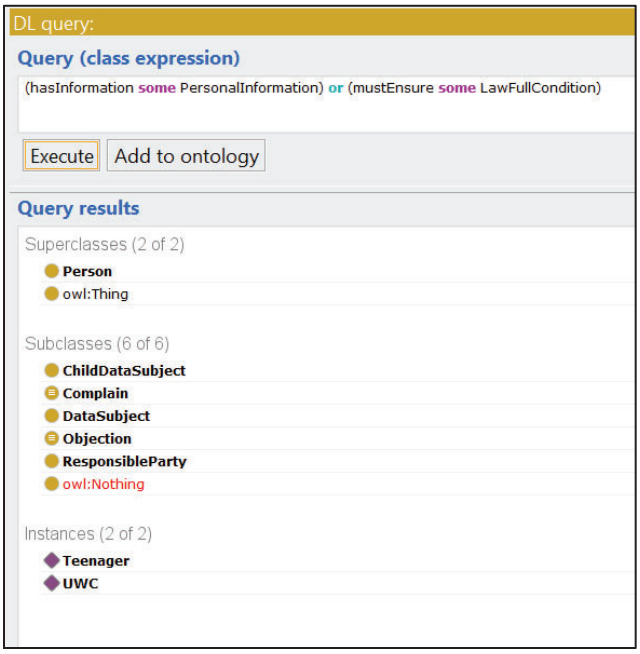
○ Show laconic justifications ○ Limit justifications to 2

Explanation 1 □ Display laconic explanation

Explanation for: UWC Type (hasInformation some PersonalInformation) and (mustEnsure some LawFullCondition)

UWC Type DataSubject
UWC Type ResponsibleParty
ResponsibleParty SubClassOf mustEnsure some LawFullCondition
DataSubject SubClassOf hasInformation some PersonalInformation

Figure 4: Explanation of the answer to the first query



(hasInformation some PersonalInformation) or (mustEnsure some LawFullCondition)

Superclasses (2 of 2)
- Person
- owl:Thing

Subclasses (6 of 6)
- ChildDataSubject
- Complain
- DataSubject
- Objection
- ResponsibleParty
- owl:Nothing

Instances (2 of 2)
- Teenager
- UWC

Figure 5: The second query and answer

8. Conclusion and Future Work

This paper highlights the importance of private information regulation in a globally connected world and the shift in the focus of legislators to enact legislation to support the privacy rights of individuals and juristic entities. The focus of this paper is the South African POPI Act and the impact it will have on various role players, such as the responsible party and data subjects. It outlines the development of a prototype ontology to provide a knowledge base on various concepts within the Act that will promote transparency and education that can aid with the inception of this Act. The development of the ontology must be continued to produce a mature ontology. The intended users of this ontology (when it has matured) include any person or organisation that will be affected by the POPIA, for example, citizens, businesses, government departments, and community centres.

The POPIA ontology is based on similar work done for the European Union’s General Data Protection Regulation (GDPR). This research demonstrates a proof of concept knowledge base on the POPIA. Going forward, this research can benefit by involving a legal domain expert in the evaluation and subsequent design decisions of the ontology. This will enable better accuracy in modelling decisions with regards to describing concepts formally. The reuse of existing relevant ontologies is a consideration for future work.

References

Bartolini, C. & Mathuri, R. (2015a). Reconciling data protection rights and obligations : An ontology of the forthcoming EU regulation (2015). Workshop on Language and Semantic Technology for Legal Domain (LST4LD), Recent Advances in Natural Language Processing (RANLP 2015).

Bartolini, C., Muthuri, R., Santos, C. (2015b). Using ontologies to model data protection requirements in workflows. Ninth International Workshop on Juris-informatics (JURISIN)

Benjamins, V.R., Contreras, J., Corcho, O., Gomez-Perez, A. (2002). Six challenges for the semantic web. International Conference on Principles of Knowledge Representation and Reasoning.

Birnhack, M.D. (2008). The EU Data Protection Directive: An Engine of Global Regime. *Computer Law & Security Review*, Vol. 26, no. 6. pp.508-520.

Botha, J., Grobler, M.M., Hahn, J., Eloff, M.M. (2017). A High-level Comparison between the South African Protection of Personal Information Act and International Data Protection Laws, 12th International Conference on Cyber Warfare and Security (ICCWS), Dayton, Ohio, USA.

Common, M.F. (2018). Facebook and Cambridge Analytica: let this be the high-water mark for impunity (07 2018), http://eprints.lse.ac.uk/88964/

Consortium, T.G.O. (2001). Creating the gene ontology resource: Design and implementation. *Genome Research* Vol. 11, no. 8, pp. 1425-1433. Retrieved from https://doi.org/10.1101/gr.180801, http://dx.doi.org/10.1101/gr.180801

Constitution of the Republic of South Africa: Section 14, (1996). http://www.justice.gov.za/legislation/constitution/SAConstitution-web-eng.pdf

Da Veiga, A., Vorster, R., Li, F., Clarke, N., Furnell, S. (2018). A comparison of compliance with data privacy requirements in two countries. 26th European Conference on Information Systems. University of Portsmouth.

Fernández-López, M., Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *Knowledge Engineering Review*. Vol 17, no. 2, pp. 129-156 (Jun 2002). Retrieved from https://doi.org/10.1017/S0269888902000462,

Horrocks, I., Patel-Schneider, P., Harmelen, F. (2003). From SHIQ and RDF to OWL: The making of a web ontology language. *SSRN Electronic Journal*, Vol 1. Retrieved from https://doi.org/10.2139/ssrn.3199003

https://doi.org/10.1017/S0269888902000462

Information Shield. (nd). International privacy laws. http://www.informationshield.com/intprivacylaws.html [Accessed July/7, 2014]

Kontopoulos, E., Bassiliades, N., Governatori, G., Antoniou, G. (2013). A modal defeasible reasoner of deontic logic for the semantic web. *Semantic Web*, pp. 140-167. https://doi.org/10.4018/978-1-4666-3610-1.ch007

Krötzsch, M. (2012). OWL 2 Profiles: An Introduction to Lightweight Ontology Languages, pp. 112-183. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). Retrieved from https://doi.org/10.1007/978-3-642-33158-94;

Lambrix, P., Habbouche, M., Perez, M. (2003). Evaluation of ontology development tools for bioinformatics. *Bioinformatic,s* Vol. 1, no. 12 , pp 1564-71 (2003)

Man, D. (2013). Ontologies in computer science. *Didactica Mathematica*, Vol. 31, pp. 43-46.

Miles, A., Pérez-Agüera, J.R.: Skos (2007). Simple knowledge organisation for the web*. Cataloging & Classification Quarterly,* Vol. 43(3-4), pp. 69-83. https://doi.org/10.1300/J104v43n03 04, https://doi.org/10.1300/J104v43n03$_0$4

Musen, M.A. (2015). The Protégé project: a look back and a look forward. *AI Matters* Vol. 1, no. 4, pp. 4-12. https://doi.org/10.1145/2757001.2757003

Noy, F., Mcguinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory 32*, Vol 01.

Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.(2018). PrOnto: Privacy Ontology for Legal Reasoning: 7th International Conference, EGOVIS 2018, Regensburg, Germany, pp. 139-152. https://doi.org/10.1007/978-3-319-98349-3$_1$1

Pandit, H.J., Fatema, K., O'Sullivan, D., Lewis, D. (2018). GDPRtEXT - GDPR as a Linked Data Resource. Lecture Notes in Computer Science. In book: The Semantic Web. DOI: 10.1007/978-3-319-93417-4_31

Pandit, H.J., Lewis, D. (2017). Modelling provenance for GDPR compliance using linked open data vocabularies. In: PrivOn@ISWC.

Parsia, B., Patel-Schneider, P., Krötzsch, M., Rudolph, S., Hitzler, P. (2012). OWL 2 web ontology language primer (second edition). Tech. rep., W3C (Dec 2012), http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

Seaborne, A., Prud'hommeaux, E. (2008). SPARQL query language for RDF. W3C recommendation, W3C http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

Sheth, A., Ramakrishnan, C. (2004). Semantic (web) technology in action: Ontology driven information systems for search, integration and analysis. *IEEE Data Engineering Bulletin*, Vol. 26.

Sidhu, A., S. Dillon, T., Chang, E., S. Sidhu, B. (2005) Protein ontology development using OWL. OWLED*05 Workshop on OWL: Experiences and Directions, Galway, Ireland.

South African Government Gazette (2108). https://opengazettes.org.za/gazettes/ZA/2018.html

South African Government Gazette: Protection of personal information act (2013), Retrieved from http://www.justice.gov.za/legislation/acts/2013-004.pdf

Srncova, Z., Babela, R., Mamrilla, R., & Balazova, Z. (2019). GDPR implementation in public health. Int Health Journal. Feb 2019 Issue. http://ihjonline.com/issue/2019-02/gdpr-implementation-in-public-health

United Kingdom Government Gazette (1998). Data Protection Act. http://www.legislation.gov.uk/ukpga/1998/29/contents

Usefulness of Cyber Attribution Indicators

Henrik Karlzén
Swedish Defence Research Agency, Linköping, Sweden
henrik.karlzen@foi.se

DOI: 10.34190/EWS.20.074

Abstract: Attributing a cyberattack to the attacker is difficult. Cyberspace is conducive to anonymity and many attackers actively hide their tracks by using false flags that point to other culprits. On the other hand, the difficulties of attribution are overstated by some, such as the media, in efforts to remain neutral. Furthermore, the requisite level of certainty in attributions depends on whether the information will be acted upon in a court of law by striking back, or merely to better understand the attack. Attributing cyberattacks would enable counter-attacking and the use of sanctions. If the attribution was to establish a nation-state as the attacker, there could be geopolitical and military implications. Attacks by nation-states may also exempt insurance companies from paying out compensation and reveal cyber weapon sales in violation of export restrictions to those nation-states. The extant academic literature on attribution is interdisciplinary, with research on technical indicators and discussions on political implications. For instance, the Diamond model includes attacker motivations and capabilities, while the Q model includes over a hundred indicators divided into more than thirty categories, such as functionality and skills. Based on the literature, this paper investigates how attribution has been used by cyber security companies to establish attacker identity in all attacks during 2017, as covered by the Council on Foreign Relations database on nation-state cyber operations. It is concluded (1) that attribution is based on indicators such as IP addresses, language settings, required resources, and the apparent (political) goal. It is also concluded (2) that these indicators may be hard to measure, e.g. due to obfuscation and false flags that point in different directions. These are often based on previous (also weak) attribution and are commonly used without scientific rigour. Finally, it is also concluded (3) that publicly pointing fingers at nation-states may be pointless or dangerous because of the power differential, and may reveal one's sources and methods for attribution.

Keywords: cyberattack, attribution, cyber operation, APT, security, usefulness

1. Introduction

This paper investigates attribution, i.e. determining who conducted a cyberattack. Attackers typically wish to remain anonymous when conducting cyberattacks while there are several reasons why others want to reveal a cyber-attacker's identity. First, the victim (and others) probably wants to change their relationship to the attacker, e.g. by using sanctions (Davis et al., 2017), or by counter-attacking (Egloff, 2018). This is arguably more important when the attacker acts out of character (e.g. attacking despite a peace treaty (Porter, 2017)). These possibilities also mean that attribution can be a deterrent (Floyd, 2018). Second, attacker identity might determine who should deal with the consequences of the attack. For instance, a nation-state attack may absolve an insurance company from paying damages to a victim. A case is currently being litigated in a US court, where the victim is claiming that the insurance company has denied payment without having proved that the attack was indeed conducted by a nation-state (Illinois, 2018). Third, a manufacturer of a cyber-weapon might only sell the weapon to certain organisations, and if others use it the manufacturer may want to take action (Ahmed and Perlroth, 2019). Fourth, knowledge of the attacker enriches one's understanding, enabling better defence (Hunker, Hutchinson, and Margulies, 2008). Fifth, cyber security companies can use attribution as a proof-of-concept or for PR (Davis et al., 2017).

Attribution is often conducted by multinational cyber security companies that provide good insight into networks (Romanosky and Boudreaux, 2019). They also have strong technical capabilities, but somewhat lack the analytical capability of intelligence agencies (Bartholomew and Guerrero-Saade, 2016). On the other hand, intelligence agencies – that have good sources both in cyber and other domains (Malik, 2017; Symantec, 2018) – cannot speak as openly about attribution, e.g. to protect methods and sources. Collaboration between organisations is limited, which is demonstrated by their confusing assignment of different names for the same threat actor (Romanosky and Boudreaux, 2019). It has been suggested that different attributing organisations work together, e.g. in a consortium of private actors (Davis et al., 2017), a cyber-IAEA (promoting peaceful cyber akin to the work of the International Atomic Energy Agency) (Neutze, 2016), or in an international court for cyber attribution (Chernenko, Demidov, and Lukyanov, 2018).

This paper aims to better understand how attacks are attributed, as well as the obstacles to attribution and its publicising. More specifically the paper answers the following research question: *how are attacks attributed in theory and practice?*

1.1 Related work

Attribution is an interdisciplinary task, with research focusing on technical indicators as well as political implications and communication. For instance, Rid and Buchanan (2015) studied the less technical side of how to attribute and proposed the *Q model* for categorising attribution indicators. In contrast, Kaspersky researchers Bartholomew and Guerrero-Saade (2016) described more technical challenges as well as false flags used by attackers to mislead. Davis et al. (2017), published by the US think tank RAND, described that attribution methodologies vary and that a standardised and scientifically transparent methodology is missing. Another RAND paper, by Romanosky and Boudreaux (2019), studied the topic of publicly disclosing attribution conclusions. The Master thesis by Boot (2019) summarised one type of attribution indicators and how attackers conceal themselves. In the paper by Pahi and Skopik (2019), presented at the 2019 European Conference on Cyber Warfare and Security, the so-called *Diamond model* for attribution was elaborated in order to address the shortcomings of attribution methods and increase the resistance to false flags. The new *Cyber attribution model* structurally separated and then combined the *how* of an attack with the *who* (attacker).

Most of the academic literature is theoretical in nature and few papers investigate larger numbers of empirical cases where attribution has occurred. As such, theories are not validated by empirical data, and the academic literature typically only investigates if a proposed model can be exemplified by a few cases rather than large collections of data. The paper by Valeriano and Maness (2014) is a rare exception; it formulated a theory on cyber conflict and evaluated the theory by collecting data on 110 cyber incidents and 45 cyber disputes between rival nations. The paper indicated that cyber combat is rarely used, causing mainly minor effects, and is primarily regional rather than global.

There is currently much interest in improving the possibilities of attribution, such as the new US Defense Advanced Research Projects Agency (DARPA) program Enhanced Attribution, running from 1 November 2016 to 1 May 2021. The program aims to decrease the time for attribution from months to hours or even seconds (DARPA, 2016a), and increase the government's capability to openly reveal its conclusions about an attack without harming sources and methods (DARPA, 2016b).

1.2 Paper structure

The remainder of the present paper describes the research method, attribution indicators in theory and practice, discussions of the practical difficulties with attribution as well as limitations of the empirical data, and finally conclusions.

2. Method

The next subsection details how cyberattack databases and cyberattacks were identified, followed by a subsection on how the empirical data was categorised.

2.1 Cyberattack databases

In order to find information on conducted cyberattacks, an internet search for databases on such attacks was performed. Many databases with information on cyberattacks were found. To choose between the databases, several criteria were applied. Similar criteria had previously been established by this paper's author together with some colleagues. These criteria were (somewhat adjusted) as follows, with excluded databases based on each criterion: They focus on *attacks* rather than actors, etc. (7 excluded: e.g. ATT&CK Groups, and APT Groups and Operations, which focused on actors; and 1 excluded: VirusTotal, which focused on malware.) They have *references* with more details for increased trustworthiness. (1 excluded: CSIS, which lacked references; and 1 excluded: Dyadic, which only had references in terms of names of news media.) They have *details* such as dates, modi operandi, attackers, and victims. (4 excluded: e.g. Security Without Borders.) They contain a large *number* of attacks rather than only a few. (1 excluded: Relevante Cybervorfälle.) Finally, they are *trustworthy* (with a described method and reliable authors). (2 excluded: Hackmageddon, which is authored by a private individual and has relatively few reliable sources, and VERIS which does not clearly describe its method for compilation.)

Only one database was not excluded: Cyber Operations Tracker (CFR), composed of 300+ operations (attacks) from the year 2005 onwards, authored by the US think tank Council on Foreign Relations. This is also the most complete database for state-sponsored cyberattacks (Romanosky and Boudreaux, 2019), and is updated quarterly (CFR, 2019). It may be noted that several of the excluded databases are (partially) based on CFR (and vice versa). It may also be noted that CFR also has its flaws, such as commonly weak attribution evidence (that the attacks were state-sponsored), and it only contains open sources that are mostly in the English language with victims who are native English speakers (CFR, 2019).

An example of how CFR describes an attack is shown in Table 1. For each attack, CFR provides a title, date, victim, attacker, description of attack, and possible victim response. There are also links to one to three sources (per attack), where more information can be found. Over half the sources are *news articles*, with the most common news media being The New York Times, The Washington Post, Reuters, and Wired. A third of the sources are *cyber security companies' reports*, with the most common one being Kaspersky, FireEye, Symantec and Palo Alto Networks. There are also *government reports* (seven percent of sources), mostly from the USA, and o*ther sources,* e.g. documents from think tanks and research institutes.

Table 1: Sample attack in CFR (shortened).

Title	**Date**	**Description**			
Compromise the North Korean nuclear program	2017-03-04	A threat actor, believed to be the United States, targeted computer networks and electronic components associated with the North Korean nuclear program to sabotage and slow its development.			
Response	**Victims**	**Sponsor**	**Type**	**Category**	**Sources_1**
[blank]	North Korea	United States	Sabotage	Military	nytimes.com/…

In order to obtain a manageable amount of sources, only the 45 attacks from 2017 (fresh but also mature) were chosen. The CFR sources were read by this paper's author together with a colleague, and statements on attribution were noted and divided into categories (as described in section 3).

2.2 Categorisations of attribution indicators

There are several categorisations of attribution indicators in the literature, such as the Diamond Model (Caltagirone, Pendergast, and Betz, 2013) that scientifically formalises the work of intrusion analysts. This model, where the diamond's four corners represent the adversary, capability, victim and infrastructure, was further developed by Pahi and Skopik (2019). A second categorisation is the Lockheed Martin Cyber Kill Chain (Hutchins, Cloppert, and Amin, 2010), which divides attacks into different stages. However, it was primarily developed to stop attacks. A third categorisation by Wheeler, Larsen, and Leader (2003) focuses on network communication indicators, with the goal to adjust systems, thus making attribution easier. A fourth categorisation by Cook et al. (2016) uses categories such as network communication, forensic analysis of client, analysis of malware, and non-technical methods, and is thus more useful for measurements. A fifth categorisation by ODNI (2018) describes the US intelligence agencies' categories for attribution, such as recurring modi operandi, infrastructure, malware, motivation, and indicators from external sources.

This paper uses a more detailed (semi-structured) categorisation developed by Rid and Buchanan (2015), called the Q model. The Q model includes, among other things, more than a hundred questions the answers to which may be considered attribution indicators. Most of the questions are divided into more specific categories provides the categories of indication used in the Q model, together with explanations and indicators mentioned by some other theoretical works on indicators. As illustrated inTable 2, the Q model fairly well covers the attribution indicators suggested elsewhere, although only if the categories are interpreted in a broad manner.

Table 2: Attribution indicator categories (first column) from the Q model by Rid and Buchanan (2015). The categories are briefly explained in the second column. The third column shows indicators from some other theoretical works

Category	Explanation	Indicators in the literature
Indicators of compromise	Suspicious behaviour that triggers the attribution effort	
Entry	Penetration techniques, vulnerabilities	
Targeting	The target and its specificity	Target specificity[a, b, e]
Infrastructure	Infrastructure such as software,	Infrastructure[f]

Category	Explanation	Indicators in the literature
	command and control (C2)	Bandwidth[d]
		Compiler[f]
		Domain[d]
		IP-address[a ,e ,f]
		Exfiltration method[f]
Modularity	Malware modularity, reused components, different teams	Tool procurement[d]
		Programming style[f, h]
		Code blocks, function calls, multi-threading, binary similarity, etc[f]
Language	System language settings, region-specific strings	System language settings[f],
		Region-specific strings[a]
		Programming language[f]
		Language skills[c]
Personas	Pseudonyms and names	Pseudonyms and names[c]
Pattern-of-Life	Time of attack, working hours	Time of attack, working hours[c]
Stealth	Detection evasion, anti-forensics	Anti-attribution method[f]
		The password chosen to hide code[f]
		Crypto-mechanism[f]
Cluster	Other attacks with similar methodologies	
Functionality	Designed purpose (modify, interrupt, etc.)	
Approval	Who approved it, target verification, collateral minimisation	Amount of collateral damage[a, b]
Mistakes	Attacker mistakes	
Skills	Skills required and their rarity	Technical sophistication[a]
Scope	E.g. as campaign	The attacker returns to the system[h]
Stages	E.g. preparatory, main, or follow-up; teams	Number of steps[e, h]
Evolution	Changes during attacking	
Claims	Claims, announcements	Claims[e]
Insider	Insider required and used	
Intelligence	Target intelligence required	Target intelligence required[g]
Cost	Cost and testing requirements	Testing requirements[g]
Significance	Significance for attacker	
Context	Political and regional context, other events, other sources	Political and regional context[a]
		Other events[a, b]
		Other sources[a]
		The judicial system makes no effort or has no success[a,b]
Benefit	Who benefited?	Matching doctrine[a]
Consequences	Damage, side-effects	Effect[b]
		Side-effects[a, b]

Sources: a = Jolley (2017), b = Ottis (2009), c = Bartholomew and Guerrero-Saade (2016), d = Cook et al. (2016), e = Mateski et al. (2012), f = Boot (2019), g = Langner (2013), h = Nicholson et al. (2015).

3. Attribution indicators in theory and practice

The CFR sources describe their attempts to attribute attacks in various detail. One example involved cyber security company FireEye attributing an attack to the Iranian government, based on a username in the code being the same as one used in a web forum with ties to the Iranian government. Furthermore, the code contained artefacts in Iranian language Farsi, the targets aligned with Iranian interests, and the time of attack aligned with the Iranian time zone and working week (Saturday-Wednesday). Similarities to code used in other attacks were also mentioned. News media usually provide fewer details and rely on claims from anonymous people or government officials who provide no evidence.

Table 3 matches the Q model categories with practical attribution examples from the CFR database sources. The Q model covers quite well the indicators used in practice, with only *insider requirements* missing in this selection of empirical data. It is also clear that many practical examples were needed to cover the entire Q model, indicating that it is rare to use the entire Q model to attribute a single attack. Indeed, only a few indicators are mentioned in each case, with the indicators varying between cases, and indicators that constitute counter-evidence are ignored or even considered false flags (as in PWC (2017)). Furthermore, the chains of evidence are long, with one attribution building on the previous one.

Table 3: Attribution indicator categories and practical examples.

Category	**Practical examples**
Indicators of compromise	Data inaccessible (Cherepanov, 2017)
Entry	Spearphishing (Great, 2017)
	Supply-chain attack (Cherepanov, 2017)
Targeting	Target fitting national interests (ThreatConnect, 2017a; O'Leary et al., 2017)
	Diverse targeting (Secureworks, 2017)
Infrastructure	Nationality of DNS servers (O'Leary et al., 2017) and IP-addresses (Reaqta, 2017; Marczak et al., 2017)
	Spoofed domains (ThreatConnect, 2017b)
	C2 (PWC, 2017; Symantec, 2017)
	Links via legitimate websites to trick filters (ThreatConnect, 2017a)
Modularity	Open source tools (Lancaster, 2017)
	Nationality of tools (O'Leary et al., 2017)
	Tool demonstration log files (Marczak et al., 2017)
	Code (Symantec, 2017)
Language	System language settings (Reaqta, 2017; Atch and Neray, 2017)
	Artefacts in specific natural language (O'Leary et al., 2017; Secureworks, 2017)
	Poor or unusual use of natural language (Great, 2017)
Personas	Recurring names in code (Symantec, 2017; O'Leary et al., 2017)
Pattern-of-Life	Reduced attack activity during certain holidays (Secureworks, 2017; O'Leary et al., 2017)
	Compilation on certain countries' working hours (PWC, 2017)
Stealth	Type of code obfuscation (Symantec, 2017)
	Type of crypto-mechanism (Hegel, 2018)
Cluster	Similar process-trees (Reaqta, 2017)
Functionality	Gather information (Goeij, 2017)
	Destroy data (Cherepanov, 2017)
Approval	Who ordered it (Reyes et al, 2017)
	Existence of kill switch (Symantec, 2017)
Mistakes	Mistakenly accessing without proxy (Hegel, 2018)
Skills	Capability (Atch and Neray, 2017)
Scope	Attacks to obtain code signing certificates for later attacks (Hegel, 2018)
Stages	Several teams cooperating (PWC, 2017; Hegel, 2018)
Evolution	First targeting diplomatic entities, then defence organisations (Great, 2017)
	Adapts and learns to evade detection (Hegel, 2018)
Claims	Former team member admission (Reyes et al., 2017)
Insider	[not used]
Intelligence	Target's email solutions (Hegel, 2018)
Cost	Massive infrastructure to decrypt and analyse (Atch and Neray, 2017)
Significance *(for attacker)*	Valuable (Secureworks, 2017)
Context	Victim a certain country showed interest in (O'Leary et al., 2017; PWC, 2017
	Statements from officials (Goeij, 2017; Huetteman, 2017; Homewood, 2017)
Benefit	A certain nation (Homewood, 2017)
Consequences	Serious for a nation's security (Goeij, 2017)
	Intelligence gathering (Atch and Neray, 2017)

4. Discussion

The following subsections discuss practical difficulties with attribution, and limitations of the empirical data.

4.1 Practical difficulties with attribution

Attribution is slow and often requires weeks or months of analysis (ODNI, 2018). Attributing cyberattacks is also difficult (Keromytis, 2016), although the news media often overstate the problems in order to appear balanced in their reporting (Bartholomew and Guerrero-Saade, 2016).

Finding an accurate and complete set of attribution indicators is a challenge. While all empirical cases studied in this paper are covered by the Q model, each case does not use even close to the full model. However, one single indicator is rarely going to be sufficient, since e.g. attacker tool use does not reveal much as tools are often openly available (Bartholomew and Guerrero-Saade, 2016).

Furthermore, some indicators are difficult to measure, e.g. code that is too difficult to understand, or jurisdictional issues restricting access to information (Hunker, Hutchinson, and Margulies, 2008; Romanosky and Boudreaux, 2019). Proficient attackers adapt to known indicators and obfuscate their behaviour, e.g. by restricting execution of their malware (Boot, 2019). Allegedly, more than half of Stuxnet development costs were for anti-attribution (Langner, 2013). The Vault7 documents released by Wikileaks (2017a) (and allegedly written by the CIA) show instructions not to leave dates and times in code.

Attackers also mislead attribution efforts by planting false flags, e.g. changing language settings (Pahi and Skopik, 2019) or adding text in different languages (Bartholomew and Guerrero-Saade, 2016). Wikileaks (2017b) suggests that the CIA could use Vault7 to write in a foreign language and make it seem like they tried to conceal that language. Wikileaks (2017c) further suggests that Vault7's "Umbrage" code snippets from others' cyber weapons could be used to blame the snippets' original authors. Another alleged use of false flags was how (outdated) certificates used in Stuxnet were later reused in different code to cast blame on Stuxnet (Bartholomew and Guerrero-Saade, 2016).

In some cases, misleading might even be the main goal in order to blame a third party (Wheeler, Larsen, and Leader, 2003). However, some attackers may wish for attribution in order to claim responsibility, revealing their capability (Floyd, 2018) and gaining status (c.f. terrorist claims in Kearns, 2019). Furthermore, Kaspersky researcher Guerrero-Saade thinks "false flags are rare and that most attribution that threat researchers do is accurate" (Zetter, 2017).

To facilitate attribution, internet protocols may be made more resistant to spoofing, and traffic that is difficult to attribute could be filtered by routers (Nicholson et al., 2015). However, this would be difficult to accomplish (Cook et al., 2016), and privacy implications stopped one such effort by DARPA (Wheeler, Larsen, and Leader, 2003). Another option is to improve the attribution process by comparing different hypotheses (ODNI, 2018), as in Fokker and Beek (2019), using independent reviews (Davis et al., 2017), or public disagreements between analysts as in e.g. Lancaster (2017).

4.2 Limitations of the empirical data

As the CFR database is based on only open sources, attacks that are not publicly attributed might be missed. There are situations when public attribution is not desirable, e.g. warning the attacker and thereby revealing one's sources and methods (Rid and Buchanan, 2015; Bartholomew and Guerrero-Saade, 2016). Attribution can also be harmful to one's reputation, or even dangerous (Romanosky and Boudreaux, 2019). Companies do not point fingers at nation-states as often as other nation-states do (Mueller et al., 2019), and US companies may hesitate to point fingers at the US government (Yadron, 2015; Romanosky and Boudreaux, 2019). Public attribution can also ruin attacks that one would like to happen, such as when being collateral damage of an anti-terrorist operation (Rid and Buchanan, 2015).

Another flaw with the CFR database and its public sources is that the attribution evidence is typically weak and likely to be insufficient for courts of law (Berghel, 2017). However, weaker evidence may lead to stronger evidence (Clark and Landau, 2010), or be sufficient for security decision-making (Tsagourias, 2012). There is rarely an assessment of evidence strength, with a few exceptions such as Hegel (2018), despite the importance of certainty assessments (Shakarian et al., 2015; Davis et al., 2017). Furthermore, while there are examples of attribution uncertainty scales, e.g. those in ODNI (2018), there is no established scale (Davis et al., 2017). Without an established scale, actors will disagree on what is sufficient evidence, e.g. depending on their ability to collect evidence (when attributing) (Schmitt and Vihul, 2017), and obfuscate (when attacking). Furthermore, it is not surprising that those who are accused consider the evidence too weak, such as the Chinese government's reaction to the US attributing attacks to China (Ministry of Foreign Affairs of PRC, 2015).

5. Conclusions

The focus of this paper was to investigate how cyberattacks are attributed in theory and practice. Theoretical works have suggested many different ways to attribute cyberattacks, e.g. the highly detailed Q model. However, the methods used in practice do not seem to use a set methodology. While the Q model fairly well covers all practical cases of attribution in the CFR database as investigated in this paper, only some of the multitude of attribution indicators are applied in each case, such as IP addresses, language settings, required resources, and the apparent (political) goal. Whether attacks required an insider was not used as an indicator in the studied practical cases.

Studying actual attribution cases has its hurdles. Gaining access to attribution data is difficult, and there are reasons to believe that the open sources do not reflect all attribution efforts (some of which are not made public). There are several reasons not to make public attribution statements. Public attribution warns attackers, tips off other potential victims (who one might want to see get hurt), harms the relationship with the alleged attacker, harms the reputation if the attribution is shown to be wrong, and reveals sources and methods. Furthermore, publicly pointing fingers at nation-states may be pointless or dangerous because of the power differential.

Attribution is difficult, requiring a hard process of measuring many different attribution indicators with scientific rigour and overcoming attackers' obfuscation and false flags as well as one's own subjectivity. Furthermore, the indicators may point in different directions – gathered attribution evidence is often weak and based on previous (also weak) attribution. Moreover, the strength of the evidence is rarely assessed in practice. There is also a lack of collaboration between organisations who seek to attribute.

References

Ahmed, B.A. and Perlroth, N. (2019) "Using Texts as Lures, Government Spyware Targets Mexican Journalists and Their Families", [online], *New York Times,* 19 June, www.nytimes.com/2017/06/19/world/americas/mexico-spyware-anticrime.html

Atch, D. and Neray, P. (2017) "Operation BugDrop: CyberX Discovers Large-Scale Cyber-Reconnaissance Operation Targeting Ukrainian Organizations", [online], CyberX Labs, cyberx-labs.com/en/blog/operation-bugdrop-cyberx-discovers-large-scale-cyber-reconnaissance-operation/

Bartholomew, B. and Guerrero-Saade, J.A. (2016) "Wave Your False Flags! Deception Tactics Muddying Attribution in Targeted Attacks", *Virus Bulletin*, October.

Berghel, H. (2017) "On the Problem of (Cyber) Attribution", *Computer*, Vol. 50, No. 3.

Boot, C. (2019) "Applying Supervised Learning on Malware Authorship Attribution", M.Sc. thesis, Radboud University Nijmegen.

Caltagirone, S., Pendergast, A. and Betz, C. (2013). *The Diamond Model of Intrusion Analysis*, Hanover, MD: Center for Cyber Threat Intelligence and Threat Research, 5 July.

Cherepanov, A. (2017) "TeleBots are back: Supply-chain attacks against Ukraine", [online], Eset, https://www.welivesecurity.com/2017/06/30/telebots-back-supply-chain-attacks-against-ukraine/

Chernenko, E., Demidov, O. and Lukyanov, F. (2018) "Increasing international cooperation and cybersecurity and adapting cyber norms", [online], Council on Foreign Relations, www.cfr.org/report/increasing-international-cooperation-cybersecurity-and-adapting-cyber-norms

Clark, D.D. and Landau, S. (2010) "The Problem isn't Attribution; It's Multi-Stage Attacks", *ACM ReArch*.

Cook, A., Nicholson, A., Janicke, H., Maglaras, L. and Smith, R. (2016) "Attribution of Cyber Attacks on Industrial Control Systems", *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, Vol. 3, No. 7.

DARPA. (2016a) "DARPA-BAA-16-34, Enhanced Attribution, Frequently Asked Questions", DARPA.

DARPA. (2016b) "Broad Agency Announcement Enhanced Attribution DARPA-BAA-16-34", DARPA.

Davis, J., Boudreaux, B., Welburn, J., Aguirre, J., Ogletree, C., McGovern, G. and Chase, M. (2017) "Stateless Attribution: Toward International Accountability in Cyberspace", RAND.

Egloff, F.J. (2018) "Cybersecurity and Non-State Actors", Ph.D. thesis, University of Oxford.

Floyd, G.S. (2018) "Attribution and Operational Art: Implications for Competing in Time", *Strategic Studies Quarterly*, Vol. 12, No. 2.

Fokker, J. and Beek, C. (2019) "Ryuk Ransomware Attack: Rush to Attribution Misses the Point", [online], *McAfee Blogs*, securingtomorrow.mcafee.com/other-blogs/mcafee-labs/ryuk-ransomware-attack-rush-to-attribution-misses-the-point/

Goeij, H.d. (2017) “Czech Government Suspects Foreign Power in Hacking of Its Email”, [online], *The New York Times*, 31 January, www.nytimes.com/2017/01/31/world/europe/czech-government-suspects-foreign-power-in-hacking-of-its-email.html?_r=1%0A

Great. (2017) “Introducing WhiteBear”, *Securelist,* [online], Kaspersky, securelist.com/introducing-whitebear/81638/%0A

Hegel, T. (2018) “Burning Umbrella: An Intelligence Report on the Winnti Umbrella and Associated State-Sponsored Attackers”, [online], 401Trg, 401trg.pw/burning-umbrella/

Homewood, B. (2017) “IAAF says medical records compromised by Fancy Bear hacking group”, [online], *Reuters*, www.reuters.com/article/us-sport-doping-iaaf/iaaf-says-medical-records-compromised-by-fancy-bear-hacking-group-idUSKBN1750ZM

Huetteman, E. (2017) “Marco Rubio Says His Campaign Was a Target of Russian Cyberattacks”, [online], *The New York Times,* 30 March, www.nytimes.com/2017/03/30/us/politics/marco-rubio-russian-cyberattacks.html

Hunker, J., Hutchinson, B. and Margulies, J. (2008) "Role and Challenges for Sufficient Cyber-Attack Attribution", Dartmouth College

Hutchins, E., Cloppert, M. and Amin, R. (2011) *Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains*, 6th International Conference on Information Warfare and Security, ICIW

Illinois. (2018) “2018 WL 4941760 (Ill.Cir.Ct.) (Trial Pleading). MONDELEZ INTERNATIONAL, INC., Plaintiff, v. ZURICH AMERICAN INSURANCE COMPANY, Defendant”, Circuit Court of Illinois.

Jolley, J.D. (2017) *Attribution, State Responsibility, and the Duty to Prevent Malicious Cyber-Attacks in International Law*, University of Glasgow

Kearns, E.M. (2019) “When to Take Credit for Terrorism? A Cross-National Examination of Claims and Attributions”, *Terrorism and Political Violence*

Keromytis, A. (2016) “Enhanced Attribution”, [online], DARPA, www.enisa.europa.eu/events/cti-eu-event/cti-eu-event-presentations/enhanced-attribution

Lancaster, T. (2017) “Muddying the Water: Targeted Attacks in the Middle East”, [online], Palo Alto Networks, researchcenter.paloaltonetworks.com/2017/11/unit42-muddying-the-water-targeted-attacks-in-the-middle-east

Langner, R. (2013) *To Kill a Centrifuge*, The Langner Group.

Malik, W. (2017) “What are the benefits of attribution?”, [online], *TrendMicro,* 16 August, blog.trendmicro.com/what-are-the-benefits-of-attribution/

Marczak, B. (2017) “Champing at the Cyberbit”, [online], The Citizen Lab, citizenlab.ca/2017/12/champing-cyberbit-ethiopian-dissidents-targeted-commercial-spyware

Mateski, M., Trevino, C.M., Veitch, C.K., Michalski, J., Harris, J.M., Maruoka, S. and Frye, J. (2012) “Cyber Threat Metrics”, *SANDIA Report,* SAND2012-2427, March.

Ministry of Foreign Affairs of PRC. (2015) “Foreign Ministry Spokesperson Hong Lei’s Regular Press Conference”, Ministry of Foreign Affairs, the People’s Republic of China, 5 June.

Mueller, M., Grindal, K., Kuerbis, B. and Badiei, F. (2019) ”Cyber Attribution”, *The Cyber Defense Review*, Vol. 4, No. 1.

Neutze, J. (2016) “The role of cybernorms in preventing digital warfare”, [online], *Microsoft EU Policy Blog*, blogs.microsoft.com/eupolicy/2016/07/08/the-role-of-cybernorms-in-preventing-digital-warfare/

Nicholson, A., Janicke, H., Watson, T. and Smith, R. (2015) “Rolling the Dice - Deceptive authentication for attack attribution”, *Proceedings of the 10th International Conference on Cyber Warfare and Security, ICCWS*.

O’Leary, J., Kimble, J., Vanderlee, K. and Fraser, N. (2017) “Insights into Iranian Cyber Espionage: APT33 Targets Aerospace and Energy Sectors and has Ties to Destructive Malware”, [online], FireEye. www.fireeye.com/blog/threat-research/2017/09/apt33-insights-into-iranian-cyber-espionage.html

ODNI. (2018) *A Guide to Cyber Attribution*, Office of the director of national intelligence

Ottis, R. (2009) *Theoretical Model for Creating a Nation-State Level Offensive Cyber Capability*, European Conference on Information Warfare and Security

Pahi, T. and Skopik, F. (2019) *Cyber Attribution 2.0 : Capture the False Flag*, European Conference on Cyber Warfare and Security.

Porter, C. (2017) “Private Sector Cyber Intelligence Could Be Key to Workable Cyber Arms Control Treaties”, [online], *Lawfare* www.lawfareblog.com/private-sector-cyber-intelligence-could-be-key-workable-cyber-arms-control-treaties

PWC. (2017) ”Operation Cloud Hopper”, [online], PwC, www.pwc.co.uk/cyber%0Ahttps://www.pwc.co.uk/cyber-security/pdf/cloud-hopper-report-final-v4.pdf

Reaqta. (2017) “A dive into MuddyWater APT targeting”, [online], Reaqta, reaqta.com/2017/11/muddywater-apt-targeting-middle-east

Reyes, G., Adams, D., Cooper, J. and Camacaro, D. (2017) “EXCLUSIVE: Panama’s expresident wiretapped Americans, according to court documents”, [online], *Univision,* 24 June, www.univision.com/univision-news/latin-america/exclusive-panamas-ex-president-wiretapped-americans-according-to-court-documents

Rid, T. and Buchanan, B. (2015) “Attributing Cyber Attacks”, *Journal of Strategic Studies*, Vol. 38, No. 1–2.

Romanosky, S. and Boudreaux, B. (2019) *Private Sector Attribution of Cyber Incidents: Benefits and Risks to the U.S. Government*, National Security Research Division. RAND, February.

Schmitt, M. and Vihul, L. (2017) “International Cyber Law Politicized: The UN GGE’s Failure to Advance Cyber Norms”, [online], Just Security, www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/

Secureworks. (2017) “BRONZE BUTLER Targets Japanese Enterprises”, [online], Secureworks, www.secureworks.com/research/bronze-butler-targets-japanese-businesses

Shakarian, P., Simari, G.I., Moores, G. and Parsons, S. (2015) “Cyber attribution: An argumentation-based approach”, *Advances in Information Security*, Vol. 56.

Symantec. (2017) “WannaCry: Ransomware attacks show strong links to Lazarus group”, [online], Symantec, www.symantec.com/connect/blogs/wannacry-ransomware-attacks-show-strong-links-lazarus-group

Symantec. (2018) “The Cyber Security Whodunnit: Challenges in Attribution of Targeted Attacks”, [online], Symantec, www.symantec.com/blogs/expert-perspectives/cyber-security-whodunnit-challenges-attribution-targeted-attacks

ThreatConnect. (2017a) “Fancy Bear Pens the Worst Blog Posts Ever”, [online], ThreatConnect, threatconnect.com/blog/fancy-bear-leverages-blogspot/

ThreatConnect. (2017b) “Parlez-vous Fancy?”, [online], ThreatConnect, threatconnect.com/blog/activity-targeting-french-election/

Tsagourias, N. (2012) “Cyber attacks, self-defence and the problem of attribution”, *Journal of Conflict and Security Law*, Vol. 17, No. 2.

Valeriano, B. and Maness, R. (2013) "The Dynamics of Cyber Conflict between Rival Antagonists, 2001-2011", *Journal of Peace Research*, February.

Wheeler, D.A., Larsen, G.N. and Leader, T. (2003) “Techniques for cyber attack attribution”, [online], Institute for Defense Analyses, October, apps.dtic.mil/dtic/tr/fulltext/u2/a468859.pdf

Wikileaks. (2017a) "Development Tradecraft DOs and DON'Ts", [online], wikileaks.org/ciav7p1/cms/page_14587109.html

Wikileaks. (2017b) "Vault 7: Projects", [online], wikileaks.org/vault7

Wikileaks. (2017c) "Vault 7: CIA Hacking Tools Revealed", [online], wikileaks.org/ciav7p1

Yadron, D. (2015) “When Cybersecurity Meets Geopolitics”, [online], *Wall Street Journal*, 23 March, blogs.wsj.com/digits/2015/03/23/when-cybersecurity-meets-geopolitics/

Zetter, K. (2017) "Masquerading hackers are forcing a rethink of how attacks are traced", [online], *The Intercept*, 4 October, theintercept.com/2017/10/04/masquerading-hackers-are-forcing-a-rethink-of-how-attacks-are-traced

Turning the Asymmetry Around: Tactical Principles of Warfare in the Cyber Domain

Mikko Kiviharju and Mika Huttunen
Finnish Defence Research Agency, Riihimaki, Finland
mikko.kiviharju@mil.fi
mika.huttunen@mil.fi

DOI: 10.34190/EWS.20.085

Abstract: This paper analyzes the applicability of five of the essential tactical principles in the logical layer of the cyber domain by cross-referencing them to four different technological elements. The viewpoint chosen here is that of offensive cyber operations. By analyzing the different combinations of technological and tactical principles, the aim is to identify new insights about changes in the tactical principles, and limitations and new opportunities arising from the technological elements. It is argued that the current direction of the asymmetry in cyberspace is not due to the properties of cyberspace itself, but due to the multiplier effect cyberspace imposes on defender and attacker skill levels and resources.

Keywords: Offensive cyber operations (OCO); Tactical principles of warfare; Manoeuvre; Surprise; Use of Terrain

1. Introduction

The use of cyberspace both in the civilian and in the military world grows exponentially. The cyber domain is used for intelligence, national defence and also in the offensive. The cyber domain differs from the conventional domains based on its origin (made by man or other cyber entities) and its rate of change (constantly accelerating, (Fildes, 2010)). Applying the principles of warfare to the cyber domain is still a somewhat unexplored area of research. Possibly the most striking difference between kinetic and cyber warfare is the environment used for warfare and "battles". Kinetic warfare takes place in the physical world, governed by well-known static laws of physics, which apply the same for all parties. Cyber war, on the other hand, is mostly waged in the artificial, man-made terrain, which is constantly under change and highly malleable by the belligerents (Parks & Duggan, 2011).

This paper discusses the applicability of the general tactical principles in cyberspace. These principles have been developed during thousands of years based on experience and learning from battles, campaigns and wars. They are fairly consistent over doctrines of modern state-level actors, when it comes to the physical domain (kinetic warfare), but it is a reasonable question, whether these principles are valid in cyberspace. There is evidence that at least parts of the principles are applicable in the information and cognitive dimensions.

The idea of applying tactical concepts to cyber operations is not by itself novel. It has also been discussed at least by Farmer (2010), McCroskey and Mock (2017) and Crowell (2017). However, the application has been either very closely tied to the physical layer or based on case studies. In general, the related work on this subject tends to be more on the line of existence theorems.

2. Definitions and Scope

In the Finnish cyber security strategy (2013), the cyber domain is defined as a combination of logical and physical layer elements. This definition thus includes also physical devices but considers technologies not currently in the mainstream to be implicitly out of scope. Any "cyber phenomenon" (see Fig. 1) can be thought to travel freely in physical, cognitive and logical layers, but there are also uniquely identifiable time-windows, when some sub-processes of the phenomenon reside solely in the logical layer. The logical layer will experience the effects from cognitive and physical layer as discontinuities, sinkholes and other types of singularities.

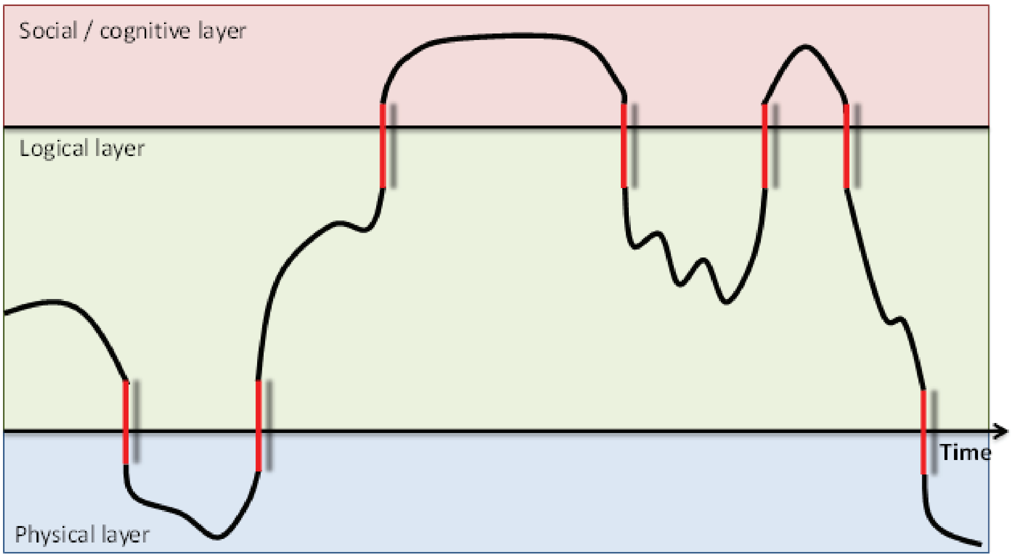


Figure 1: The path of a cyber phenomenon, with discontinuities when crossing the border between social (/cognitive), logical and physical layers

The more precise reasons for handling the logical layer of cyber domain, and scoping the physical and cognitive layers out of this study are as follows:

- To understand and further process the theory typically becomes more efficient by partitioning the problem space.
- Functions such as the creation or moving objects about are fundamentally different on the logical layer compared to the physical. This makes attempts to formulate any cross-layer principles futile, if the layers are not considered separately first: for example movement in cyberspace consists of discrete "jumps" of an entity from one address space to another, which can be physically and logically far apart

It appears that there is an essential gap in understanding cyber-tactical issues in the logical layer. The logical layer is referred to as the particular layer containing the actual "cyber terrain" (McCroskey, 2017).

The concept of "battle" purely in the logical layer of the cyber domain is also somewhat undefined, considering the conventional requirement for use of violence. Here, battle in the cyber terrain is defined as: the intentional (or hostile) confrontation of two or more opponents, which may involve interaction. The main difference with conventional battle definitions is that the time element is allowed for larger variation (case-by-case) in cyberspace, and the interaction need not be violent. Also, the definition of "Military operation" is taken to be a series of battles, their support and preparatory actions following common mission or campaign motives and bounded by common goals, time and place (space).

Current terminology in cyber operations includes such concepts as offensive cyber operations (OCO), defensive cyber operations (DCO) with such flavours as response actions (RA, in practice active defence) and internal defensive measures (IDM, passive defence and defender-network internal actions).

The general tactical principles of warfare are believed to be nearly completely independent on technology and to remain the same from one era to the next. The evolution of these general principles, however, has not come to a halt - instead they are being developed further in different nations worldwide. The goal of this development is to identify those principles that improve national military capabilities (Farmer, 2010).

Some conventional examples of the general principles of warfare are Surprise, Deception, Offensive, Economy of force, Manoeuvre, and Concentration (of forces) . These principles can be perceived to be means applied on all levels of warfare, i.e. battles, operations and the whole war. However, although the means are applied at different levels, they are not the same throughout (Huttunen, 2010).

One of the purposes of this paper is to show, how logical cyberspace phenomena and the principles of conventional warfare can be analyzed. To achieve this goal, a set of cyberspace technological elements were selected, which all affect the general principles of warfare.

The number of different tactical principles (of warfare) per doctrine is between eight and fourteen, depending on the national doctrine or historical reference. The total number of all of the principles across different armed forces' doctrines is approximately 40 (Huttunen, 2010). Five of the most common and fruitful principles have been chosen here to be analysed. Additionally four technological elements/principles have been selected, based on the property that they have specific relevance on tactical principles.

3. Technological principles

3.1 Lack of principle of conservation of energy

The seemingly limitless energy principle of the cyberspace logical layer is called here simply the "energy principle". In physical layer battles, the use and projection of energy has a central role. In cyberspace the very concept of "energy" is different, but not completely independent of physical layer energy. In the tactical and operational frameworks the energy resources of the physical layer can also be seen as strategic resources that need to exist as a whole, and bring other extra resources to cyberspace tactical level at almost no cost.

The common misconception is that due to this technological element, there are no constraints at all on the logical layer. However, the constraints are merely not straightforward to see, such as:

- Limited hard disk space on the physical layer constrains storage space and memory swap space on the logical layer
- Limited number of physical cables or allocation of frequencies gives rise to bandwidth restrictions
- Limited number of processing units on the physical layer prevents actors on the logical layer from using limitless computational power (to e.g. cross security domain boundaries with non-credential-based access or to find zero-day-vulnerabilities efficiently[1])

The capability to transfer and project energy is a universal property of the physical layer, which is not replicated in cyberspace.

3.2 Differences in the distance measures

The pathways on the logical layer consist of physical layer connections, which implies less than 0,1s delay on any possible contiguous pathway on the Earth. Thus, as long as the cyber phenomenon lasts tens of seconds or more, the delays within the planet are irrelevant. On the other hand, the nodes of the network make different kinds of checks and prioritizations to the data packets, which cause extra delay to communications.

The distance measures or metrics of cyberspace are different from the physical layer. Instead of geometric distance, a valid metric is rather the number of connections in the different subnetworks in the logical layer and different limitations posed by connecting nodes than. Also the "location" of an entity is not one single point, but rather a presence on multiple different sublayers and networks, or a collection of logical addresses.

When estimating distances (infinite, if there is no connection) it is important to consider the actual distance instead of the defender's idea of distance: in spite of security policies, communication pathways may exist, e.g. by carrying USB drives. This principle is referred to here as the "distance principle".

3.3 Modifiability of the environment

As man-made environment cyberspace is basically arbitrarily modifiable with very little use of energy: e.g. whole address spaces can be relocated or terminated. Limitations for modifiability are mostly due the size of the difference of the required change, compared to a pre-planned change. There is a very strong dependency between the environment and actions, whereas on the physical layer the environment can be considered independently. Here, this is the "modifiability principle".

This technological element has also some tactical level implications:

- The capability to change the environment (/terrain) is asymmetric: The advantage is held by the environment administrator (the holder of "root" credentials and active root session)

[1] New software vulnerabilities can be found by testing different paths of program execution via input interfaces. In principle, limitless computational power can enumerate all the possible inputs and find all the exploitable vulnerabilities in logical layer elements

- The meaning of accurate and timely intelligence is heightened, forcing the intelligence cycle to need to function faster. This also offers the defender the possibility to avoid hits by controlled environmental changes and fast mobility within defensive address space.

3.4 Discontinuity of the environmental factors

There are many types of discontinuities in cyberspace, due to its networked nature (see Fig. 2): Discontinuities in time; in connections and pathways; in the existence of sublayers and subnetworks of the cyberspace; and in the control of sublayers and subnetworks. This principle is referred to as “discontinuity principle”.

The functional logic in software is almost always stepwise: processors advance from one state to another atomically. On the other hand, environmental factors may change even between state transitions. It then follows that in cyberspace, it is possible for such changes to happen, which are impossible to detect based on environmental, logical layer factors only.

Discontinuity of connections is not completely unknown on the physical layer either (nor in the conventional tactical processes): the logistics change from sea to land transportation is one example. The physical layer can always offer some kind of connection, which gives more manoeuvrability in the meaning of bridging a connection discontinuity (but offers less manoeuvrability in distance).

Possibly one of the less intuitive discontinuities in cyberspace is connected to the control of the environment. Unlike in the physical layer, cyber-terrain is by default always in the control of some entity. However, contrary to the intentions of the original owner or developer of the system, another party can seize the control to herself. In this case, a hostile or inaccessible environment may change into friendly terrain.

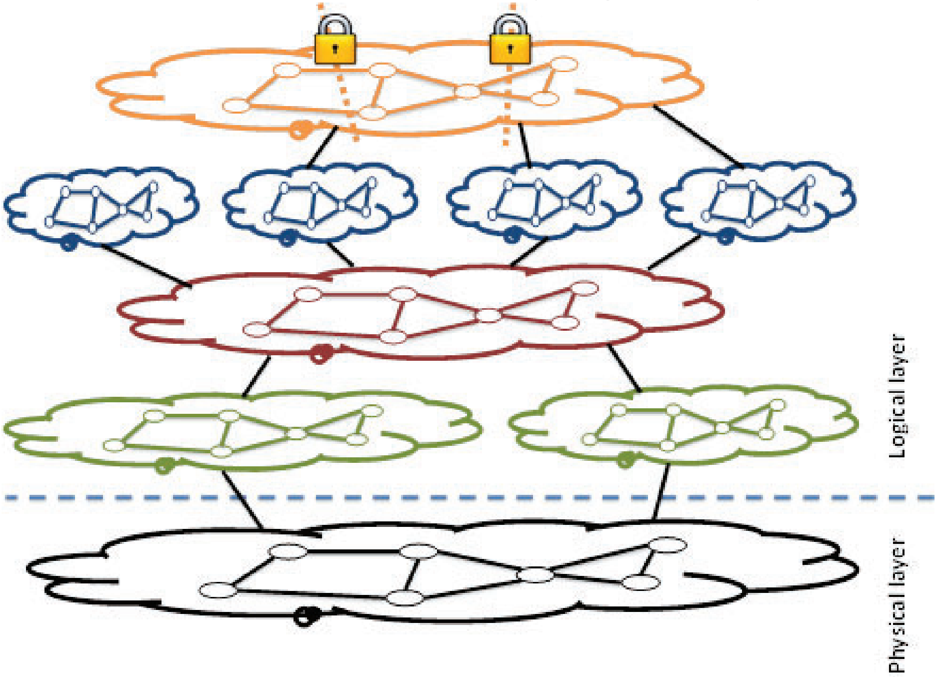


Figure 2: The networked, layered and discontinuous structure of cyberspace. Networks separated by the technology are drawn as separate, and those separated by security policies are shown with locks.

4. General Principles of War

This research is focused on five central (tactical) principles of war: Concentration of forces, Economy of force, Surprise, Use of Terrain and Manoeuvre. Deception is also an important part of operations in cyberspace, and it contributes significantly to the principle of Surprise; however, as a separate principle, the use of Deception is not considered to exhibit essential differences between cyberspace and the physical layer.

Concentration of Forces aims to focus the effect and power of multiple different weapons systems simultaneously to the same target or area. The target or area depends on the level of warfare considered and on the troops and weapons systems used. The concentration presumes that the systems are within range of the target. The goal of concentration of forces is a decisive effect on a battle or conflict, resulting in destruction or disruption. Today, the effect can be concentrated in a network-based manner without having to co-locate troops and systems participating in the concentration effort.

Economy of Force strives to direct the right resources to correct locations in the theatre. There is not enough combat power (as defined in US JCS (1999)) to distribute everywhere – better to be in the right place at the right time. The goal of economic use of force is to complete the mission with minimal losses. As a principle, the Economy of Force is developed from lack of infinite resources, such as troops, platforms, weapons, ammunition, etc.

Surprise is used to find such ways of action, for which the enemy is not prepared or is able to respond slowly to, which translates favourably to the actions of the other party. Surprise is one of the most central and useful principles of war. It is not just a way to improve physical efficiency in the battle, but it has a specific cognitive significance. Thus, the surprise need not be very extensive to have an effect. There can be multiple methods to achieve surprise: deception, timing, placing, tactics, mass and quality, e.g. use of disruptive technology (West, 1969).

Use of Terrain is used here as shorthand to encompass other environmental factors as well, such as weather and time of day. The use of terrain has typically been a method to increase combat power against a superior force. The aim is to direct battles to such location or time, where the superior enemy is not able to use their capabilities fully due to features of the terrain, weather or other environmental factors. Knowing the terrain and environment beforehand increases the combat power for example in the form of increased protection or mobility. The principle includes the implicit assumption that the terrain (or the physical environment) contains larger energy reserves and sinks than any of the parties of the battle can independently control.

Manoeuvre is meant to direct forces and effects in a coordinated way to the flank or rear of the enemy forces, or more generally to such areas that cause a multiplier effects or break the battle plans of the enemy. Manoeuvre aims to achieve such a position and situation that weakens the possibilities for the adversary to fight in coordination. It should be noted that manoeuvre is different from mobility: manoeuvre is tightly coupled with the concept of battle and is understood to be movement in relation to the opponent, or tactical movement (Huttunen, 2010).

5. Art of War in the Cyber Domain

The use of the tactical principles is context-dependent (e.g., relying on legislation and other external factors), and on the tactical level the principles are used together with combat functions and campaign types (Huttunen, 2010). Therefore, it is not straightforward to extend the principles independently to a new domain, such as cyber.

It was found in this research that solely on the logical layer, the cyber domain will change the possibilities for the use of tactical principles, partly strengthening and partly weakening them, partly also modifying the underlying concepts. The following text discusses these factors individually.

5.1 Economy of Force

On the physical layer the law of conservation of energy strongly limits the production and mobility of weapons, ammunition and supplies. In cyberspace, the conservation law does not manifest the same way, making also the requirements and definitions for “economic use” change. For example, one can easily launch network intrusion attacks (i.e. project force and effects) from one location simultaneously to multiple different targets without restriction in geographical distance. This is not possible on the physical layer.

As another example of the differences (to physical layer), it is possible to match resources to follow the actual needs almost exactly and in real time. In conventional warfare, it takes significantly larger amount of time to stop applying weapons, even if the operational picture and communications were on ideal level. In cyberspace, the components producing the effect may itself be intelligent and follow e.g. command channels or time or other logical layer variables, when working in offline mode.

The seemingly limitless energy may make the necessity of the whole principle of economy of force questionable. This is, however, not categorically true: if physical layer tactical analogies are used, this may be the case, but there are restrictions elsewhere, such as:

- The party planning OCOs (or DCO-RA) has only a limited amount of address space available, which typically will wear down in the course of the operation (as opponent's forces will block the offending addresses). This may (case-by-case) be sufficient, but it is highly dependent on the environment.
- The detailed technical methods in OCO are a very limited resource, which is being consumed during an operation. Unlike on the physical layer, where the same force projection mechanism using energy transfer can be applied repetitively, it is possible to "defuse" cyber technologies the first time they are seen. The administrative policies and malware detection efficiency of the defender vary, but the cyberspace itself makes the detailed methods a limited resource.
- The cyberspace targets require typically long periods of intelligence and preparation to make effects possible. For static targets in a visible address space this is still feasible, but during cyberspace manoeuvres new targets will emerge, which may not have as much time for preparation. The time for preparation for operational effects may also be viewed as a resource requiring economic use of force.

The most probable consequence of the changes in the restrictions in projecting force is that the cost functions used during the planning phase need to be rethought.

5.2 Concentration of Forces

The principle of concentration of forces changes radically, if ”forces” still refer to physical space analogies, e.g. to the payload of an exploit or to a malicious data packet. In cyberspace, the component causing the desired effect is most typically a software component. The use of force, when using physical layer analogies, can be then concentrated to multiple targets simultaneously and the effect can be sustained almost limitlessly without drops in efficiency (without defender actions or that the energy demands for the physical layer become intolerable). Furthermore, increasing the power of the attack is as easy as starting it and location of the point of attack can be moved even physically without having to move the attacking components.

However, some of the resources are limited even in cyberspace. These types of resources are needed, for example, to manufacture (/code) weapons systems (/malware) and to deliver the attack to the target. These resources include a limited address space, the instantiations of the attacking components (actual malware), and their exact operation and techniques. A good strategy would then be that "good" malware resources are directed more toward only valuable resources (e.g. Stuxnet used four different zero-day vulnerabilities, which were a rare commodity at that time (Fildes, 2010)). This makes an instance of concentration of forces, only with a different definition of "force".

Force protection can be made easy for a defender even during an attack, as long as the logical layer constraints are considered. To see this, consider that "weapons platforms" can easily (within each address subspace) be moved, replicated and hidden. Replication and hiding can be combined and taken so far as making it very difficult for the attacker to find real targets, which also diminishes the defender losses in that particular address space.

Also, concentration of forces (or actually, their effect) can in cyberspace be temporarily prevented for example by shutting down the system or the part of network phasing the attack (although this may acceptable only in a very limited number of cases).

Cyberspace lacks a universal way to project force (such as projectiles or explosives in the physical world). This means that in cyberspace, the concentration of force cannot be universally accomplished with a similar size "cyber" arsenal as is customary for physical attacks. Instead, to accomplish a sufficient effect on the opponent, the number of attack methods needs to be many times larger and their tempo than on the physical layer. In cyberspace two arbitrary locations do not, by default, have a connecting path to all or any parties, but the connections need to be created in the planning phase to enable the use of force.

Effects occurring in cyberspace share the property that the target recovers generally very quickly, and repeating the exact same effect is challenging (as the defending system is by default able to adapt to the previous type of attack). The attacker usually needs to change address spaces and/or the attacking components' fingerprints / footprints between consecutive attacks.

5.3 Surprise

The use and success of surprise is largely dependent on the opponent's capabilities for planning, preparation and ISR functions. The key here is the ability to monitor and detect cyberspace environmental factors. Whereas in the physical layer the laws of nature and thus the environment provide equal opportunities for all parties, in cyberspace the ISR methods themselves depend on the environment. On the other hand it is possible to modify the cyber environment to such a degree that possible event chains are minimized, which makes surprises less likely.

From the technological principles of energy and distance, it follows that it may not be necessary to move troops or platforms physically at all, or that physical movement is not seen on the logical layer. So movement does not affect detection and thus the surprise as much in cyberspace, although for example creating new physical communications channel for the attacker may create movement relevant for detection (and surprise).

The technological principle of modifiability leads to the property that the different OCO preparatory actions can be detected considerably less efficiently in the cyberspace (than in the physical layer) and that the timeframe in the different attack phases is orders of magnitude shorter than in the physical layer. Due to these factors, cyber-attacks usually come as a surprise.

The discontinuity technological principle acts as an opposite force for the attacker: the possibilities for complete control of defender's own systems enable categorical preparation and quick reactions for attacks, possibly with some loss to defender's capabilities (e.g. following a very tightly defined security policy, cutting off connections or resetting platforms when detecting anomalies). The same tactics can work in the physical layer as well, but the possibilities for use on the logical layer are significantly wider.

5.4 Use of Terrain

The physical layer meaning of ”terrain” refers to such properties of the domain that cannot be affected, but that could still be used. Due to the man-made nature of cyberspace, there is always some party that is capable of affecting any desired environmental factor. This does not mean, though, that the current actor has modification rights to intended factors (a firewall can be shut down, but it requires administrator privileges).

Due to the asymmetry connected to access privileges, the use of terrain in cyberspace is also very asymmetric: the holder of the access rights to the environment has orders of magnitude better possibilities to use the "terrain" than those, who do not (the attackers, by default). This, again, is a cyberspace property: depending on the skill levels of the attacker and defender, the asymmetry may be actually for the benefit of the attacker.

When the logical and physical layers are compared to each other (without asymmetries between opponents), the free space between physical points enabling free movement offers more possibilities for the use of terrain in general. For example, connections between two points need to be explicitly severed (by placing obstacles), whereas in cyberspace they need to be explicitly created. Thus, based on the properties of cyberspace itself, the possibilities to attack are much more limited than on the physical layer: in a typical attack scenario the attacker must advance its presence in the "terrain", of which the defender by default has possibilities for superior knowledge and more access rights.

Additionally, after the attack has been detected, the new exploitation possibilities can be assumed to diminish, since the defender can add limitations by removing access rights, or terminate parts of the terrain completely. This again adds to the asymmetry for the benefit of the defender, not the adversary.

5.5 Manoeuvre

In this text, the manoeuvring entity is considered to be the "presence" of the attacker or defender at each set of logical locations (having activity in those sublayers and logical addresses). Additionally, "movement" is considered to be analogous to that of the physical layer as a change of location of the entity with respect to the environment, and "manoeuvre" to be movement relative to the opponent. However, the layered structure of cyberspace should be considered. Location on each sublayer is typically well-defined (at least if the sublayer is addressed) but the activity on all of the levels of presence does not manifest itself simultaneously, let alone change addresses. This implies the following properties for manoeuvres in cyberspace (compared to the physical layer):

- Movement often is discontinuous, i.e. activity will suddenly appear/disappear on one sublayer and address. In the physical world the troops stay even after a manoeuvre at some location or continue the combat, but this need not be the case in cyberspace: manoeuvres can be stopped instantaneously.
- To get a complete operational picture of the manoeuvre, all of the sublayers of the activity need to be considered (e.g. networking level and persona level).

Isolating actions or troops is another manoeuvre, which is not critically different in the cyberspace: in cyberspace the actions always need connectivity, but unlike on the physical layer, connection is not available by default. Isolation is thus probably more efficient in cyberspace, but on the other hand the success and verification of success of the isolation manoeuvre are difficult.

Manoeuvres in kinetic warfare may attempt to add degrees of freedom to support blue force actions, e.g. by searching for alternative routes to affect targets. In cyberspace tactical manoeuvres this is normal activity, e.g. “lateral movement” across target networks. Tactical manoeuvres to increase degrees of freedom for the attack have thus more significance in cyberspace: without them it may well become impossible to continue the attack, and can be used to create beneficial asymmetry for the defender.

Evasion manoeuvres are an order of magnitude more efficient to conduct than in conventional warfare. For example, in the Georgian war cyberspace operations in 2008 the central government web pages were simply copied far enough with sufficiently many copies to prevent effects (Sydney Morning Herald, 2008). Another instance of this is botnets performing DoS attacks jumping frequently through the address space.

Since new effects can be created at will to nearly anywhere and anytime, why is manoeuvring needed at all? In practice, however, there are restrictions: the systems’ bandwidth and CPU-power are limited and the defender practices counteractions, such as access control. Then the points of presence of the attacker are limited resources for the attacker, which need to be manoeuvred. If the defender has lower skill, the energy principle will simply enable more manoeuvres from one point.

6. Conclusions

At this moment, military operations in cyberspace are more cost-effective for the attacker, even to the degree that it can be termed asymmetric warfare. Viewing purely from the technical platform and the theory connected to it, this does not have to be: the “state-machine” space of the cyberspace is still far simpler than possible actions on the physical layer, and nothing happens against the (actual, instantiated) policy on the logical layer. The opportunities to attack exist only as long as the system owners have an incomplete picture about the actual rules of the system or do not have the resources to instantiate their intent properly. However, based on the current, somewhat extensive experience in software development industry, there simply is not enough resources to design systems securely or to monitor them sufficiently.

This paper postulates that the situation in cyberspace currently is analogous to a situation in conventional warfare, where the opponent has developed superior firepower and mobility compared to the defender, such that old fortifications become obsolete, e.g. German blitzkrieg in the WWII. Then the answer is not to focus the energy on better fortifications, but to increase blue forces mobility and firepower. In other words, the cyber attacker has currently better realized the potential of cyberspace for warfare, and the defender is only slowly catching up. In the long term, the fortifications themselves may again become adequate, as has happened with cryptographic algorithms in the 90s and 00’s.

In conclusion, the findings of the tactical principles in the cyber domain (or its logical layer) can be summarized by the following observations:

- Due to the technological energy principle, attrition warfare is not a meaningful concept. It may then be that manoeuvre is now one of the most important tactical principles, which manifests itself as necessary nearly in all known cyber operations.
- The seemingly endless energy on the logical layer also implies a new connection between the strategic and tactical level: the mere existence of a strategic asset (the physical layer is able to provide energy) enables the near-free creation of resources on the tactical level.
- One of the reasons the cyber attacker is currently winning, is that the defender is not using fully the possibilities offered by the cyber terrain modifiability. Trends looking to rectify this situation include

the concept of Software-Defined-Networking (SDN), which can also be used in cyber defence. This means that for the defender, the “use of terrain” is the most productive principle to focus on.

- The economy of force is likely to be affected most by the difference in the limited resource types between cyberspace and the physical world. As one example, the replication of “weapons platforms” is not a bottleneck, but the development capability of the platforms is.
- The principle of concentration of forces will also face changes, and these can be derived from the change in those resource types, which actually are limited. For example, the novelty of an attack component in cyberspace is a tactical resource that will not be replenished as effects with the new component will only have limited time of use. The concentration of forces cannot in all cases be continuously maintained.
- The cyber terrain should better be taken into consideration during the planning process. The faster and the more innovatively each party of the battle can modify the terrain, the better chances they have of winning.

According to this study, the general principles of warfare are valid in the cyber domain as well, but their resource types and cost functions (at least) will have to be rethought. Thus the art of war and its methods developed during thousands of years can still be applied in cyberspace, as well as in cyber warfare.

References

Crowell, R.M (2017) “Some Principles of Cyber Warfare: Using Corbett to understand War in the Early Twenty-first Century” Corbett Paper N.19, The Corbett Centre for Maritime Policy Studies, Kings College, University of London. January 2017.

European Union Military Staff (2016). “EU Concept on Cyber Defence for EU-led Military Operations and Missions”. EEAS(2016)1597. 22.11.2016.

Farmer D.B. (2010) “Do the Principles of War Apply to Cyber War?”, School of Advanced Military Studies, United States Army Command and General Staff College, Fort Leavenworth, Kansas, AY 2010

Fildes, J. (BBC News - Technology). (2010) “Stuxnet worm targeted high-value Iranian assets”. http://www.bbc.com/news/technology-11388018, 23.9.2010.

Huttunen, M. (2010) ”Monimutkainen taktiikka”, (”The Complex Military Tactics”, in Finnish) Publications of the Department of Warfare (National Defence University) 1, Nr. 1/2010, EditaPrima Oy Helsinki. 2010..

McCroskey, E.D., Mock, C.A. (2017) “Operational Graphics for Cyberspace”. Joint Force Quarterly (JFQ) 85, 2nd Quarter 2017, pp. 42-49. 2017.

Multinational Capability Development Campaign (MCDC) 2013-14: “Combined Operational Access, Cyber Implications for Combined Operational Access Guide and Specifications for the Analysis of the Cyber Domain v 1.0”, Report, 2014

Parks. R. C., Duggan D.F (2011) ”Principles of Cyber Warfare”, IEEE Security and Privacy, vol. 9. no 5, 2011.

Sydney Morning Herald, Moses, A. (2008) “Georgian websites forced offline in ‘cyber war’”, Sydney Morning Herald, Technology, Fairfax Media. 12.8.2008. http://www.smh.com.au/news/technology/georgian-websites-forced-offline/2008/08/12/1218306848654.html, (Referenced 16.8.2017)

Secretary of the Security Committee, Finnish Ministry of Defence (2013). “Finland’s Cyber Security Strategy”, in: https://turvallisuuskomitea.fi/en/finlands-cyber-security-strategy/, 24.1.2013.

West, J. (1969) “Principles of War: A translation from the Japanese”. (Reprinted by Combat Studies Institute) U.S. Army Command and General Staff College, Fort Leavenworth, Kansas.

US JCS, (1999) “Dictionary of Military Terms”, Greenhill books, London. ISBN: 1-85367-386-2, 1999.

US JCS (2014) “Joint Intelligence Preparation of the Operational Environment, p. I-3”. https://fas.org/irp/doddir/dod/jp2-01-3.pdf. , 21.5.2014 (referenced 12.7.2017)

Finnish View on the Combat Functions in the Cyber Domain

Mikko Kiviharju [1], Mika Huttunen[1] and Harry Kantola[2]
[1]Finnish Defence Research Agency, Riihimaki, Finland
[2]Finnish Army Signal School, Riihimaki, Finland
mikko.kiviharju@mil.fi
mika.huttunen@mil.fi
harry.kantola@mil.fi

DOI: 10.34190/EWS.20.086

Abstract: In the wake of development of technology and warfare in general it has become necessary to consider the validity of the traditional combat functions in other domains than kinetic warfare. This text will discuss the applicability of combat functions – fires, mobility and protection – to the cyber domain and cyber conflicts. Compared to the conventional domains, which cover e.g. the physical terrain and circumstances, the cyber domain ”expires” quickly. Military actions in the cyberspace require real-time operational picture and the capability to comprehend the effect caused by changes in the domain. It is important to be able to map the conventional military terminology and concepts between the physical payer and cyberspace, such that people with military background can more easily adapt their knowledge to cyberspace events and also realise, which actions are impossible to perform in the cyberspace and which are completely new. Each of the combat functions is researched independently from the viewpoint of military tactical principles and by delving deep into the very definitions of the concepts. We posit that use of fires is analogous to offensive cyber operations (OCO), mobility to a change in the “cyber terrain” and protection to defensive cyber operations (DCO). In order to make these analogies, we first need to identify the different components making up the combat functions in the physical world and then lift these concepts (not the components themselves) up to the logical layer of cyberspace. Details of each analogy are considered via known cases of cyber operations or operations with a cyber component. Our findings suggest new content to the conventional combat functions, but also that their principles of use may still remain the same. The quick pace of change in cyberspace does not remove the need for planning and planned actions. On the contrary, the ability to understand the effect of the changes adds to need for planned and skillful operations in the cyber domain.

Keywords: Cyber domain; Combat Functions; Fires; Mobility Protection; Principles of War

1. Combat Functions: Fires, Mobility and Protection

The combat functions have conventionally stated to be fires, mobility (or movement) and protection (e.g. Fuller, 1926 and U.S. DoD, 2001). They appear globally across the doctrines of different nations, such as with the U.S.A and Russia. The U.S. military terminology includes concepts such as firepower (fires), protection and mobility / movement[1]. According to Huttunen (2010), in Russia, the corresponding concepts are огонь (ogon’, “fire”), манёвр (manjevr, mobility) and укрытие (ukrytiye, physical protection). The term “combat function” is from NATO (AJP-3.2), but they are also called “tactical functions” (in the British Army Doctrine) or “elements of battle” (Finnish use of the concept) and the aforementioned three functions are the most commonly used, although national extensions exist as well (Huttunen, 2010).

In this paper, we will research each of the combat functions independently from the viewpoint of military tactical principles and by delving deep into the very definitions of the concepts we strive to make analogies to the concepts in the physical world. We first need to identify the different components making up the combat functions in the physical world and then lift these concepts (not the components themselves) up to the logical layer of cyberspace. Details of each analogy are considered via known cases of cyber operations or operations with a cyber component.

Firepower and mobility are physical means to affect the enemy directly with kinetic means. The prerequisite for this is that the combat capability of blue forces is protected. This follows from the fact that in battle both parties will try to project effect to the opponent using similar means. Thus, the combat power is conventionally weakened during the battle both because of expenditure of ammunition and from sustained effect of the opponent’s use of firepower. To sustain combat power requires sufficient blue forces protection (Huttunen, 2010, p.76).

[1] There is a subtle difference between the concepts “maneuver” and “mobility”, which we will discuss later. For the main purposes of this text, we will use the original term in (Fuller, 1926): “mobility”, when we refer to capability and “movement” otherwise

In battle the enemy kinetic effects are directed towards the enemy. These effects have traditionally been built from firepower, such as direct and indirect fires, missiles, mines or other destructive or disruptive weapons effects deployed against enemy troops, structures and weapons systems (including tanks, aerial vehicles, battleships, communications systems, etc.). The effectiveness of the weapons systems used for the kinetic effects has been dependent on the prevalent level of technology (Huttunen, 2010, p.78).

Effects produced by the use of fires can be divided into two types: destructive and disruptive effects. Destructive effects aim to inflict damage toward enemy troops and annihilate targets. Disruption is based, in addition to the destructive effect, on cognitive effects such as psychological phenomena (fear). The objective in the use of fires is to destroy and disrupt opponent targets or to exhaust the enemy. During the attack, use of fires is organized into periods of destruction by firepower. Effectiveness is achieved with high frequency of the use of fire, surprise, concentration of firepower into the most important targets, and wide movement of firing components and wise fire control (Ahromejev, 1986).

Firepower and the use of fire are fairly narrow as concepts, since they describe kinetic effects. Non-kinetic methods, such as electronic warfare, psychological operations and cyber-attacks have traditionally been left outside these concepts.

Mobility is traditionally (since Sun Tzu) been considered as the means to gain advantage in combat. Mobility and movement aim to concentrate battle power, such as troops and weapons system effects, as well as gain dominance of the battlefield (U.S DoD, 2001). Movement can be used to search for cover, or it can also be thought of as an attack.

Mobility represents change, since it can be used to bring changes to the prevalent situation in the battlespace. Movement is a way to gain or bring the opponent to such a position, where it is possible to generate an effect against the opponent. The effect may be, for example, slowing or wearing down, stopping or destroying the enemy (Riihijarvi, 1984).

One should note that mobility (/movement) as a combat function is different from maneuver, which is seen more as a tactical principle of warfare. The most important differences between these concepts (in the physical world) are:

- Mobility is always movement of troops, materiel, weapons platforms etc. in relation to the environment (and thus always an event requiring activity), whereas maneuver is movement in relation to the opponent, and can thus be also passive.
- The concept of mobility is not associated with goals as tightly as the concept of maneuver, where the goal is to gain tactical level advantage over the opponent.
- Mobility is a very straightforward action, i.e. moving from point A (in the environment specified) to point B. Maneuvers, on the other hand, may be highly complex.

We depict the differences between mobility (/movement) and maneuver in Fig. 1. This text only addresses mobility.

In battle, the basis for maintaining and sustaining combat readiness is protection. Protection can be produced via active or passive methods. Active methods include those immediate preventive measures aimed at stopping enemy activity and weapons effects. Passive protection is meant to increase the resiliency of targets or to make them harder to detect (Finnish Army Command, 2004). By extending passive and active protection to different dimensions and domains of warfare the goal is to achieve blue force protection and prevent or weaken different offensive capabilities of the enemy (Huttunen, 2010, p.89).

The combat functions have so far been describing mostly phenomena from physical or kinetic warfare, and their exact meaning has been varying in the course of technological history. More and more dimensions have been identified in the general warfare, however. It is then natural to ask, whether it is still meaningful to use concepts such as “fires”, “mobility” and “protection”, and whether these concepts are still applicable to warfare containing both kinetic and non-kinetic elements.

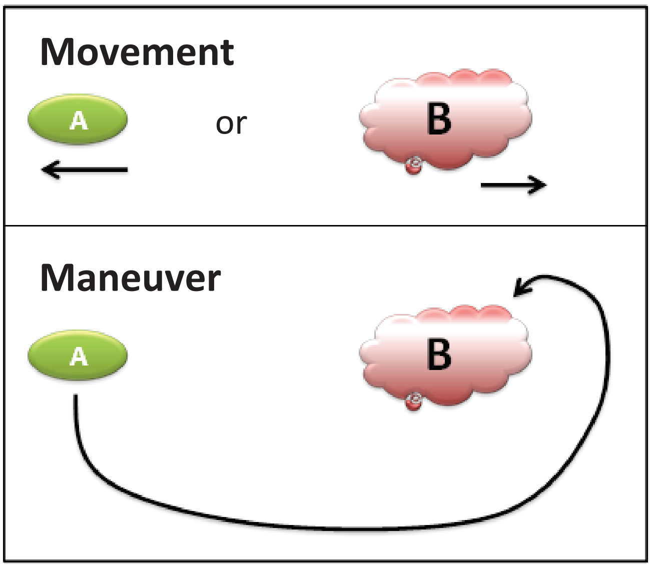


Figure 1: movement and manoeuvre

2. Analogies of the Combat Functions in the Cyber Domain

In order to research, whether the combat functions can be moved to the cyber domain as such (if at all) it is at first necessary to define them closely, with special focus on the purpose of the definitions. After this step, it is natural to try to make layer-independent analogies (i.e. analogies that are as general as possible, when moving them from the physical layer to the logical layer of cyberspace).

The cyber domain itself is a concept understood in quite a many ways. In the Finnish use of the concept, it is possible to use the definition for governmental purposes, which is taken form the National Cyber Security Strategy. This definition includes both layers (physical and logical) to be part of the cyber domain, as long as they are intended for information processing.

A particular feature of the cyberspace is that it is not possible to move conventional, physical layer troops, weapons systems, ammunition or warriors as in other domains, since they are purely physical objects, whereas the cyber domain primarily exists in the logical layer (Applegate, 2012, p.186). On the logical layer, troops of this type are not moved per sé, but the sources of effect and their targets are positioned with each other in the virtual environment according to the planned effects (Parks and Duggan, 2011).

In all of the conventional domains, the actors need to be able to control the basic elements of the battlespace (such as time, location and internal dependencies between the troops) in order to create opportunities to act and fight. Similar interdependencies can also be found in the cyber domain, although the dependencies follow a different patter than what has been traditionally understood. Once control of the battlespace (i.e. the cyber ”terrain”: certain address spaces in the logical sublayers and possibly also some physical nodes) has been acquired, degrees of freedom to perform movement and effects also increase. This type of control in the cyber domain logical layer may in the correct circumstances also enable effects on the physical layer, by controlling weapons systems, influencing the will, actions or decision-making processes of the opponent (Crowell, 2017). In the cyber domain, the control of the environment has bounds in time and location, but they are to be understood differently from the conventional domains (of warfare). In the cyber domain, multiple parties can hold parts of the control from the same system in the same environment (/”terrain”) at the same time. An analogy from ancient sea warfare would be a (hypothetical) situation, where intensely foggy seas would enable multiple warships to move very close to each other without awareness of each other, having the operational picture of complete control (of the local seas).

Another central factor here includes the concept of time. An often quoted statement is that in the cyber domain, events progress extremely fast and that effects can be created and directed by the push of a button (Buchanan, 2016, p.42). This statement is only partly true, since it considers the deployment part of the operation only (the Assault-phase, according to terminology in the work of Andress and Winterfield (2013)). This is preceded in the cyber domain by a significant amount of information gathering and other preparations, which may take a considerable amount of time (Buchanan, 2016). In addition, operations in the logical layer need always be connected with events on the physical layer in the conventional domains, and time (to be considered) needs to be related to these – either to support them or to run parallel with them (Crowell, 2017). These factors need to be taken into account as well, when different combat functions are discussed.

In the following, we will separately discuss the analogies and applicability of different combat functions (fires, mobility and protection) while considering the previously mentioned characteristics of the cyberspace. This part of the study is implemented comparatively by considering the definition of the particular combat function and identified, existing operations. We use the following assumptions of the equivalencies in cyberspace:

- Use of fires corresponds to cyber-attack
- Mobility / movement corresponds to changes in the environment
- Protection corresponds to either cyber defence or change in the environment

2.1 Fires and Use of Fires –Offensive Cyber Operations

The use of firepower is directed to destroy, disrupt or exhaust the enemy or enemy targets (Ahromejev, 1986). Use of fires is, however, only one method to generate effects against the opponent (a very significant method, nevertheless, and historically even the only possible method in addition to psychological effects). The goals of offensive cyber operations (OCO) also aim to destroy or disrupt, but additionally to degrade, deny or deceive opponent services (Applegate, 2012). As concepts the offensive operations on the physical and logical layer are analogous, but this does not directly imply that the cyber domain exhibits a single type of significant method to generate effects: the kinetic use of fires is based on focused transfer of energy in an environment that does not place general barriers for energy transfer. In cyberspace (or the logical layer) nothing is transferred without existing network connections. This missing degree of freedom implies that, for example, there is no direct analogy for "projectiles" on the logical layer.

In order to be able to study the analogies of use of fires at all, the active components ("actors") need to possess a meaningful analogy in the cyberspace. We propose here that the analogy for projectiles can be found via their nature: on the physical layer, projectiles are typically of the type "fire-and-forget" – even though they can contain some intelligence in guiding them to the target and deliver status information to the mission control, it is not usually possible or efficient to control them by the launcher of the projectile. The corresponding actor in the cyberspace (the logical layer) would be a virtual component not using a command-and-control channel, for example viruses, worms and communication packets that do not expect replies.

According to this definition, "cyber firepower" is the planned use of these virtual components as a part of other operational actions. For example, DDoS-attacks usually try to disrupt or degrade services, and according to the definitions here, "use of fire" begins only when the data packets have been sent towards the target (network), not while the other preparatory actions connected to it are underway.

Use of fires in cyberspace targets systems connected to the cyber domain and their functions. Affecting these systems with (cyber-) firepower will accomplish either change in the systems themselves, of in the users of the systems. Examples of cyber-attacks with effects on the physical layer include the attack on a German steel-mill in December 2014 (BSI, 2014) and the disruption of Ivano-Frankivski power grid in December 2015 (Zetter, 2017). In these cases the attacks involved both interactive hacking (logistics and movement) and use of fires. The use of firepower in the steel-mill attack could be understood to be the spear-fishing emails that were sent to the employees of the mill; or a malware using MS Word macros in the case of the Ukrainian power grid, which accomplished altered activity in the users and systems of the grid.

Use of cyber-firepower currently has primarily disruptive effects on target systems. Destructive effects can be directed mostly on virtual (logical layer) targets, but the effects can also be directed – with careful planning – towards physical layer (e.g. Stuxnet).

The cyber-version of the use of firepower here is very detailed activity, i.e. it lies on the tactical level. It is often necessary to use firepower in order to achieve the goals of the operation, but it is rare to generate complex effects with single instances of use, such as deception (modern cyber operational picture is formed via multiple sources). We then distinguish four subgroups for use of fires in the cyber domain: destroy, degrade, alter and block / repel.

Use of fires can be organized as periods also in the cyberspace. The cyber-attack against Ivano-Frankivski did not just impact the power grid functions: firepower was used against the power grid functions at the same time when network and power operator call centres and customer service functions were disrupted (Buchanan, 2016, p.78). Thus, the power companies were not able to receive status information about the operational level of the grid, nor coordinate the repair work. This example proves the existence of properties of surprise, concentration and control of fire, which are criteria for the elements of use of fire.

2.2 Mobility – Change in the Cyber Domain

Mobility and movement are possibly the most difficult combat functions in cyberspace to grasp. In the cyber domain physical troops or soldiers are not moved, but the actions occur primarily in virtual environment. However, cyber domain also has subjects and objects of actions which can generate effects and which can be affected: program code, network data packets, etc. The entities formed by these subjects often encompass several logical sublayers, possibly reaching even down to the physical layer. The location of the entities is well-defined, within the address space of each sublayer, and the entities are moved around, created and deleted. These can be considered to be movement of a set of components / entities with respect to the address space or movement on the logical layer in relation to the environment. An example of an active entity in cyber environment is given in Fig. 2.

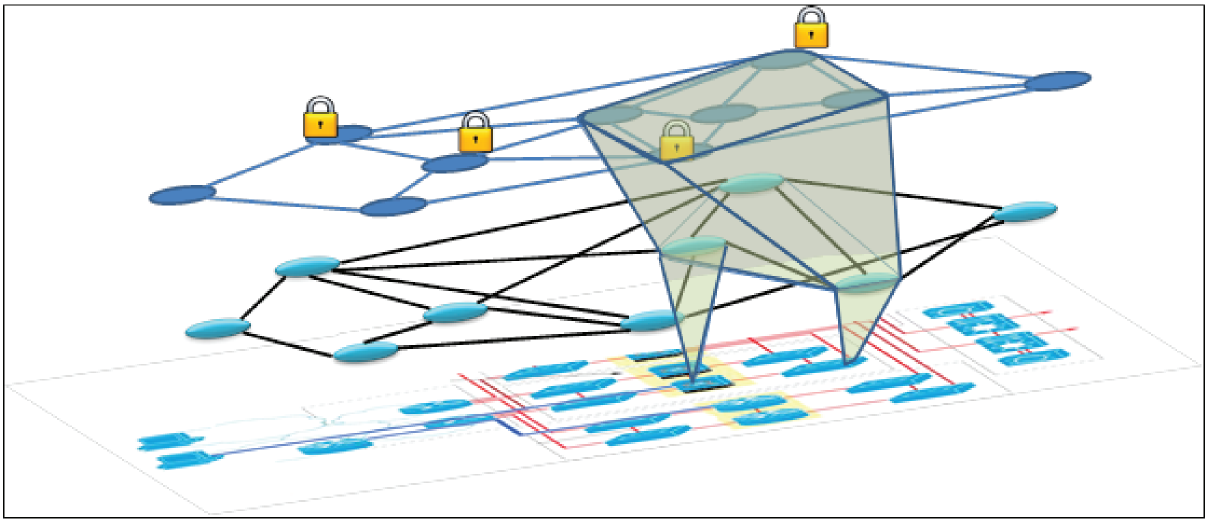


Figure 2: An example of a virtual entity in the cyber environment encompassing different layers from physical layer to several logical sublayers (green shows the area of nodes, where the entity has credentials to operate). Movement of the entity occurs in relation to different parts of the cyber environment, which also changes the ”form” of the entity, and produces conventional movement. All the changes are discrete.

For example, ”compromising a system” (a complete takeover) under these definitions means that the attacker is able to execute their (or some other) program code with the credentials of the system’s original owner on those levels and address spaces of cyberspace covered by the system.

Movement (with respect to the environment) is always enabled by data communications, which in turn requires compatible interfaces and generally existing physical links (also physical ”packets“, such as USB drives apply – an example of a case where movement on the logical layer also requires movement on the physical layer).

Making an analogy from movement to virtual (offensive) components is not, however, the complete picture of the matter: also the environment (“terrain”) itself can be modified, so the components can also move in relation to the terrain even if they are passive. This does not remove the need for data communications, since the reconfiguration of the environment requires communications.

We then propose to have some combination of the following elements to be the analogy of movement in the cyber domain:

- Virtual component changing addresses on different levels (program code, network data packets, entity credentials, etc.). The addresses may be very fine-grained, and not just IP-addresses. In the lowest level “addressing” can also refer to memory addresses on the level of threads of execution.

- Removing a virtual component
- Adding a virtual component
- A communication mechanism allowing “moving”, removing or adding a virtual component
- A (local) change in the environment of the virtual component

With the definitions above, the virtual components are unique, but possibly partly overlapping, unless very detailed and fine-grained sublayers and addressing are used (which we are not aware to be possible in the current operational picture applications).

In general, the analogy of movement in cyberspace can be considered to be the change of virtual components in relation to the cyber domain. Arbitrary changes cannot, however, be considered as movement: it should generate significant enough changes to some addresses of the component, relevant to the operation in battlespace.

Compared to e.g. ground warfare, system compromises are equivalent to building a forward operating base (Applegate, 2012, p. 187). This metaphor is not, however, completely valid in the cyber domain: occupying a tactically important strip of terrain secures a full control of the surroundings for the aggressor. In the cyber domain such situations may arise, where a certain spot in the terrain (i.e. computer system) is occupied by multiple parties, each unaware of each other. A practical example of this is the cyber-attack into the internal information systems of the U.S. Democratic Party during the presidential elections of 2016, where two different actors had penetrated to the same system simultaneously (Alperovitch, 2016).

In the current ICT-networks multi-layer protections are very common. This means that data and services have been divided into different sensitivity levels (classification levels, if they handle governmental information and confidentiality), and between two systems of different sensitivity levels there are always extra controls. The cyber-attack terminology acknowledges here different types of manoeuvres (which are still called movement): lateral movement, where the attacker moves along the protection layers; and horizontal movement, where new (and more sensitive) layers are penetrated. This, however, is movement primarily in relation to the defender’s protection mechanisms, which is why we classify it as manoeuvre. We also would like to note that lateral and horizontal movement both are valid in current rather statically configured networks and may not be meaningful in SDN-controlled and very dynamic networks any more.

With movement, or changes of active components in relation to the address spaces of the cyber domain, it is possible both to search for a better position in the “terrain” (part of mobility) and evade the opponent (protection). In cyberspace, movement can also be generated by modifying the terrain and positioning the (presence of) the parties differently, or moving virtually in and between systems. Change in the cyber domain may exhibit properties from two different combat functions. Mobility can enable access to targets and the use of fires (certain phases of cyber-attack) or enhance passive protection by deforming the terrain. As in the physical world battles, movement can be used for protection or understood as an attack. This principle is analogous to the cyber domain as well.

2.3 Protection – Defensive Cyber Operations or Change

Defensive Cyber Operations (DCO, cyber ”protection” in the Finnish terminology) probably resembles protection in the conventional domains of warfare the most. Protection in cyberspace has been based on firewalls, hardware- and software-based protection (physical airgaps, logical segmentation of networks), and preventing traffic not conforming to security policies from traversing the protected network segments (Applegate, 2012, p.190). These methods are understood rather similarly from the points of view of both the entities in cyberspace and physical world.

The cyber domain can offer an additional method for protection from the conventional domains: the ability to quickly and radically alter the location and form of the terrain itself (Applegate, 2012). This method of protection, along with the concept of counter-attack, is called active cyber defence[2], where one modifies the network environment under their control in order to slow down or degrade opponent’s operations or degrees of freedom, or to directly affect the opponent (Maybaum, 2014).

[2] Or DCO-RA: Defensive Cyberspace Operations, Response Actions

The protection function has generally been considered as inferior to answer all the different types of cyber-attacks (Applegate, 2012). We postulate, though, that this asymmetry does not follow directly from the nature of the cyberspace, but from the different skill levels between the parties and the multiplier effect created by the cyberspace. By considering the possibilities to generate different kinds of movement by changing information system configurations, network topologies, address spaces, user / role credentials etc. defender posture can be strengthened, when the adversary is forced to perform reconnaissance regularly anew and re-customize their attack tools to match the new environment. In the physical world this would equate to placing the attacker into a new terrain with an old or completely inaccurate map. In this case, the attacker would not even need to move, only the terrain would be modified or switched to be unfriendly to the attacker.

Modification of the terrain can also be realized by directing the attacker(s) to a worthless (to the defender) terrain with full of sensors. To the cyber defenders these kinds of false spots in the terrain are known as "honeypots" or "honeynets". These honeynets resemble the actual battlespace to a degree, but contain no critical information, and an attack targeting the honeynet does not compromise any defender services or operations. Using honeynets is more like deception, and it is used also to monitor the detailed actions of the attacker and thus to create better protection methods (Maybaum, 2014, p.22). Honeynets represent the kind of change (movement) that can create new types of protection. Analogously in the physical world, the opponent could be directed to fake targets by camouflaging the neighbourhood of the targets, while the blue forces move to prepared battle stations to observe and strike the attacker only when the circumstances become beneficial.

Honeynets can also be developed to slow down opponent's movement, e.g. by altering the response times of communications and systems to longer than necessary (so-called tar pit technologies - if the delays are taken to other extremes, attackers will merely select other paths and targets). Using these techniques in combination with honeynets, however, may be in conflict with the actual purpose of the honeynets, as the most sophisticated attackers with interesting techniques to monitor are also the most careful. It is therefore important to distinguish those methods of protection intended for deception and those intended for delaying the opponent. An example of the latter is using different SDN technologies.

Attackers also need to protect their operations. In the cyber domain this implies, e.g., covering the virtual tracks, hiding attack code (in defender networks) and protecting the command-and-control channels of the components. Experts have estimated that it is unlikely that the attacker is able to modify the attack without revealing its presence or tactics, if the defender performs changes in the environment in the middle of the attack (Applegate, 2012, p. 186). Every action in the cyberspace can be monitored on some level. Often these signs are not followed appropriately, due to missing sensors or because the signs cannot be combined into meaningful operational picture, if there is not enough capability for analysis and sensor fusion (Applegate, 2012, p.187). Also, different system design errors, such as inadequate protection of log files due to incorrect role-assignment (sysadmin vs. security admin), may result in insufficient operational picture.

According to the definition, active measures of protection consist of those immediate defensive actions that are intended to stop enemy operations and weapons effects. Passive protection is meant to increase the target's resiliency against attacks or to weaken the possibilities to detect them (Finnish Army Command, 2004). In the light of the previous examples we can say that the cyber domain exhibits both active and passive defensive measures. The cyber domain itself (especially the logical layer) offers also a unique type of possibility for protection by using movement (change of and in relation to the environment), by modifying the domain itself. This capability for modification is orders of magnitude larger than in the conventional domains, due to the coverage of the possible modifications and possibly very short time intervals for changes.

3. Conclusions

Implementing cyber defence and operations requires mutual understanding between different cyber actors. This may turn out to be challenging, if personnel and troops on different levels do not understand what is actually meant, when they mention "cyber-attack", "defence", "effect", "disruption", "deception" and other concepts. The operators or the actual technical people speak a very concrete, technical language. This language differs somewhat from the military leaders with less technical background. Also common understanding of the concepts between technically and tactically oriented soldiers varies.

Liles et al claimed in 2012 in an international cyber conference that:

> *"Looking at the conventions of land warfare and the principles of war that constitute strategy and tactics it becomes obvious that there is a substantial disconnect when considering cyber warfare."*

Based on the results of this study, we find that this claim does not hold closer scrutiny. It would seem that there is a difference between two schools of thought concerning some of the concepts and their content. The erroneous claim above likely results from different language used by technical people and general commissioned officers, again due to different terminology and concepts. When we held the combat functions, their meaning and definitions in closer scrutiny, we found that the cyber domain does not differ essentially from other domains of warfare, contrary to some suspicions. This is in concordance with other studies on e.g. general tactical principles of warfare.

When the kinetic battle crosses the border from physical to the logical layer, the contents of the combat functions change technically. Kinetic and non-kinetic environments and systems used in them are different. Similar analogies can be found e.g. between different military branches. For example, when we consider a single weapons platform, there is a distinct difference between the turn of an air fighter in the air domain and a turn of a tank in the ground domain; or between indirect fire of a navy ship to a pointwise target in the sea domain and shelling of field artillery against a larger target area in the ground domain. The combat functions can be found in all of the domains of warfare, but different domains enable or restrict the implementation of the functions differently. This holds true in the cyber domain as well, where planned combat requires (including battles with defensive and offensive focus) deep understanding of the internals of domain. The military planners need only to recognize and understand both the limitations and possibilities of the domain, and be able to be use fires, mobility and protections to the advantage of the mission.

The Finnish approach to the cyber operations follows closely the western world model. Comparing this model to some completely different ones, e.g. the Russian model (Russian Security Council, 2016) is difficult: the Russian military doctrine does not consider "cyber" at all, but rather that information is a vital part of any operation, and it needs to be both protected and manipulated (on the adversarial side). Following this doctrine, it is evident that the Russian approach to e.g. DCO stems from the physical world analogies, namely isolation of domains, even large ones. There is e.g. evidence that the Russian Federation is preparing to take the whole country or a at least some critical areas physically and/or logically off the public, global Internet (Kukkola, 2019).

Even though the model of the cyber domain appears different in the Russian system, compared to the Western view, this does not mean the tactical level actions would not follow the same logic – there are number of ways to arrive to the same courses of action: judging from the Ivano-Frankivski power grid incident in the Ukrainian theatre (Zetter, 2017), the criteria for the elements of use of fire, were all fulfilled and apparently understood by the belligerents.

Finding equivalences between different terminologies in different domains is important. In a U.S. experiment this was expressed as follows (US Army Cyber Command, 2017):

> *"Perhaps the most important lesson learned from the pilot thus far is that cyber effects can be understood and executed much like other warfighting effects such as combat engineering or intelligence operations, allowing cyber planners and operators to use maneuver terms and symbols tactical commanders are familiar with, and to integrate seamlessly into their staffs."*

When we studied the concepts of warfare and their content in the cyber domain, we can note that they have not yet been fully structured. According to a Finnish professor of philosophy, Ilkka Niiniluoto (1980): "*avoiding purely verbal arguments presumes we know and understand the topics of conversation*". Thus the cyber warfare concepts and their contents need to be researched and precisely defined, so we can fully understand them. This principle is not limited only to the concepts of combat functions, but covers also other tactical and operational terminology.

As a conclusion, the trichotomy fires-mobility-protection can be understood in the cyber domain via the analogy of effect-change-protection. By joining the content of these "cyber combat functions" to the terminology of conventional functions, it is possible to create a common view of the content and context of these concepts to all their users.

References

Ahromejev S. F. et al (eds.): ”Vojennyj Entsiklopeditšeskij slovar’” (”Military Dictionary”, in Russian), Vojennoje Izdatelstvo, Moskva, 1986.

Alperovitch Dimitri: ”CrowdSource”. A presentation in CyCon US, 10.10.2016.

Applegate Scott D. “The principle of Maneuver in Cyber Operations”, in 4th International Conference on Cyber Conflict, 2012.

The British Ministry of Defence, Land Warfare Development Centre, “Land Operations”, Army Doctrine Publication AC 71940

Buchanan Ben, “The Cybersecurity Dilemma: Hacking, Trust and Fear between nations”, C. Hurst & co Ltd, London, 2016.

Crowell Richard M: “Some Principles of Cyber Warfare: Using Corbett to understand War in the Early Twenty-first Century” Corbett Paper N.19, The Corbett Centre for Maritime Policy Studies, Kings College, University of London. January 2017.

Federal Office for Information Security (BSI): “The State of IT Security in Germany 2014”, https://www.bsi.bund.de/EN/Publications/SecuritySituation/SecuritySituation_node.html , Referenced in 01.09.2017.

J.F.C. Fuller: “The Foundations of the Science of War”, Hutchinson & Co, London, 1926.

Huttunen, M.: ”Monimutkainen taktiikka”, (”The Complex Military Tactics”, in Finnish) Publications of the Department of Warfare (National Defence University) 1, Nr. 1/2010, EditaPrima Oy Helsinki. 2010.

Liles, S., Rogers M., Dietz E. J, Larson D., “Applying Traditional Military Principles to Cyber Warfare”, 2012 4th International Conference on Cyber Conflict.

Maavoimaesikunta: ”Taistelutekninen suoja” (”Technical Protection in Battle”, in Finnish) in: Military Technical Forecast 2020 (STAE 2020), part 2, Edita Prima Oy, Helsinki, 2004.

Maybaum M., “Responsive Cyber Defence: Technical and legal analysis”, NATO CCD CoE, Tallinn, 2014.

Multinational Capability Development Campaign (MCDC) 2013-14: “Combined Operational Access, Cyber Implications for Combined Operational Access Guide and Specifications for the Analysis of the Cyber Domain v 1.0”. Report, 2014.

Niiniluoto I., ”Johdatus tieteen filosofiaan”, (”Introduction to Philosophy of Science”, in Finnish) Helsinki, Otava, 1980, s. 164

NATO, ”Allied Joint Doctrine for Land Operations”, AJP-3.2

Parks. R, C, ja Duggan D. F, ”Principles of Cyber Warfare”, IEEE Security and Privacy, vol. 9. no 5, 2011.

Riihijärvi P.: ”Tiedon käyttö johtamisessa”, (”Using Information in Leadership”, in Finnish) A Master’s Thesis in National Defence University Course for General Staff Officers nr. 45, Aug/ 1998.

Sun Tzu: ”Sodankäynnin taito” (Finnish translation of ”Art of War”), Tietosanoma oy, Juva 1998.

US Army Cyber Command. “Integration of cyberspace capabilities into tactical units” https://www.army.mil/article/163156/integration_of_cyberspace_capabilities_into_tactical_units, referenced 20.7.2017.

U.S. DoD: “Field Manual FM 3-0, Operations, Department of Army”, Washington DC., 4.7.2001.

Zetter, K., “Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid”, 2016, Wired, 3.3.2016. https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/, Referenced 1.9.2017.

Andress, J., Winterfield, S.: “Cyber Warfare: Techniques, Tactics and Tools for Security Practitioners”, Syngress, October 2013.

Russian Security Council (Совет Безопасности Российской Федерации), “Doctrine of Information Security of the Russian Federation”, http://web.archive.org/web/20191026180449/http://www.scrf.gov.ru/security/ information/DIB_engl/ 2016.

Kukkola, Ristolainen, Nikkarila (eds.): “Game Player: Facing the Structural Transformation of Cyberspace”. Finnish Defence Research Agency Publications 11, 2019. Available online: https://puolustusvoimat.fi/en/web/ tutkimus/publications. 2019.

Cyber Wargaming on the Technical/Tactical Level: The Cyber Resilience Card Game (CRCG)

Thorsten Kodalle
The Bundeswehr Command and Staff College, Germany
thorstenkodalle@bundeswehr.org

DOI: 10.34190/EWS.20.022

Abstract: NATO understands cyber within a cognitive, virtual and physical domain and on technical, tactical, operational, strategic and political levels. The NATO Research Task Group 129 "Gamification of Cyber Defence/Resilience" explores the advantages of gamification, especially Serious Games in an example in the form of manual wargames but also card-driven games and matrix wargames to support training and education in all domains and on all these levels. Within their task is the development of specific prototypes for these specific domains and levels. The Basic Staff Officer Course of The Bundeswehr Command and Staff College (Führungsakademie der Bundeswehr) developed within their competence-based training and education system the “Cyber Resilience Card Game” (CRCG) for training and education on the technical/tactical level. The CRCG was already used several times for training at the NATO Centre of Excellence Defence against Terrorism (NATO COE DAT) in the “Terrorist Use of Cyber Space” course. It is also used within the curricula of the Bundeswehr Command and Staff College and is currently under development for a localised German version to increase awareness and resilience. This paper describes the history of the development of the game, the different iterations, evaluations from the development team, test groups and actual course participants and insights from the most recent research and development workshop in June 2019. It will examine how gamification can be used to incentivise player behaviour that is officially recommended. Essential cyber hygiene measures like Two Factor Authentication (2FA) are usually not applied in real life because they are too inconvenient. It will also illustrate how cyber awareness can be enhanced using a card-driven game. Insights are based on observed player behaviour and the subjective evaluation of players concerning the effectiveness to reach specific cyber-related cognitive learning goals. It describes a good practice approach to a research and development and education (RD&E) approach to awareness building and competence-based training. It provides a blueprint to implement a student-centred and student-driven learning and teaching approach to circumvent death by PowerPoint. This is the first of three articles for the Conference Proceedings of ECCWS 2020. There are redundancies in each introduction, and this article examines terminological uncertainties more.

Keywords: Gamification, Serious Games, Wargames, Cyber Training, Cyber Education, Cyber Awareness

1. Introduction Into NATO SAS 129

The North Atlantic Treaty Organization System Analysis and Study (NATO SAS) 129 Research Task Group (RTG) “Gamification of Cyber Defence/Resilience” started its activity on 12 June 2017 and the current activity end date is 12 June 2021. A NATO SAS RTG is working within the framework of the NATO Science and Technology Organization (NATO STO), and their activity is proposed in a Technical Activity Proposal (TAP) (NATO Science and Technology Organization, 2016) that needs to be approved. NATO SAS 129 TAP was approved in 2016 and described background and justification, objectives, topics, deliverables and the administrative framework of NATO SAS 129. This article will put a particular focus on the demonstrator that was developed at the Bundeswehr Command and Staff College: The Cyber Resilience Card Game (CRCG).

1.1 Background and Justification (Relevance to NATO):

The Wales Summit Declaration in 5 September 2014 (NATO, 2014) provides a basis of justification for this research task group: "We are committed to developing further our national cyber defence capabilities, and we will enhance the cybersecurity of national networks upon which NATO depends for its core tasks, in order to help make the Alliance resilient and fully protected. Close bilateral and multinational cooperation plays a key role in enhancing the cyber defence capabilities of the Alliance. … Technological innovations and expertise from the private sector are crucial to enable NATO and Allies to achieve the Enhanced Cyber Defence Policy's objectives. We will improve the level of NATO's cyber defence education, training, and exercise activities.” (NATO, 2014) Since the Wales Summit NATO issued a cyber pledge and designated Cyber as the 5th domain on the Warsaw Summit in 2016 (NATO, 2016). For NATO a serious cyber-attack could trigger Article 5 (Stoltenberg, 2019), and on the Brussels Summit, 2018 (NATO, 2018a) NATO allies agreed to set up a new Cyberspace Operations Centre (NATO, 2019).

In 2016 the TAP predicted that "Cyber threats and attacks will continue to increase in numbers, sophistication and the potential damage and this activity will contribute to cyber defence resilience in modern NATO environments." (NATO Science and Technology Organization, 2016). This prediction turned out to be accurate; cyberspace is always on, and "NATO is a target three times over." (Omand, 2019), not only the networks of the organisation are targets but also NATO members and soldiers private mobile devices (Grove *et al.*, 2017). Individual soldiers are particularly vulnerable targets (Bay and Biteniece, 2019) and can be "catfished" (Lapowsky, 2019). NATO SAS 129 developed this assessment further based on NATO expectations that the future theatre of war is expected to be mega-cities (Strategic Analysis Branch, 2017, p. 38), where the environment is rich with cyber assets and is developing a Multi-Domain Future Urban Wargame addressing these issues on the tactical and operational level. However, simple cyber hygiene efforts are still the baseline of cyber defence and resilience according to NATO General Secretary Stoltenberg: "Some of the biggest cyber-attacks have only been possible because of human error. Such as picking up an infected USB Drive placed in a car park, and plugging it into a computer. Or clicking on a bad link in a 'phishing' email. It is time we all woke up to the potential dangers of cyber threats." (NATO, 2018c). The TAP recognised in 2016 that "many real-world solutions are available for training and education of cyber experts, there is a lack of training and education of cyber defence/resilience in general. Not many solutions are available for training and education of clients such as end-users, policymakers and military decision-makers." Today there is a significant amount of open-source games available, that specifically target young end-users (Coenraad *et al.*, 2020).

"Conventional methods for raising general awareness is often either costly or ineffective. Therefore one of the possible solutions for training and education is developing serious games and gamification applications." This 2016 TAP statement still rings true today. Advertisement for professional Game-Based Learning (GBL)/Serious Games solutions are built on the premise that PowerPoint presentations for teaching, training, and raising awareness are boring and ineffective. Humans are still the weakest link in any cyber defence, and the low level the adoption of multi-factor authentication (MFA) or two-factor authentication (2FA) is still a concern (Das *et al.*, 2019). Compared with the required vaccination rate for measles and pertussis (92-96%), rubella (84-88%) and mumps (88-92%) (Anderson and May, 1985) we are far from herd immunity although the adoption rate of 2FA increased from 2017 (28%) significantly to 2019 (53%). Also, user awareness rose significantly from 44% to 77% (Engler, 2019). However, opinions differ on the effectiveness (Colnago *et al.*, 2018) (Covello, 2019). There are also different national attitudes concerning providing 2FA for customers from the private sector. The willingness in Germany is lower (t3n Redaktion, 2019) than in the U.S. (ThumbSignIn *et al.*). In the end, our society still provides a big attack surface.

It is assumed in the TAP that Serious Games and Gamification can contribute to solving this problem. It is also assumed "that games are available across platforms and can be designed in a way that it attracts the general audience". For NATO as an organisation, it is essential to understand complex cyber resilience/defence/incident management scenarios. Based on the premise, "that gamification techniques can be useful in training and education regarding different cyber defence/resilience scenarios in a joint and high-pressure environment" the conclusion is drawn "thus, gamification provides opportunities to understand the possibilities inherent in cyber defence and train or educate people while they are having fun, contributing to the goals set forth by the Wales Summit."

1.2 Objective(s):

The stated objective of the TAP is "To effectively enhance information security and cyber defence education and training through the use of serious gaming and gamification approaches."

1.3 Topic To Be Covered:

In the first 2016 TAP the proposed work consists of three main topics:

1. The definition of Serious Game and Gamification and the examination of advantages and disadvantages, common problems during development, gamification characteristics, game mechanics and technologies, and defence applications.
2. The understanding of the big picture of cyber defence and resilience as a baseline for the specification and prioritisation of cybersecurity subjects and user groups that can benefit from utilisation of Gamification and Serious Game applications, classification of operations and decisions in cyber defence and resilience, and examples of cybersecurity training and education.
3. The development of Gamification and Serious Game methodology guidelines for cyber defence and resilience, including the implementation of prototype demonstrations.

1.4 Deliverables

NATO SAS 129 is supposed to submit a final technical report documenting findings on gamification, describing cyber defence and resilience baseline information, and game methodology guidelines with prototypes developed.

1.5 Administrative Framework: Lead Nation, Team Leader and Participants

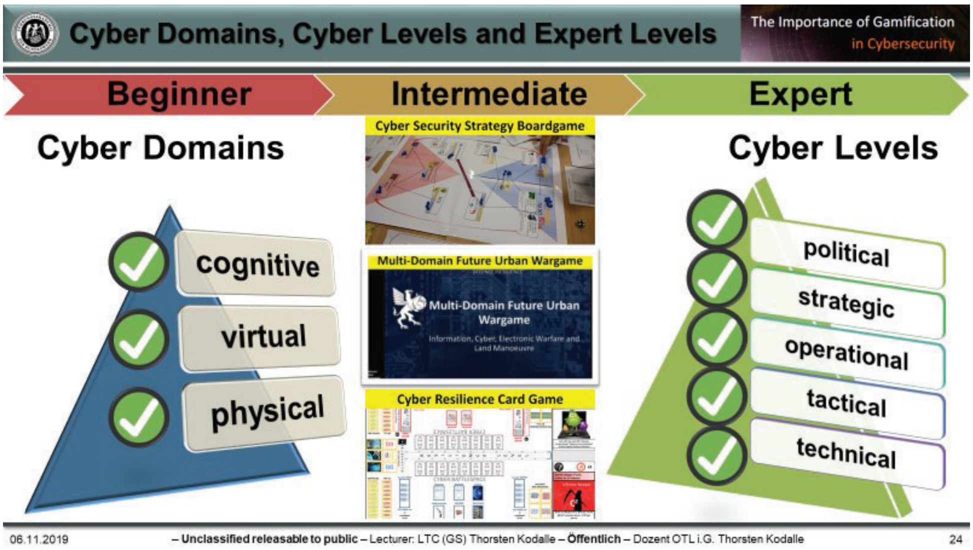
The lead nation for NATO SAS 129 is Turkey, chaired by Mr Levent Berke Capli. Nations and organisations participating are USA, Great Britain, Germany, Netherlands, Turkey, and NATO Cooperative Cyber Defence Center of Excellence (NATO CCD COE) whose efforts include enhancing information security and cyber defence education awareness and training.

2. Terminology: About Cyber and Games

This paper offers an overview of the current discussion of the terminology used in this paper. There is a terminological grey area of the featured CRCG which is a hybrid form of Wargaming, Serious Games, Game-Based Learning, and depending on the usage also Gamification.

2.1 Cyber Defence And Cyber Resilience

There is still an absence of an internationally accepted and harmonised definition of cybersecurity and definitions of the related concepts (Buřita, 2019). NATO as an organisation with 29 different member states and several different languages unsurprisingly differs on when to write “cyber security”, “cybersecurity”, “cyber defense”, “cyberdefense”, “cyberdefence” or “cyber defence”. A keyword search on NATO.int on the 10 November 2019 resulted in 221 entries of “cyber security” and 30 entries of “cybersecurity”. There are nine entries of "cyber defense", one entry of "cyberdefense" (American English), eight entries of "cyberdefence" and 603 entries of “cyber defence” (British English). “Cyber resilience” is used 15 times, “cyberresilience” 0 times, “cyberspace” is found 118 times and “cyber space” 34 times. British English outnumbers American English on NATO.int around 5:1. A search on the 27 February of 2018 resulted in 1670 entries of “defense” and over 9280 entries of “defence”. However, the maximum number of search hits has been limited to 1000, so this relation is no more than obvious. Without going into to details, the author as a certified NATO Cyber Defense Advisor offers his understanding of Cyber as a domain with different subdomains (Cyber Domains) and different levels (Cyber Levels) (see Figure 1).



Cyber Domains, Cyber Levels and Expert Levels

The Importance of Gamification in Cybersecurity

Beginner

Intermediate

Expert

Cyber Domains

cognitive

virtual

physical

Cyber Security Strategy Boardgame

Multi-Domain Future Urban Wargame

Multi-Domain Future Urban Wargame

Information, Cyber, Electronic Warfare and Land Manoeuvre

Cyber Resilience Card Game

Cyber Levels

political

strategic

operational

tactical

technical

06.11.2019 – **Unclassified releasable to public** – Lecturer: LTC (GS) Thorsten Kodalle – **Öffentlich** – Dozent OTL i.G. Thorsten Kodalle 24

Figure 1: Cyber Domains, Cyber Levels and Expert Levels

Figure 1 is a crucial result of the research and development (R&D) Workshop NATO SAS 129 conducted in a joint endeavour with the German Institute for Defence and Strategic Studies (GIDS) 18 -20 June 2019. Three prototypes developed within the framework of NATO SAS 12) are featured in the middle of the slide with the CRGC at the bottom. Three categories "Beginner", "Intermediate" and "Expert" at the top refer to the level of expertise a participant in these games has, either as a cyber expert and as a gamer.

On the left, there are three "Cyber Domains". The "physical" domain includes actual hardware like computers, routers, wires et al. The "virtual" domain includes intangible software and in example logical layers, protocols et al. The "cognitive" domain includes all human mental or mind-based, conscious and unconscious junctures to the other domains. Specific attack vectors focus on specific domains. Social engineering is implemented in the cognitive domain. Logical bombs are planted in the virtual domain and centrifuges in uranium enrichment facilities are blown up on the physical level. These cyber domains also differentiate into the technical, tactical, operational, strategic and political "Cyber Level" (at the right of Figure 1). The private mobile device of a NATO soldier is hardware (physical domain) on the tactical level. A command and control (C^2) server in a corps headquarter (HQ) could be hardware (physical domain) or virtualised in the cloud (virtual domain) on the operational level. The email system of the government is in the virtual domain and on the political level. Disinformation campaigns are effective in the cognitive domain on the tactical level, maybe with huge political effects. Thinking about cyberspace always requires a holistic approach, because everything is connected, literally. The different stages of a unified kill chain (UKC) illustrate this interconnectedness: 1. Reconnaissance, 2. Weaponisation, 3. Delivery, 4. Social Engineering, 5. Exploitation, 6. Persistence, 7. Defence Evasion, 8. Command & Control, 9. Pivoting, 10. Discovery, 11. Privilege Escalation, 12. Execution, 13. Credential Access, 14. Lateral Movement, 15. Collection, 16. Exfiltration, 17. Target Manipulation, 18. Objectives (Pols, 2017, p. 87). Therefore the concept of cyber resilience (or resilience in general) is more important. Resilience in this context means "Having sufficient capability, capacity, and will to endure adversity over time, retain the ability to respond, and to recover quickly from strategic shocks or operational setbacks. (NATO, 2018b, A62). Resilience should include the idea for an organisation to be still functional, even suffering successful attacks, which close to "the ability of a functional unit to continue to perform a required function in the presence of faults or errors. NATO Adopted 2005-03-01 (NATO Standardization Office, 2019, p. 3731).

2.2 Gamification

There is no clear definition of Gamification. According to Merriam-Webster, Gamification is "the process of adding games or gamelike elements to something (such as a task) so as to encourage participation" (Merriam-Webster Dictionaries, 2019). This definition is used in business (Ballou, 2017). The common consensus on Gamification, however, is the exclusion of complete games, which separates Gamification from Serious Games and Game-Based Learning (Werbach, 2019a). The term Gamification is often closely linked to marketing techniques like in this definition: Gamification is "the application of typical elements of game playing (e.g. point scoring, competition with others, rules of play) to other areas of activity, typically as an online marketing technique to encourage engagement with a product or service." (Lexico Dictionaries, 2019). That is the reason why critics among game developers named Gamification "Pointification" (Robertson, 2010). Some game developer avoided the term at al (McGonigal, 2015). This controversy is sometimes lost in translation. McGonigal's book was translated into German and marketed as the first book in German on Gamification (McGonigal, 2016). The TAP defines gamification as "the use of game thinking and game mechanics in non-game contexts to engage users in solving problems and increase users' self-contributions." Wikipedia used this definition in 2015 (Malokin, 2015).

2.3 Serious Games

Clark C. Abt introduced the term "Serious Games" in 1970 (Abt, 1970). The term Serious Games is, according to Abt, an oxymoron: "The oxymoron of Serious Games unites the seriousness of thought and problems that require it with the experimental and emotional freedom of active play." (Abt, 1987, p. 11). Today, Serious Games are closely linked to digital games and digital simulation, and computers are required: "Serious games are games that use computer game and simulation approaches or the technologies for, primarily, non-entertainment purposes (Jansz and Slot, 2019). The TAP defines Serious Games as "simulations of real-world events or processes designed for understanding problems. Although serious games can be entertaining, their main purpose is to train or educate users."

2.4 Wargames and Wargaming

Peter Perla scholarly defines a "wargame" as "a warfare model or simulation whose operation does not involve the activities of actual military forces, and whose sequence of events affects and is, in tum, affected by the decisions made by players representing the opposing sides.' (Perla, 1990, p. 164). Philip Sabin defined recreational wargaming as "military simulation games" (Sabin, 2014, 359). The most recent definition of "Wargaming", provided by James "Pigeon" Fielder, combines previous definitions. Wargaming is "a synthetic decision-making test under conditions of uncertainty against thinking opponents, which generates insights but not proven outcomes, engages multiple learning types, and builds team cohesion in a risk-free environment."

(Fiedler, 2020). NATO SAS 129 does not define Wargaming, neither does NATO SAS-139 Research Task Group on NATO Analytical War Gaming. Another example, that NATO can not agree on the same writing for Wargaming.

2.5 Game-Based Learning (GBL)

Game-Based Learning is another area of terminological confusion. In private business “Game-based learning is training that uses game elements to teach a specific skill or achieve a specific learning outcome.” (Findlay, 2016), whereas scholarly interpretations commonly imply the usage of a complete game for educational purposes: “An educational game is defined as a game being designed and used for teaching and learning.” (Al-Azawi *et al.*, 2016, p. 132). These games used for Game-Based Learning are seen as Serious Games (Noemí and Máximo, 2014), implying that they are digital.

2.6 Terminological Conclusion

The development of a none digital game, with elements of a (cyber)wargame, developed for a serious purpose, used in an educational setting covers all aspects of the terms above. In the end, in cyberspace attribution is difficult as is the attribution of games, mainly, if the picture includes the development process.

3. History of Development

3.1 First Idea

The first idea for a card-driven game was developed at the kickoff meeting of NATO SAS 129 12 to 14 June 2017. NATO SAS 129 was introduced to the commercial of the shelf (COTS) cyber card-driven game “Android Netrunner” developed by Richard Garfield. Android Netrunner is an asymmetric two-player game that plays in a dystopian future about 300 years from now. It is entirely card-driven, with no dice, much tactical depth and was played in international competitions like Richard Garfield's "Magic The Gathering", the first trading card game ever and still prevalent. Clausewitz compared card games and war: “makes war of all branches of human activity the most like a game of cards.” (Clausewitz, 1873) (Clausewitz, 2016, p. 26). The author, who is a member of the Bundeswehr Command and Staff College stationed in the "Clausewitz Kaserne" in Hamburg, was intrigued by the idea to develop a cyber-related card-driven game prototype for the present.

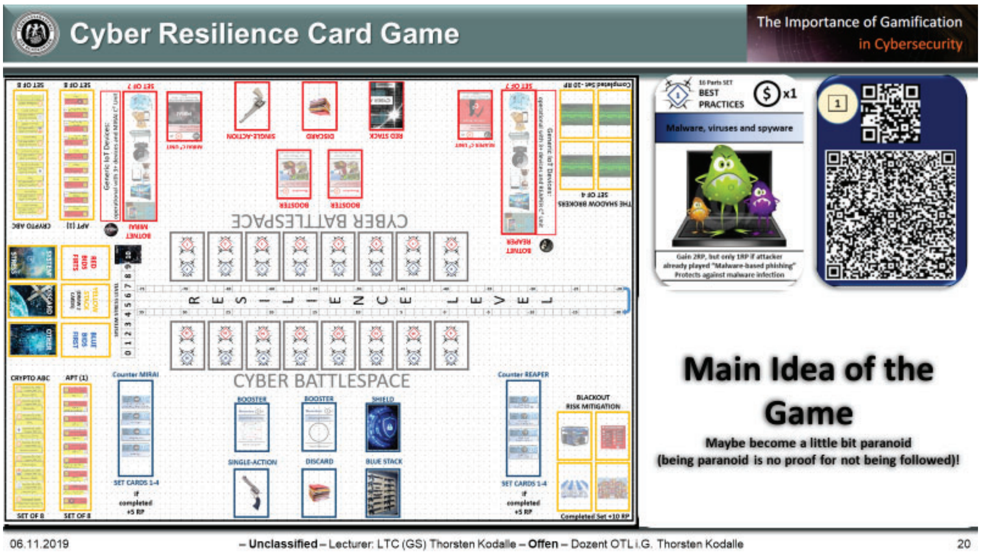
3.2 Development And Proof Of Concept

The author followed a basic Design Thinking approach. Design is a process, a general approach to tackling challenges that are particularly useful for gamification. Design Thinking is a targeted, iterative process that is geared towards people and seeks a balance between analytical and intuitive thinking, i.e. between algorithms and heuristics. An essential aspect of design practice is starting with a rough prototype that conveys the basic structure of the game and testing with real people in order to further develop and refine the game based on these experiences. This typical design process is also suitable for gamification (Werbach, 2019c). According to Andrzej Marczewski (Marczewski, 2018), the design thinking approach follows five steps: 1. Define, 2. Empathy, 3. Ideas (“developing ideas”), 4. Experiment (experimentation) and 5. Test (testing). Marczewski also developed the simple gamification framework GAME. It describes the necessary steps: 1. Gather (gathering information), 2. Act (Act - Prototype Development), 3. Measure (the measurement of feedback) and 4. Enrich (enrichment of the system in an iterative process). To develop a cyber resilience card-driven game that would work for NATO, the author and his students aimed to identify the successful game mechanics, game principles and game elements from Android Netrunner. For this endeavour, the Bundeswehr Command and Staff College acquired ten sets of Android Netrunner, and the 3rd class of the Basic Staff Officer Course (Basislehrgang Stabsoffizier – BLS) of 2017 got the task to develop a current version of Android Netrunner that they would like to play. Marking the starting point of development of the "Cyber Resilience Card Game" (CRCG).

The first step was learning to play the game “Android Netrunner” and embracing the fun of playing the game, the tactical depth and understanding what makes playing the game fun. Based on observational data, the game was fun; indeed, because the students played the game also in their free time for fun. Which provides the insight that engaging learning material has the potential to drive student motivation even after class is over. However, one of the first insights while playing Android Netrunner was, that learning to play a game with deep tactical depth is consuming much time. Students needed about five hours of playtime to understand game mechanics completely.

3.3 Iterations

After playtesting the very first prototype with a control group within the BLS 3-2017 and its first implementation in the "Terrorism Usage of Cyber Space Course" at the NATO COE DAT, the game evolved regularly based on player feedback. The first significant milestone of development was the creation of a game board by students. The first version of the game board already depicted the basic conceptual ideas about card arrangement on the table, but it developed considerably. Figure 2 shows the most advanced version. The game board provided visual clues and is a "scaffolding" technique.



Cyber Resilience Card Game

The Importance of Gamification in Cybersecurity

Main Idea of the Game

Maybe become a little bit paranoid (being paranoid is no proof for not being followed)!

06.11.2019 – Unclassified – Lecturer: LTC (GS) Thorsten Kodalle – Offen – Dozent OTL i.G. Thorsten Kodalle 20

Figure 2: Cyber Resilience Card Game Board Layout

Players who are unfamiliar with card game concepts like Rummy are sometimes confused and have a fundamental question of where to put a card on a table. There was also a specific demand by test players to write more details into the rules. The author encountered test players who never played cards in their life before. The rule to shuffle the deck before playing was not considered necessary from the developers, until the first contact with these specific group of players. The game evolved in three significant steps during the BLS classes of 1st 2018, 2nd 2018,1st 2019 and a specific effort during an R&D NATO/GIDS workshop in June 2019. Alongside were three playtests at the NATO COE DAT in Ankara, 2 in seminars at The Bundeswehr Command and Staff College and presentation at the ECCWS 2018 and ICCWS 2019 and within the Bundeswehr Cyber Defence Community. The NATO/GIDS workshop in June 2019 provided the current status quo. The game is ready for implementation in a seminar setting and requires a trained facilitator and about 3h of time for introduction and gameplay. It can be extended for longer by adding another layer of R&D by the players in the sense that player is identifying content that could be added to the game or even creating cards on their own.

3.4 Status Quo

The most recent updated edition of the CRCG is available in this dropbox folder (Kodalle, 2020).

4. Evaluations

As a metric for evaluation, the author used Teresa de la Hera Conde-Pumpido's "Conceptual Model for the Study of Persuasive Games" (de la Hera Conde-Pumpido, Teresa, 2013). The author conducted a survey-based on this concept and asked players how much they felt persuaded on the specific levels of "Signs", "Systems" and "Contexts". The CRCG was compared with the web-based training course "Cyber-Hygiene Check-Up" (Aufbaustab Cyber- und Informationsraum, 2016, p. 36) of the German Ministry of Defence (MoD), the French MoD board game "Cyber Strategia" (Ministère de la Défense et des Anciens combattants, 2016, p. 10) and the U.S. Post Naval Graduate School's Serious Game (digital) "CyberCIEGE" (Thompson and Irvine, 2011).

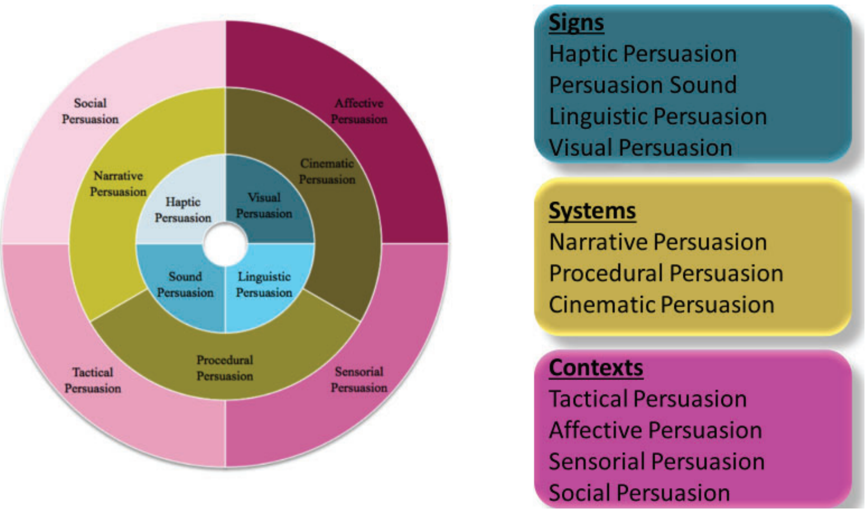


Figure 3: Conceptual Model for the Study of Persuasive Games", adapted from de la Hera Conde-Pumpido, Teresa, 2013

The author discussed the approach with experts in the area of wargaming in general and in developing cyber wargames in particular. During playtesting and seminars, the author made field observations, and "lessons learned" were identified. Player feedback was collected, and easy adaptions of rules were sometimes implemented instantaneously during the game.

4.1 Development Team, Test Groups And Course Participants

The survey was conducted with the development team (17 students) and a control group (17 students) in the same BLS. Also, the author played the game at the 2018 “Terrorist Use of Cyberspace Course” at NATO COE DAT and surveyed with the participants (54 students). The survey also included questions about the applicability of the CRCG to cover the topics of the NATO’s 2016 “Cybersecurity A Generic Reference Curriculum” (Partnership for Peace Consortium Emerging Security Challenges Working Group, 2016).

4.2 NATO/GIDS Workshop June 2019

The NATO/GIDS Workshop (23 international diverse experts covering cybersecurity, history, wargaming) provided valuable insights as pointed out above in 2.1. More details are provided in an accompanying paper.

4.3 Results

The development team valued the level of persuasion in the different areas the highest, particularly in comparison with the other games. The German control group evaluated the CRCG at a similarly high level without playing any other games in comparison. The majority of the participants in all groups recognised a higher degree of expertise in the topic of cyber after playing the game. The international group at NATO COE DAT felt significantly less persuaded than the German groups. However, participants evaluated the game at the highest rate in comparison to other course content. The author expected these results. Playing a game is usually more fun and engaging than listening to a PowerPoint presentation. The proof of concept was the question if the CRCG could successfully address different areas of the NATO’s Generic Reference Curriculum (Partnership for Peace Consortium Emerging Security Challenges Working Group, 2016) from the participant's point of view. This is, in essence, a design thinking approach asking the students if learning goals can be reached and implying that they would prefer to learn topics in a gamified or game-based learning approach. They were provided with 74 topics from the curriculum and evaluated 51 suitable to be addressed with the CRCG. From the author's point of view, this is the proof of concept for addressing topics in cyber education with gamified and game-based learning approaches in general and advancing the CRCG in particular.

5. How To Incentivize Player Behavior Through Persuasive Game Design

Player behaviour can be incentivised in two main ways, either by extrinsic motivation or by intrinsic motivation. Extrinsic motivation in the worst-case exploits addiction loops in our brain and manipulates the player (Werbach, 2019a). In extreme cases, this approach leads to demotivation. That is the main reason why some game developers refer to Gamification as Pointification. On the other side, players get incentivised by meaningful action, purpose and self-determination in a game. In this case, Self Determination Theory is leveraged to make players play a game (Werbach, 2019b). Considering these two perspectives, a developer has a different development framework available. The two most commonly used Gamificationframeworks are Kevin Werbach's "D6" (Werbach and Hunter, 2015) and Yu-Kai Chou's "Octalysis" (Chou, 2015). In addition to these two prominent frameworks, there are other frameworks available, "The Persuasive Game Design method" (PGD) in particular provided useful inputs for the development of Serious Games (de la Hera Conde-Pumpido, Teresa, 2013).

6. How The CRCG Enhances Cyber Awareness And Increases Resilience

A particular feature of the CRGC is the implementation of cyber hygiene measures during gameplay as real-life actions. Players have to check their Amazone account, download a password manager and the Google Authenticator. With all these in-game activities, the player leaves the CRCG more resilient than she entered. Real-life action has a higher learning retention rate than passive consumption of learning material (presentations, reading, watching videos). On the cognitive level, players learn about threats, and on the emotional level, players fell threatened because they realise how much ground they still have to cover to close all the open attack vectors they just experienced during gameplay and were not able to close with real-life actions.

7. Conclusion

In contemporary private business, there is the anticipation of a "results only working environment" (ROWE) with the capability of "always online real-time access" (AORTA). For education, there is the idea of providing a holistic learning environment that encourages student-centric learning behaviour: self-driven and without limitations to time and space. The gamification of (cyber) education embraces this framework. A card-driven game, playable only on-premise is not the ideal contribution for this environment. An app for mobile learning would be the ideal solution. The specific feature of the CRCG, enhancing the individual, private cyber resilience of the individual player in managing their smartphone during gameplay, using the expert knowledge that is available on site, makes it an ideal contribution for a seminar setting and enhances game-based learning. On the content level, exploring content and acting upon it is an educational value in itself. However, on the next level, students can create their own content and thereby training their 21st 4C century skills (communication, collaboration, creativity and critical thinking) in the domain of cyber. Content needs to be verified for relevance and actuality, new content can be added, and new ideas re-engineered into new game mechanics. The endeavour of developing the CRCG in several iterations is thereby a never-ending educational story. The facilitator is the storyteller, and he tells a story of purpose and deep meaning for the individual player and the sense of being distinguished as a multiplier for (especially older generations) a particular set of skills around private cyber resilience. Thereby the player plays a vital role in closing capability gaps in our societies and reducing the attack surface of the whole society.

References

Abt, C.C. (1970), *Serious games,* 1. publ, Viking Press, New York NY.

Abt, C.C. (1987), *Serious games*, University Press of America, Lanham, London.

Al-Azawi, R., Al-Faliti, F. and Al-Blushi, M. (2016), "Educational Gamification Vs. Game Based Learning: Comparative Study", *International Journal of Innovation, Management and Technology*, pp. 131–136.

Anderson, R.M. and May, R.M. (1985), "Vaccination and herd immunity to infectious diseases", *Nature*, Vol. 318.

Aufbaustab Cyber- und Informationsraum (2016), "Abschlussbericht Aufbaustab Cyber- und Informationsraum. Empfehlungen zur Neuorganisation von Verantwortlichkeiten, Kompetenzen und Aufgaben im Cyber- und Informationsraum sowie ergänzende Maßnahmen zur Umsetzung der Strategischen Leitlinie Cyber-Verteidigung".

Ballou, E. (2017), "How to Use Gamification in Mobile App Onboarding", *Boxes and Arrows LLC,* 17 October, available at: http://boxesandarrows.com/how-to-use-gamification-in-mobile-app-onboarding/ (accessed 1 March 2020).

Bay, S. and Biteniece, N. (2019), "The current digital arena and its risk to serving military personnel", *NATO STRATCOM COE*, pp. 7–18.

Buřita, L. (2019), "Online glossary of cyber security", in Cruz, T. and Simoes, P. (Eds.), *ECCWS 2019 - PROCEEDINGS OF THE 18TH EUROPEAN CONFERENCE ON CYBER WARFARE AND*, ACPIL, [S.l.], pp. 72–77.

Chou, Y.-K. (2015), "Gamification & Behavioral Design. Octalysis: Complete Gamification Framework", available at: https://yukaichou.com/gamification-examples/octalysis-complete-gamification-framework/ (accessed 4 January 2020).

Clausewitz, C.v. (1873), "On War. Book 1, Chapter 1", available at: http://www.clausewitz.com/readings/OnWar1873/BK1ch01.html#a (accessed 1 April 2020).

Clausewitz, C.v. (2016), *Vom Kriege: Vollständige Ausgabe der acht Bücher,* 1. Auflage, Contumax; Hofenberg, Berlin.

Coenraad, M., Pellicone, A., Ketelhut, D.J., Cukier, M., Plane, J. and Weintrop, D. (2020), *Learning Cybersecurity One Game at a Time: A Systematic Review of Cybersecurity Digital Games: under review*.

Colnago, J., Devlin, S., Oates, M., Swoopes, C., Bauer, L., Cranor, L. and Christin, N. (2018), ""It's not actually that horrible"", in CHI (Ed.), *CHI 2018: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, April 21-26, 2018, Montreal, QC, Canada, Montreal QC, Canada, 21.04.2018 - 26.04.2018*, ACM, New York, NY, pp. 1–11.

Covello, B. (2019), "Opinion: Back to the Start for 2FA Adoption?", available at: https://www.tripwire.com/state-of-security/security-awareness/back-to-the-start-for-2fa-adoption/ (accessed 1 April 2020).

Das, S., Wang, B., Tingle, Z. and Camp, L.J. (2019), *Evaluating User Perception of Multi-Factor Authentication: A Systematic Review*, available at: https://arxiv.org/pdf/1908.05901.

de la Hera Conde-Pumpido, Teresa (2013), "A Conceptual Model for the Study of Persuasive Games", *Proceedings of Digra 2013: Defragging Game Studies*.

Engler, M. (2019), "State of the Auth".

Fiedler, J. (2020), "Reflections on Teaching Wargame Design", *War on the Rocks,* 1 January, available at: https://warontherocks.com/2020/01/reflections-on-teaching-wargame-design/ (accessed 1 March 2020).

Findlay, J. (2016), "Game-Based Learning vs. Gamification: Do You Know the Difference?", available at: https://trainingindustry.com/articles/learning-technologies/game-based-learning-vs-gamification-do-you-know-the-difference/ (accessed 6 October 2019).

Grove, T., Barnes, J.E. and Hinshaw, D. (2017), "Russia Targets NATO Soldier Smartphones, Western Officials Say", *Wall Street Journal,* 4 October, available at: https://www.wsj.com/articles/russia-targets-soldier-smartphones-western-officials-say-1507109402 (accessed 1 April 2020).

Jansz, J. and Slot, M. (2019), "The definition and characteristics of serious games - Introduction and Definition | Coursera", available at: https://www.coursera.org/lecture/serious-gaming/the-definition-and-characteristics-of-serious-games-Aa56e (accessed 1 March 2020).

Kodalle, T. (2020), "Cyber Resilience Card Game – Dropbox", available at: https://www.dropbox.com/sh/w5q77ag34vbyg6e/AACI6av0SeR6zGMlnffo-cS6a?dl=0 (accessed 1 April 2020).

Lapowsky, I. (2019), "NATO Group Catfished Soldiers to Prove a Point About Privacy |", available at: https://www.wired.com/story/nato-stratcom-catfished-soldiers-social-media/ (accessed 4 January 2020).

Lexico Dictionaries (2019), "Gamification | Definition of Gamification by Lexico", available at: https://www.lexico.com/en/definition/gamification (accessed 14 September 2019).

Malokin, A. (2015), "Gamification for Public Transit - MARTA", available at: https://leadership.saportareport.com/transit/2015/08/11/gamification-for-public-transit/ (accessed 1 March 2020).

Marczewski, A. (2018), *Even Ninja Monkeys Like to Play: Gamification, game thinking and motivational design,* Unicorn edition, Independently published.

McGonigal, J. (2015), *SuperBetter: A revolutionary approach to getting stronger, happier, braver and more resilient--powered by the science of games*, Penguin Press, New York.

McGonigal, J. (2016), *Gamify your life: Durch Gamification glücklicher, gesünder und resilienter leben**, *Dank revolutionärer Ansätze aus der Games-Wissenschaft, Herder, Freiburg, Basel, Wien.

Merriam-Webster Dictionaries (2019), "Definition of GAMIFICATION", available at: https://www.merriam-webster.com/dictionary/gamification (accessed 14 September 2019).

Ministère de la Défense et des Anciens combattants (2016), "OBSERVATOIRE DU MONDE CYBERNÉTIQUE", available at: https://www.defense.gouv.fr/content/download/492200/8404221/file/OBS_Monde%20cybern%C3%A9tique_201612.pdf (accessed 1 April 2020).

NATO (2014), "Wales Summit Declaration. issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Wales", available at: https://www.nato.int/cps/en/natohq/official_texts_112964.htm?selectedLocale=en (accessed 10 November 2019).

NATO (2016), "Warsaw Summit Communiqué. Issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Warsaw, 8-9 July 2016", available at: https://www.nato.int/cps/en/natohq/official_texts_133169.htm?selectedLocale=en (accessed 10 November 2019).

NATO (2018a), "Brussels Summit Declaration. issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Brussels, 11-12 July 2018", available at: https://www.nato.int/cps/en/natohq/official_texts_156624.htm?selectedLocale=en (accessed 10 November 2019).

NATO (2018b), "FRAMEWORK FOR FUTURE ALLIANCE OPERATIONS", available at: https://www.act.nato.int/images/stories/media/doclibrary/180514_ffao18.pdf (accessed 4 January 2020).

NATO (2018c), "NATO - Opinion: Speech by NATO Secretary General Jens Stoltenberg at the Cyber Defence Pledge Conference (Ecole militaire, Paris), 15-May.-2018", available at: https://www.nato.int/cps/en/natohq/opinions_154462.htm (accessed 1 April 2020).

NATO (2019), "Cyber defence", available at: https://www.nato.int/cps/en/natohq/topics_78170.htm (accessed 10 November 2019).

NATO Science and Technology Organization (STO) (2016), *Technical Activity Proposal: TAP*.

NATO Standardization Office (NSO) (2019), *Terminology extracted from NATOTerm*.

Noemí, P.-M. and Máximo, S.H. (2014), "Educational Games for Learning".

Omand, D. (2019), "Cyber Threats to NATO", *Ares & Athena*, No. 16.

Partnership for Peace Consortium Emerging Security Challenges Working Group (ESCWG) (2016), "Cybersecurity. A Generic Reference Curriculum", available at:

https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2016_10/20161025_1610-cybersecurity-curriculum.pdf (accessed 4 January 2020).

Perla, P.P. (1990), *The art of wargaming: A guide for professionals and hobbyists / Peter P. Perla*, Naval Institute Press, Annapolis, Md.

Pols, P. (2017), "The Unified Kill Chain: Modeling Fancy Bear Attacks. Designing a Unified Kill Chain for analyzing, comparing and defending against cyber attacks", Cyber Security Academy (CSA), 2595 AN The Hague, 7 December.

Robertson, M. (2010), "Can't Play, Won't Play", available at: https://kotaku.com/cant-play-wont-play-5686393 (accessed 12 October 2019).

Sabin, P.A.G. (2014), *Simulating war: Studying conflict through simulation games,* Kindle edition, Bloomsbury Academic an imprint of Bloomsbury, London, Oxford, New York, New Delhi, Sydney.

Stoltenberg, J. (2019), "NATO WILL DEFEND ITSELF. The alliance will guard its cyber domain—and invoke collective defence if required", *Prospect*, October 2019, p. 4.

Strategic Analysis Branch (2017), "Strategic Foresight Analysis. 2017 Report", available at: https://www.act.nato.int/images/stories/media/doclibrary/171004_sfa_2017_report_hr.pdf (accessed 17 January 2020).

t3n Redaktion (2019), "Deutsche Firmen drücken sich vor Zwei-Faktor-Authentifizierung", available at: https://t3n.de/news/deutsche-firmen-druecken-1203083/ (accessed 1 April 2020).

Thompson, M. and Irvine, C. (2011), "Active Learning with the CyberCIEGE Video Game".

ThumbSignIn, One World Identity and Gluu, "Customer Authentication Practices 2019 Survey", available at: https://thumbsignin.com/customer-authentication-report-2019/ (accessed 1 April 2020).

Werbach, K. (2019a), "6.2 Dangers of Behaviorism", available at: https://www.coursera.org/lecture/gamification/6-2-dangers-of-behaviorism-r1ZNz (accessed 4 January 2020).

Werbach, K. (2019b), "6.5 Self Determination Theory - Motivation and Psychology", available at: https://www.coursera.org/lecture/gamification/6-5-self-determination-theory-DI0di (accessed 4 January 2020).

Werbach, K. (2019c), "7.1 The Design Process - Design", available at: https://www.coursera.org/lecture/gamification/7-1-the-design-process-n6DGi (accessed 2 January 2020).

Werbach, K. and Hunter, D. (2015), *The Gamification Toolkit: Dynamics, Mechanics, and Components for the Win*, Wharton Digital Press, Chicago.

Using Machine Learning for DNS over HTTPS Detection

Michal Konopa[1], Jan Fesl [1,2], Jiří Jelínek[1], Marie Feslová[1], Jiří Cehák[1], Jan Janeček [1,2] and František Drdák[1]

[1]University of South Bohemia, České Budějovice, Czech Republic

[2]Czech Technical University in Prague, Czech Republic

mkonopa@seznam.cz
jfesl@prf.jcu.cz
jjelinek@prf.jcu.cz
f.marienka@seznam.cz
janecek@fit.cvut.cz
fdrdak@prf.jcu.cz

DOI: 10.34190/EWS.20.001

Abstract: DNS over HTTPS (DoH) is a new standard that is being adopted by most of the new versions of web-browsers. This protocol allows translating the canonical domain name to an IP address by using the HTTPS tunnel. The usage of such a protocol has many pros and cons. In our paper, we try to evaluate these aspects from different points of view. One of the most critical disadvantages lies in the much more complicated possibility of network traffic logging. Our team has created a machine learning-based approach allowing automated DoH detection, which seems to be pretty well usable in advanced firewalls.

Keywords: DNS over HTTPS, detection, machine learning, IPFIX

1. Introduction

Domain Name System is the principal part of the Internet infrastructure. Its primary role is to translate *domain names* to numerically defined *IP addresses*. An IP address is, simply speaking, numeric address of a resource (e.g. web server) on a network that is exclusively used by Internet protocols to route data over the network. A domain name is, simply put, a text alias of a particular IP address. The reason for introducing domain names is to make resources accessible and more user-friendly so that the user is not forced to work with not easy-to-remember IP addresses. With this system, the user has the option (in most cases) to make use of much more familiar domain names. A typical example is web browsing, where the required resource is accessed by its name rather than directly by its IP address.

The problem associated with using DNS is the insecure transmission of its messages – they are simply transmitted as plain text. Anyone who can monitor these messages (for example, an Internet service provider) can also easily find out who is visiting which web pages and when bypassing the fact of visitors using encrypted HTTPS. This situation has led to the development of several technical solutions with the primary purpose of eliminating privacy problems due to the use of unsecured DNS. One of these solutions is *DNS over HTTPS* (DoH) protocol, which encrypts DNS messages through their transmission over HTTPS. However, using DoH can substantially reduce the ability of network administrators to monitor and control network traffic. It is highly desirable to control DoH traffic in the network in some way, which requires the ability to detect that traffic. There exist several possibilities of DoH traffic detection in the literature (e.g. Hjelm (2019)), more or less reliable in a specific situation. However, we have not yet discovered, to our best knowledge, any efficient solution (on either working or experimental level), employing artificial intelligence methods, and taking full advantage of the considerable advances in this field over the last years.

The main contribution of this work is an illustration of an attempt to use artificial intelligence methods, more specifically, machine learning, to create a system that reliably and effectively detects Doh network traffic in practice.

2. DNS

As mentioned above, the main function of DNS is to translate network resource domain names to their equivalent (from the user's point of view) IP addresses. The protocol used for this translation is called the *DNS protocol,* and it simply works on the query-response principle.

An application, typically a web browser, will query a web page via a domain name. The query is first processed by an on-host DNS resolver (a resolver), a program on the client computer that is responsible for initiating and managing the query. The resolver will ask for the DNS server of the local network (further referred to as “local DNS server” in this section), whose IP address is set in the client's network configuration. If the local DNS server does not know the response directly, it queries a set of name servers. A *name server* is basically a directory that stores information, especially IP addresses, about a specific set of domain names. If the server knows the response directly, it will send it back to the interviewer. Otherwise, it will send the interviewer information about where to search - a list of IP addresses of other name servers, if any. After the query processing is finished, the local DNS server sends the final result back to the resolver, which then transmits it to the application. Finally, the application uses the returned IP address to access the queried resource, in the case of a web browser, it displays the content of the desired page. If an IP address was not found for the specified domain name during query processing, an error and its description are returned. The basic concept of DNS operation is described in Mockapetris (1987).

Query processing uses the results of already searched queries that DNS servers keep in their *cache* to reduce the load on DNS servers. Each of these results is associated with a *time to live* (TTL) entry that defines how long the result is usable in response to other queries. TTL is defined by the administrator of the *authoritative DNS server* of the domain and typically ranges from a few seconds to weeks. Some recommendations on how to set up TTL value are described in Barr (1996).

The primary protocol used to transmit DNS queries and responses is UDP on port 53 and in some cases (e.g. too much response data) TCP on the same port.

Details and specification of DNS are described in Mockapetris (1987); the principle is depicted in Figure 1.

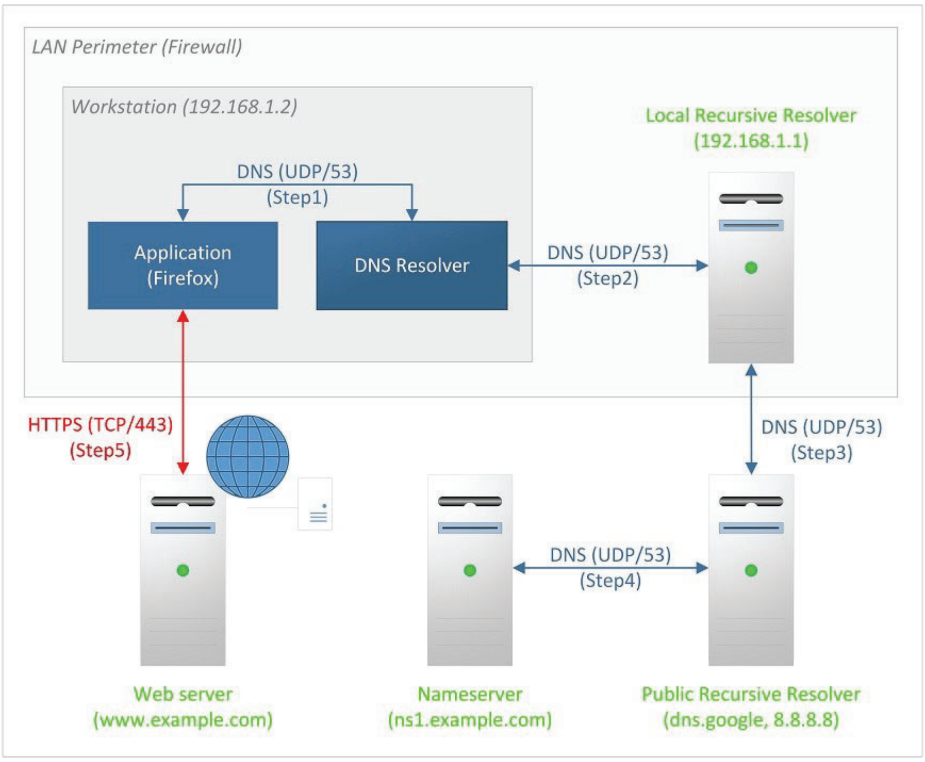

LAN Perimeter (Firewall)
Workstation (192.168.1.2)
DNS (UDP/53) (Step1)
Application (Firefox)
DNS Resolver
DNS (UDP/53) (Step2)
Local Recursive Resolver (192.168.1.1)
HTTPS (TCP/443) (Step5)
DNS (UDP/53) (Step3)
DNS (UDP/53) (Step4)
Web server (www.example.com)
Nameserver (ns1.example.com)
Public Recursive Resolver (dns.google, 8.8.8.8)

Figure 1: DNS system request and response retrieving, a standard solution without encryption

2.1 DNS over HTTPS

DoH solves problems associated with the unsecured transmission of DNS messages by the usage of HTTPS protocol. Unlike traditional DNS querying, the application can, by the usage of HTTPS, directly connect to a public DNS server supporting DoH. This server then processes the DNS request of the application and returns the requested response. Therefore, the application is not forced to go to the DNS server of the local network. The most well-known and widely used implementations of this protocol are now provided by Google and Cloudflare.

The difference in comparison to DNS can be seen in Figure 2. The complete definition of DoH is described in Hoffman et al. (2018).

Key characteristics of DoH:

- Encrypts DNS messages using HTTPS;
- Each DNS query-> response pair is transformed into an HTTPS request-> response pair;
- Applications send queries directly to the public DNS server and bypass the DNS server of the local network, making it difficult for administrators to monitor and control users' web requests;
- There are various implementations; e.g. Google and Cloudflare, which both support wire format (RFC 8484 and RFC 1035), Google also supports JSON;
- Standard defined in RFC 8484;

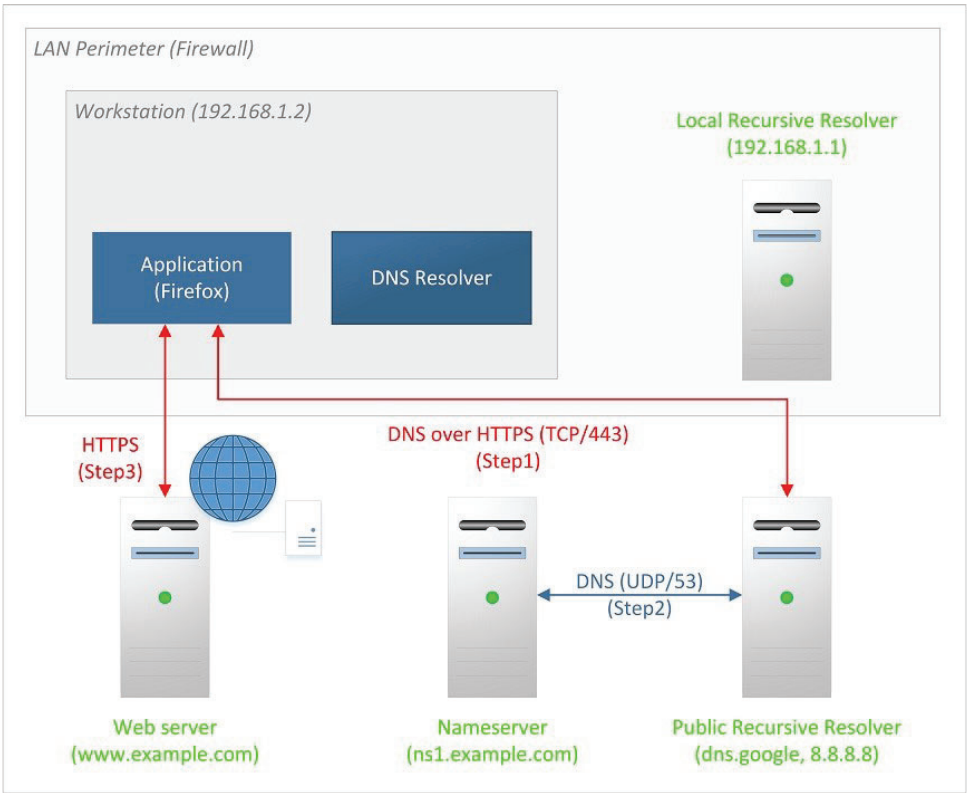


Figure 2: DNS over HTTPS principle, the request is generated by the commonly used web browser

2.2 Transport Layer Security protocol and Server Name Indication

Transport Layer Security (TLS) is a cryptographic protocol used by HTTPS to encrypt communication between a client and a server. Before the actual encryption of the communication, a TLS *handshake* takes place, which is the initial phase during which the server sends a TLS *certificate* to the client. The client then verifies it through the certification authority, and then both parties agree on specific encryption parameters and keys and finally start the encrypted communication.

Server Name Indication (SNI) is a TLS extension that allows users to use TLS in *virtual hosting*. The principle of virtual hosting, in short, makes it possible to share multiple websites on a single server using one IP address, where each of the websites has its hostname (part of a domain name). This allows to save IP addresses and share the resources of the physical server (memory, processor time, etc.). However, since the Internet routing protocols work exclusively with IP addresses, the client needs to specify somehow the website it wants to access. This is done by settings the *Host* field in the HTTP request header, where the client sets the domain name of the destination website. This way works well for HTTP but is unusable for HTTPS that uses TLS internally. It is because the TLS handshake takes place before the server reads the HTTPS request headers so that the client can only access those websites that are covered by the certificate that the server sent to the client during the handshake, which is very impractical. SNI solves this problem by merely specifying the required domain name information directly in the handshake. The server then sends a certificate covering that name. Of course, the client and the server must support SNI. The definition of SNI is described in Eastlake (2011).

Using TLS and SNI alone does not hide information about user-visited domains. The SNI field, as well as the server certificate, are transmitted unencrypted during the TLS handshake. The TLS 1.3 version newly introduces transmission of encrypted server certificate during the TLS handshake. An extension to TLS 1.3, called *Encrypted Server Name Indication* (ESNI), encrypts the SNI field transmitted during the TLS handshake. In order to encrypt the SNI field and then send it to the server during the TLS handshake, the client must have the appropriate encryption key. The client obtains this key by sending the DNS query on the domain name of this server. However, since the DNS communication is unencrypted, using the encrypted SNI field would make no sense. For this reason, DoH is used; the client obtains an encryption key to the given server through a public DoH resolver. More information about ESNI can be found in Rescorla et al. (2020).

2.3 Possibilities of DoH traffic detection

Since this is a relatively new issue, much information (including scientific articles) is currently absent. The most complete source we have found on this topic is Hjelm (2019). He mentions the following detection options:

- TLS inspection;
- Profiling of encrypted connections;
- Checking connections to servers from the list of DoH providers (frequent updates are needed)
- Logging on the level of network application;
- Whitelist of network applications and their proper configuration;
- Statistical processing of metadata (e.g. number of bytes transmitted, connection length, average packet length, etc.)

As Hjelm (2019) mentions, with the advent of TLS 1.3, where certificates are encrypted and therefore not transmitted in plaintext, as in earlier versions, TLS inspection will be made much more difficult. Another extension of TLS 1.3 encrypts the SNI field so that it is no longer possible to “lookout” the domain name from the TLS handshake.

Another possibility is to use artificial intelligence methods. Unfortunately, we could not find any information about a system based on these principles, which would be able to facilitate DoH detection. Siby et al. (2018) describe that they are currently working on a system that will shed some more light on this subject.

3. DoH Traffic Monitoring

This section will evaluate some aspects regarding the monitoring of the DoH sessions and the DoH concept in general.

3.1 Impact of DoH Using and Monitoring by Third-party Solutions

The main motivation for the DoH deployment lies in the network traffic anonymization. On the other hand, DoH anonymizes primary the traffic coming from the DoH client. This fact means that the information about the traffic can be easily stored and utilized by the DoH server provider. Browsers like Mozilla or Chrome use the DoH implicitly now, so the next increase of such traffic type is probable.

3.2 DoH Detection Scenario & proposed Solution

For achieving real results and the possibility of how to test the proposed solution, we have created the following network topology (Figure 3). The basic network items are the L3 switch (called IPFIX), which can report the traffic information via the NETFLOW protocol and a ROUTER containing the advanced firewall based on Mikrotik RouterOS v 6.46.3. The set of various virtual machines (VPC) runs on the dedicated cluster. All the traffic generated by the VPC set is forwarded through the IPFIX switch, which delivers the traffic description headers to the IPFIX collector. Collector stores all records into the processing pipeline (in our implementation Apache Kafka is used) and further analyzed. If some traced or potentially malicious activity is detected, then the manager generates specific rules for its suppression.

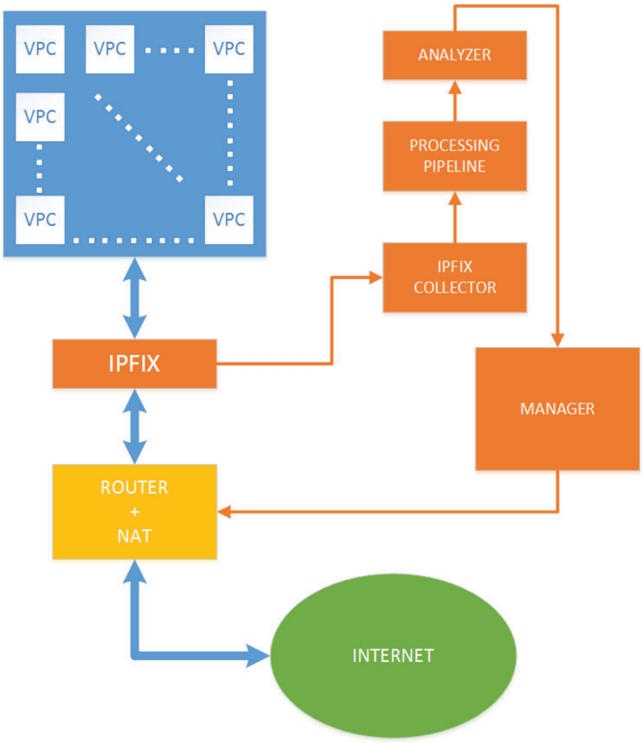


Figure 3: The scheme of the IPFIX data capturing. The traffic was generated in the virtual machines, which were connected to the outside via a NETFLOW reporting switch.

3.3 How to Make DoH Traffic Really Anonymous For Third Parties

The detection and following suppression of the DoH traffic is just only one possibility of how to limit the fact that the world's global traffic information can be used for the private benefit of big companies. The previously meant way can be too strict because the entire DoH traffic can be limited. On the other hand, the detection algorithm need not be perfect for various reasons, and some amount of DoH traffic can be classified wrongly. The following solution may be used to limit the fact that the maintainer of the DoH server can have a clear idea about the traffic information coming from DoH client. The DoH client can add some "noise information" into the DoH requests, and this can rapidly make difficult the possibility of how to predict which domains are visited by it. The DoH client program can contain some bigger (typically tens of thousands of records) data set (DS) with known domain names. Between the real DoH requests, random requests for some domains contained in the DS may be artificially inserted. The insertion of requests can be performed randomly or according to some distribution like Gaussian or Poisson etc.

4. Measurements and Results

The above-meant scenario depicted in figure 3 was used as the primary information source. The model data set (D) contained three basic data flow types. The first type was the DoH, which was generated for the dataset D containing ca. 10 000 domain names. The second was the HTTP/HTTPS traffic, which was created for each record from D, and the third type was all other traffic.

Each reported record of the IPFIX collector contained specific information about packets within a particular time interval. The entire consequence of packets expressed a concrete connection between the DoH server and the client.

Classified transmitted packets in the selected time window were used as the input data for the analysis; the monitored traffic was limited to DoH, HTTPS, and HTTP.

As the first step in DoH data analysis, it was necessary to detect the sessions in the packet stream between each pair of communication participants. The detection was done by comparing the source and destination IP addresses (in both directions) and by comparing the delay between transmitted packets. Currently, for the proof of concept, the limit value of this delay to mark packets as belonging to the same session was set to 1 s. The value was derived from the loading times of typical web pages where 1 s is enough to separate loading requests of different pages.

Each session was described by a vector of values that characterized it. The coordinates of this vector were:

- The total length of the session (the difference between the time of transfer of the first and last packet in the session);
- The number of packets in a given session;
- The total size of all packets in a session given by the sum of their sizes;

Based on the above settings, a total of 12,780 sessions were found that were divided between the training set (10,224 sessions, of which 10% is validation set) and test set (2556 sessions).

The standard multilayer forward network with the layer structure 3 - 4 - 2 - 1 and the Adam stochastic optimization (Kingma, 2015) was used for the classification (several optimization methods were tested, but on our data, the learning process was fastest with Adam optimizer). The model was implemented in Keras / Tensorflow environment. The first three layers used neurons with the ReLU activation function; the last layer uses the sigmoid function. After the first and second dense layers were added Gaussian noise layers with value 0.001. The network learning process is illustrated in figures 4 & 5. We see two jumps in the target value; this testifies to existing local minima in the values of the quality function that have to been overcome.

The achieved accuracy of classification on the test kit was 80.7%. A more detailed analysis revealed that 49.1% of the examples were classified as DoH and 31.6% as other session types. The incorrect categorization of DoH sessions as different session types caused almost the whole classification error.

Using standardized data, the quality of the model improved significantly, and the accuracy of the test kit classification was 94.4%. Correctly as DoH was classified 62.1% of the test set sessions and as other types by 32.3% of the examples. Again, the classification error was caused by incorrect categorization of DoH sessions.

It is evident that in future work, it will be necessary to focus on improving the classification quality and reducing the number of false-negative classifications. One of the possible ways for that may be to modify and expand the network inputs with additional information, as examples of this expansion can be mentioned adding the time relation between sessions or packets, or more detailed information about the structure of the session.

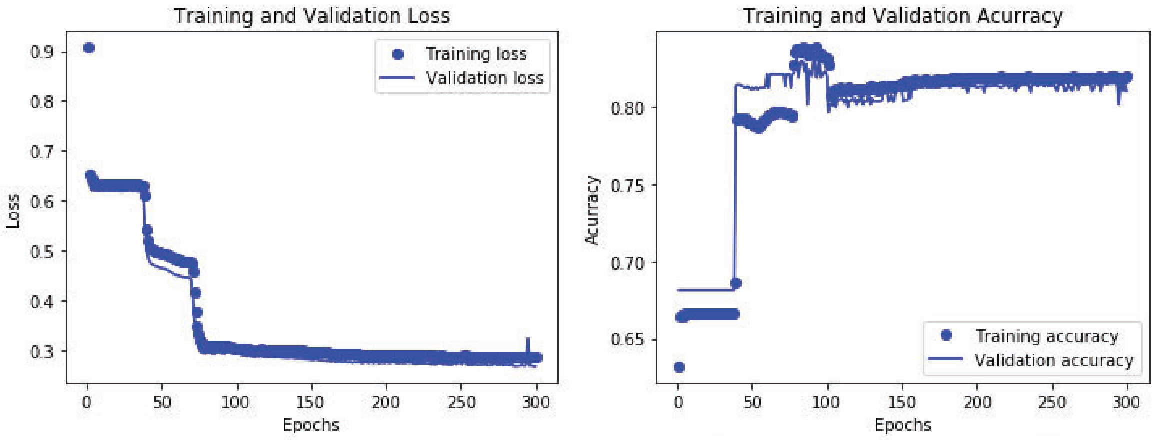


Figure 4*:* Training and validation accuracy and loss for the no -normalized data set; the proposed model can achieve the accuracy of measurement of more than 80%.

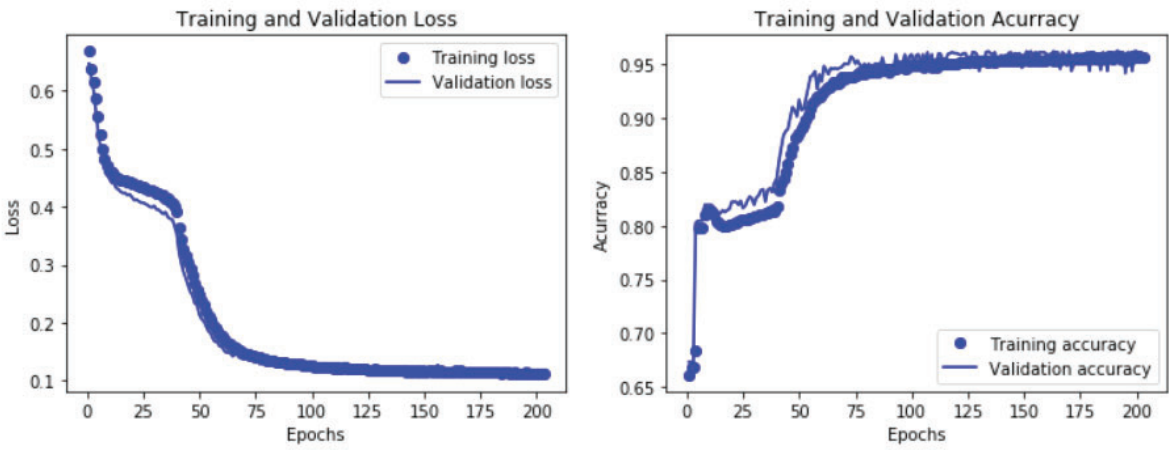


Figure 5: Training and validation accuracy and loss for the clean and normalized data set; the proposed model can achieve the accuracy of measurement of more than 95%.

5. Conclusion

In this paper, we outlined the problem of using Doh and gave a brief summary of the possibilities of DoH detection using different approaches. We proposed and verified a solution based on neural networks. We proposed and verified a solution based on neural networks, the input of which is data obtained from the edge-routers via IPFIX/NETFLOW. The proposed neural network-based detection model achieved the accuracy of prediction of more than 80% for the non-normalized data set and more than 95% for the cleaned and normalized data set. Finally, we briefly outlined the idea of an easy-to-implement protection scheme capable of significantly reducing the possibilities of analyzing network traffic generated by Doh's client applications.

References

Apache Foundation, “Apache Kafka: A distributed streaming platform,” https://kafka.apache.org/

Barr, D. (1996) “Common DNS Operational and Configuration Errors,“ RFC 1912, https://tools.ietf.org/html/rfc1912

Borgolte, K. et al. (2019) “How DNS over HTTPS is Reshaping Privacy, Performance, and Policy in the Internet Ecosystem,” *SSRN Electronic Journal*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3427563

Eastlake, D. (2011) “Transport Layer Security (TLS) Extensions: Extension Definitions, “ RFC 6066, https://tools.ietf.org/html/rfc6066

Hjelm, D. (2019) “A new Needle and Haystack: Detecting DNS over HTTPS Usage,“ SANS Institute, Information Security Reading Room, August.

Hoffman, P., McMannus, P. (2018) “DNS Queries over HTTPS (DoH),“ RFC 8484, https://tools.ietf.org/html/rfc8484

Kingma, Diederik P., Ba, J. (2015) “Adam: A Method for Stochastic Optimization,” Paper published at the 3rd International Conference on Learning Representations, San Diego, https://arxiv.org/abs/1412.6980

Mockapetris, P. (1987) “Domain Names – Concepts and Facilities,” RFC 1034, https://tools.ietf.org/html/rfc1034

Mockapetris, P. (1987) “Domain Names – Implementation and Specification,” RFC 1035, https://tools.ietf.org/html/rfc1035

Rescorla, E. et al. (2020) “Encrypted Server Name Indication for TLS 1.3,“ Internet-Draft, https://tools.ietf.org/html/draft-ietf-tls-esni-06

Siby, S. et al. (2018) “DNS Privacy not so private: the traffic analysis perspective,“ In: The 11th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2018), 27 July 2018, Barcelona, Spain.

Van Heugten, J.H.C (2018) “Privacy analysis of DNS resolver solutions,” Master’s thesis, Master of System and Network Engineering, University of Amsterdam, The Netherlands.

Towards Cyber Sensing: Venturing Beyond Traditional Security Events

Artūrs Lavrenovs, Kimmo Heinäaro and Erwin Orye
NATO CCDCOE, Tallinn, Estonia
Arturs.Lavrenovs@ccdcoe.org
Kimmo.Heinaaro@ccdcoe.org
Erwin.Orye@ccdcoe.org

DOI: 10.34190/EWS.20.062

Abstract: Host and network-based events are the backbones of any modern IT monitoring and detection system. The number of lower priority security events is significant and might contain weak indicators of cyber attacks; by combining host and network events with sensor data that are not part of conventional IT security, we are able to elevate otherwise missed events to discover hidden cyber attacks. The sensor data is fed into a situational awareness system which augments traditional alerts. This technique is primarily applicable for critical infrastructure, military, government and large organisations where the adversary is sophisticated enough to bypass existing detection methods. We discuss operational and strategic implications by using this type of sensor. We have implemented these principles in two scenarios tested in cyber exercises. In the first proof of concept we focused on sensor fusion by integrating existing non-IT sensor systems with IT security and correlated the collected data. This enabled the Blue Team to detect well-hidden Red Team attacks against a simulated power grid and counteract them. In the second, we explored a large variety of sensors monitoring individual personnel and their operating environment. Sensors used in this research are categorised into biological, environmental and EM spectrum. Biological sensor data includes heart rate, stress level and brain wave monitoring. Environmental sensors monitor the RF spectrum, CO_2 level, VOC level, temperature, humidity, infrared, ultraviolet, visible light, noise level, proximity and vibration.

Keywords: host-based events, network-based events, IDS, security events, sensors

1. Introduction

Critical infrastructure, government and military networks are under ever-increasing threat of cyber attack. The defence solutions market is experiencing growth and vendors are constantly developing new and more advanced solutions that use cutting-edge approaches like Artificial Intelligence (AI). But most of these solutions rely on traditional sources of data - host and network-based events. As adversaries in this scenario are up to state-level actors, they have the resources to investigate, adapt to and overcome new advanced defences. This calls for a widening of our view and exploring additional sources of data.

It does not necessarily equate to acquiring another new and expensive solution containing both hardware and software components, but rather evaluating what sensor data is already available. Sensors can be viewed from system-centric and human-centric perspectives. The former primarily focuses on the states of the systems, inferring human properties indirectly when possible, while the latter addresses biological data that can be measured directly via dedicated sensors. We refer to the combining of traditional and other sensor data for the purpose of detecting cyber attacks as cyber sensing.

In Section 2 we review existing research addressing human behaviour analysis and training models. In Section 3 we explore data sources acquirable from different sensors. In Section 4 we describe a proof of concept correlating traditional security events with building automation and gathering data from environmental sensors. Section 5 discusses operational and strategic implications of cyber sensing, and Section 6 presents our conclusions.

2. Related work

There are two related streams of work in this field. The first is from a human-centric perspective and the second concerns how computer models can be used to predict human behaviour.

Sensor fusion mimics the human brain combining multiple senses. Lin et al. (2004) propose a neural network architecture to integrate data from several physiological and behavioural sensors to improve reliability and resistance to impersonation (multimodal verification system).

Augmenting cyber-physical systems with sensors monitoring humans results in cyber-physical-human systems (CPHS) (Sowe et al., 2016). CPHS is not investigating defensive or offensive applications researched by us but the safety of humans (Gelenbe et al., 2012) and security of the sensors and produced data.

2.1 Human behaviour analysis

The analysis comes in three stages: human behaviour measurement, modelling and analysis (Carus et al., 2014). In straightforward research, a relatively small number of people are observed for a long time using many sensors. Typical of this research is people living in smart homes, where we build up personal data that can be classified under 'normal behaviour' and from that point, we find a threshold where deviating behaviour can be identified using supervised or unsupervised machine learning. Examples of this are:

- A multi-view setup of cameras used to recognise people's behaviour based on human action recognition with multi-view scenarios and real-time execution (Chaaraoui et al., 2014).
- Analysing relationships between behavioural patterns and energy consumption by correlating home-based activities with electricity use (Chen, Cook and Crandall, 2013).
- Analysis of the behaviour of elderly people in their homes by detecting patterns in sensor data in an unsupervised manner (Rieping, Englebienne and Kröse, 2014).

Another popular research topic is the abundance of data created by mobile phones. In modern cell phones, there are often accelerometers, gyroscopes, magnetometer, GPS, barometers, proximity sensors and an ambient light sensor. An example of an academic analysis of human behaviour using location knowledge obtained from data sent by heterogeneous moving objects can be found in Yus et al. (2014).

Closing in on the effect of human behaviour on cybersecurity, most of the related work focusses on the protection of one's own IT-systems. An example is studying the human factors aspect of cyber defence. Gutzwiller et al. (2015) focus on education, training and situational awareness to enable people to pick up cyber attacks that have been missed by automated tools.

Modelling human behaviour to anticipate insider attacks via system dynamics (Greitzer and Hohimer, 2016) is one of the few papers that looks at cyber effects from the perspective of an insider threat, trying to predict the probability by using a diverse set of data sources from the cyber domain and inferred psychological and motivational factors of a malicious insider activity.

2.2 Training models

Computer-aided learning methods depend on the kind of data that is available. If data for both inputs and outputs is available, supervised learning is often a good approach. If only input data is available, unsupervised learning or artificial intelligence is usually used. In cases where no direct access to modify the output is possible, but some measurement of the quality of an output that follows an input is, then reinforcement learning can be used. In the realm of human behaviour, it is very difficult to predict the outcome with a known set of inputs. Therefore, most research currently focusses on unsupervised learning.

Bass (1999) described how detection systems will fuse data from heterogeneous distributed network sensors to create cyberspace situational awareness. He says 'situational awareness technology is alarmingly primitive relative to protecting our critical electronic infrastructures'.

An analysis of unsupervised activity discovery and primitive sequence information in this field can be found in Seiter et al. (2014). Feature representation of sensor data that does not rely on prior expert knowledge of human activity and using unlabelled data for activity recognisers is examined in Bhattacharya et al. (2014). A behavioural activity recognition model for mini-wearable devices is used to recognise activities of daily living in Hu et al. (2014), human activity recognition from video images in Jalal et al. (2017) and training a neural network to predict whether video frames and audio are temporally aligned to predict sound source localisation and audio-visual action recognition in Owens and Efros (2018).

3. Exploring data sources

This paper is about exploring new ways to detect cyber incidents. A wide variety of sensors is used, without any assumptions about whether they can be related to the subject or not. Data collected from the sensors is correlatable with knowledge and the timeline of Red Team activities in a cyber exercise. It might be possible to detect cyber incidents from improbable sources. Further studies are needed to check which sensors could be used for early warning of an attack and which are merely an indication of an incident already discovered.

Data sources used can be divided into system- and human-centric. System-centric data sources are based on the integration of existing security and automation systems and are further studied in Section 4. Human-centric data sources are based on individual sensors that are described here. These can be divided into three classes: environmental, electromagnetic spectrum and biometric.

3.1 Environmental sensors

The environment of the defending Blue Teams could be monitored with a wide variety of different sensors. The positioning of environmental sensors is critical for accurate measurements. In this study, no assumptions of values have been made and sensors were situated at random locations near the Blue Teams. The following environmental data sources were monitored: vibrations (acoustic noise level and acceleration); atmosphere (air quality, temperature, humidity and atmospheric pressure); and computer use (current consumption).

Acoustic noise level can be measured with a microphone, digitized with an AD converter and averaged over time. Acceleration can be measured using integrated circuits containing miniaturised mechanical structures (microelectromechanical systems, MEMS). An accelerometer can detect movement, vibration and shock. (Scanaill et al., 2013)

Air quality was monitored by measuring CO_2 level, volatile organic compound (VOC) level and alcohol level. Gas content in the air can be measured with miniature semiconductor sensors that contain a heated thick-film metal oxide layer on a base that contains electrodes for measuring the resistance of the circuit. Composition and structure of the metal oxide layer are adapted to the target gas (Scanaill et al., 2013). CO_2 is produced by human respiration and the amount increases with physical activity (Fucic et al., 2012). VOCs are carbon-based chemicals that are considered as indoor air pollutants. They are typically emitted from building materials and furniture (Scanaill et al., 2013). Alcohol level is normally measured as breath-alcohol content from a single person. Temperature can be monitored by measuring voltage over a semiconductor p-n junction under constant current (Scanaill et al., 2013). Relative humidity can be determined by measuring the capacitance of material exposed to air (Sazonov, 2014). Atmospheric pressure can be measured with a piezoresistive MEMS device, where pressure bends material causing changes in resistance (Scanaill et al., 2013).

Current consumption of mains voltage powered devices (e.g. workstations and servers) can be measured with Hall effect sensors. The magnetic field produced by the current in a wire is measured by measuring voltage over a semiconductor under constant current, and the mains voltage is isolated from the measuring circuit (Scanaill et al., 2013).

3.2 Electromagnetic spectrum sensors

In this study, the aim was to monitor as wide a range of the electromagnetic spectrum as possible. With simple sensors, the following frequency ranges were covered: ultraviolet, visible and infrared light. Radio frequencies were monitored in the range of 0-6 GHz.

Ultraviolet band radiation is produced by the sun or special lamps and can be detected with a photodiode that converts light into electricity with a semiconductor p-n junction. The visible light level can be monitored with a Light Dependent Resistor (LDR), which is a semiconductor that changes conductivity according to the light level (Scanaill et al., 2013). With simple sensors, it is only possible to detect the average level of the band measured. Infrared band radiation is emitted by objects near room temperature. The IR sensor used is a proximity sensor that senses movement in the vicinity of the sensor.

The RF spectrum can be monitored with a Software Defined Radio (SDR). If the frequency range to be monitored is wider than the bandwidth of the SDR, spectrum scanning can be used. Scanning rate (Hz/s) is limited by the hardware and will leave gaps in the time axis for real-time spectrum monitoring. With scanning, it is possible to log all signals detected above a certain threshold amplitude.

3.3 Biometric sensors

Monitoring biometric values of a person in the Blue Team can reveal stress level or changes in behaviour. It can be used to detect cyber incidents as they start to cause problems. In this study, the following biometric sensors were used: heart rate sensor, cortisol level sensor and brainwave sensor.

Heart rate monitoring can be used in sports training or measuring health status and mental stress. Optical wearable sensors (photoplethysmography, PPG) can be made with an LED light source, photodetector and using a signal processing algorithm (Sazonov, 2014). They are available in the form of a wristwatch and are easy to wear in daily activities. Cortisol is a stress hormone, along with adrenaline and noradrenaline. It helps the body to function in stressful situations but is also harmful to health if the stressful situation lasts for a long period (Bonetti, 2014). Brain activity can be monitored with electrodes attached to the scalp (electroencephalography, EEG). The differential signal between electrodes is amplified and filtered to produce waveforms that can also be visualised. Signals can be categorised as gamma (indicating learning), beta (logical thinking), alpha (deep relaxation), theta (sleep) and delta (deep sleep) based on their frequency range (0-100 Hz) (Le et al., 2018).

3.4 Further development

The sensors selected for this study were mostly modern, semiconductor-based, inexpensive and widely available and most were factory calibrated with digital output. Most of the values measured possibly cannot be used to detect cyber incidents, even after further studies. The secondary target of the study was to create a universal sensor platform that can measure and log as wide range of physical values as possible. Measurements that might be useful in detecting cyber incidents, but are not part of this study, include adrenaline level, human body temperature, blood pressure, the consumption rate of beans in the coffee machine and thermal camera imaging of the Blue Team. Collecting and combining data from biometric sensors on all Blue Team members might also reveal new correlations with cyber incidents.

4. Cyber sensing proof of concept

We have implemented cyber sensing of explored data sources into a proof of concept from a system-centric and human-centric perspective. These two perspectives differ significantly therefore we independently developed prototypes for the cyber exercises Crossed Swords and Locked Shields.

4.1 System-centric cyber sensing

System-centric cyber sensing can use pre-existing systems and sensors, thus not necessarily requiring additional hardware components. Every modern building has significant automation controlling the heating, ventilation and air conditioning (HVAC), energy efficiency (e.g. smart lights), safety and security systems. At best, only the building's security system might be integrated with the organisation's cyber defence systems. Usually these sensors can only infer human presence and activity indirectly, commonly by detecting movement, IR radiation or temperature changes.

Critical infrastructure protection has been one of the main components of scenarios in cyber defence exercises for the last few years. Power grids and generation have become classical components of both defence and attack scenarios. And rightfully so, as attacks against power grids have already been seen in the real world and proven highly effective (E-ISAC, 2016).

For the proof of concept, we used a simple building automation system consisting of access control, smart lights, fire alarm and security alarm. This is representative of what is already commonly available without any additional hardware or software changes in existing building automation systems as long as the status of the sensors can be acquired. Building automation systems typically use measured values for controlling factors such as temperature, and security systems typically use only one-bit information sources (e.g. door opened or closed). In this study, we used the built-in web server of a PLC to expose four one-bit values from the security control of a single room to the IT situational awareness system: security alarm armed, security alarm triggered, safety alarm triggered, lights on. The security systems included CCTV cameras that are accessible from the network, either directly or via a recording device. Figure 1 presents the developed prototype including operator of the engineering workstation and building automation system.

The defended system was a Windows OS computer running SCADA software controlling physical RTU and circuit breakers representing a power grid engineering workstation. In the scenario cyber sensing is defending against engineering workstation compromise that would provide the attacker with initial access which is technique T818 according to the MITRE ATT&CK for ICS Matrix (MITRE Corporation, 2020). The circuit breaker is supplying power to the defender's cyber command and when opened the power is physically cut bringing the defender's network down. The workstation is feeding all OS security events into the situational awareness system.

Correlation with the traditional network and host-based events is crucial to the system-centric approach. We integrated these event streams into the exercise situational awareness system Frankenstack (Kont et al., 2017). Low priority events like authentication that otherwise would be lost in the stream of data can now trigger high priority alerts when nobody is actually at the workstation. When these high priority alerts are raised, the adversary has an option to respond by rebooting or terminating potential pivot points that were used to access the engineering workstation, thus countering the Red Team.

The four input bits from the building automation allowed us to define five rules relevant to the scenario and the real world. The monitored trigger is a standard Windows logon security event, which is a low priority event thus not causing any alert. The five implemented scenarios when the alert is raised when the logon occurs:

- The security alarm is triggered: attacker inside the armed location is attempting to access the system.
- The security alarm is armed: attacker has either bypassed physical security system or is accessing the workstation remotely.
- The state of the building automation not updated recently: potential compromise or failure of sensor integration.
- Safety alarm triggered evacuation: attacker might be exploiting chaos caused by the evacuation.
- Smart lights are off: attacker might be accessing the workstation remotely.

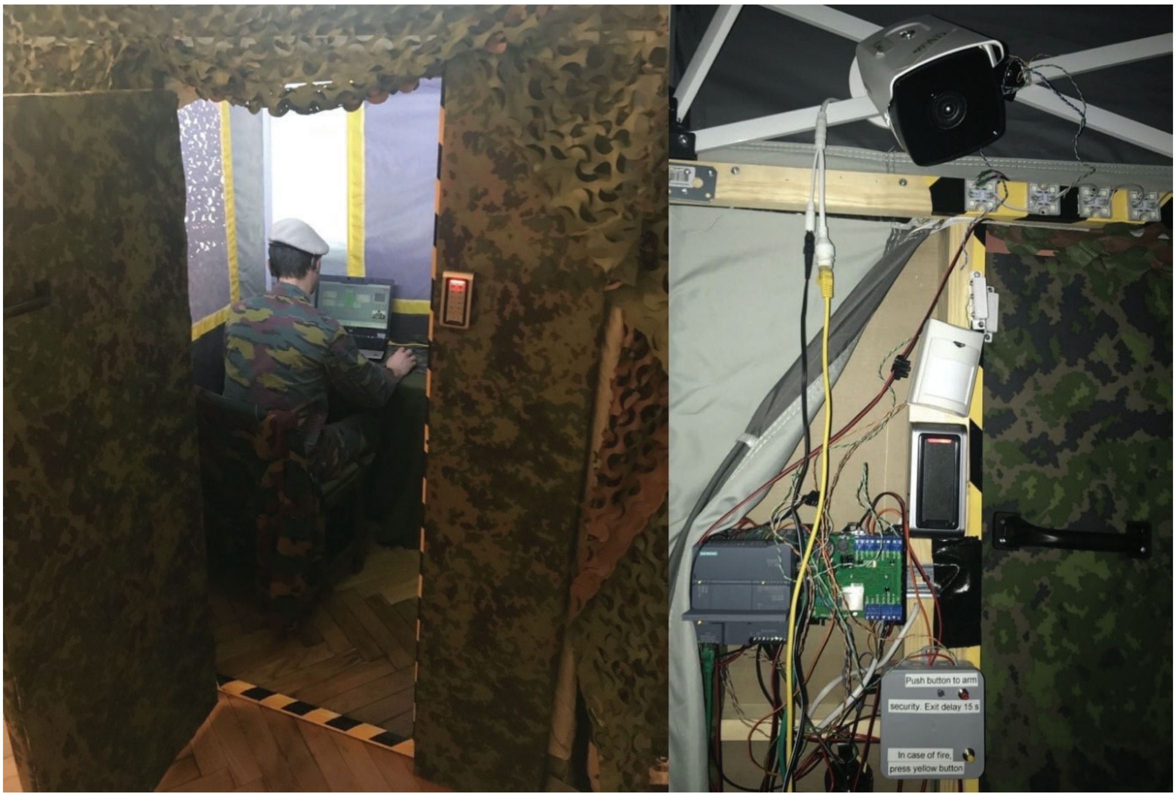


Figure 1: System-centric cyber sensing proof of concept

Initially, the Red Team has no knowledge of any non-traditional cyber defence systems being used. Through acquiring documentation or triggering the alert and therefore the defender's response, it becomes clear that the attack has to be specially crafted. Since the defender's actions are manual, an option is to craft a payload that executes so quickly that it will cut the power off, no matter if it is detected.

A stealthier option is to acquire additional information through compromising the building's CCTV, allowing the attacker to identify when the operator is in the control room but is not monitoring the displays. In this case, an alert will not be raised and it demonstrates the main weakness of the system-centric approach. It is impossible to get precise information about the operator; only their presence is inferred, but it is not known what they are doing. The operator may be monitoring the displays, taking actions that are directly related to the input of the screens or being distracted and paying attention to their personal phone. The third option is to attack the sensors and the building automation itself to disable the sensor readings, change them to hide the attack or trigger an avalanche of false alerts.

System-centric cyber defence has been demonstrated to be simple, cheap and effective as long as the attackers do not know it is being used. Security rules can be reduced to one or more sensor status bit correlations with one or more host or network-based events using a sliding time window. A set of static rules defining specific defence scenarios can suffice without the need for AI.

4.2 Human-centric cyber sensing

Human-centric cyber sensing focuses on measuring the biological properties of humans directly or through the properties of the environment around them. The volume of generated data is significantly higher than in the system-centric approach as not only might many sensors be necessary, but these sensors will have a wide range of values instead of a single bit and the sampling rate can be high. In this case, creating rules manually becomes challenging and might require AI. If the human-centric approach evolves, it can be expected that there will be a limited number of widespread universal solutions containing sets of different sensors in a single unit. Therefore, we developed one such integrated unit to collect data in various exercises and simulated scenarios.
Environmental values measured change relatively slowly, but biometric sensors and RF spectrum produce a lot of data. Commercial high data rate sensors were deployed to collect data with timestamps (Figure 2).

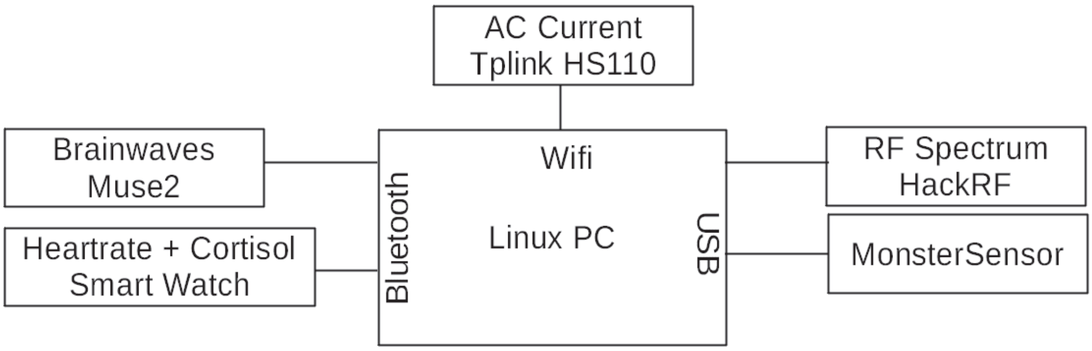


Figure 2: Integration of all human centric sensors

For slow-varying environmental data, we created a multi-sensor logging device called MonsterSensor (Figures 3 and 4). It is based on an Arduino computing unit and low-cost sensor modules which are integrated with an RTC and memory card module for timestamps and logging. The device can either be used autonomously to log values on the memory card or could be connected to a PC for real-time monitoring.

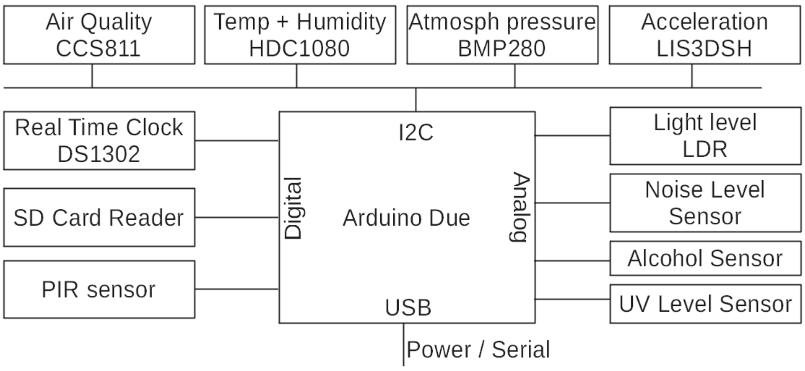


Figure 3: Multi-sensor logger (MonsterSensor) architecture

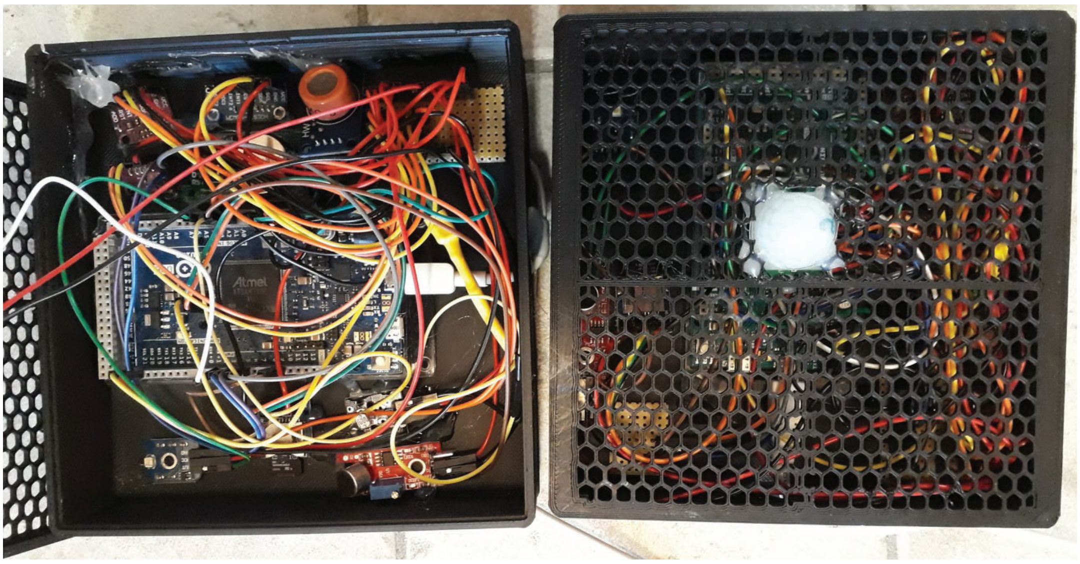


Figure 4: MonsterSensor prototype construction and housing

5. Operational and strategic implications

It is compelling to have out-of-band information, because there is almost always a person involved with cybersecurity, even when a lot of the processes are automated. Having this information adds to operational advantages by identifying and limiting risks or increasing opportunities for offensive cyber actions. The people who will be 'sensed' are mostly employees, but can be end-users, state officials, management or intruders. The monitoring of human behaviour adds to the knowledge we currently have on the effect that individuals have on cybersecurity.

Some infrastructures or smart buildings already have a lot of system-centric sensors installed. Centralising and processing all the data that is already produced by those sensors will provide insights into the correlation between human behaviour and cybersecurity.

Human-centric sensors such as heart rate monitors, accelerometers and pedometers are used widely nowadays. Smartwatches are built to collect, store and process data. However, the information that these devices produce is private and not legally accessible for analysis. Installing human-centric sensors or getting access to data of personally owned devices will have a deep impact on privacy and ethics.

The questions that need an answer, and where cyber sensing can be a breakthrough:

- Is there someone in the room?
- Are they authorised to be there?
- Is the person paying attention to their work?
- Is the person taking action if an event occurs? When the event happens, are there correlations to inputs to the system?
- Are there behavioural differences among people that are correlated with cybersecurity?

One should at least be cautious, because this information could also be used to target individuals rather than optimising a process.

5.1 Defensive perspective

The most straightforward use of this information would be from a defensive perspective. It is clear that looking at sensors that are under one's own control is easy to achieve. Achieving correlations between successful cyber attacks and behaviour of people will increase awareness and lead to improved cybersecurity. This could be achieved by better procedures, leadership, technology and knowledge development.

With enough data and good learning algorithms, it is possible to analyse human behaviour and find weak spots in a system or individual. Since in cybersecurity the attacker only has to find one weak spot and the defender has to protect every part of the system, this knowledge will help in establishing priorities to mitigate the threats.

Redeveloping Section 4.1 power grid scenario in a human-centric sensing requires addressing wide range of open questions:

- What would be the sensors and the setup?
- Where would data go? How it would it be transmitted?
- What would be the simple rules and what would be the AI?
- What happens if adversaries target sensors? What if they target data? What if they target AI?
- What if one targets private data from wearables? What if one feeds data into their own AI and find terminally ill employees even before they know it?

5.2 Offensive perspective

Target monitoring, although not the intended use, can be accomplished if patterns of human activity are known when the attacker is conducting cyber attacks. Using this information it makes it possible to attack the target when human defences are at their lowest.

If sensors are implemented from a defensive perspective and generate data when certain sensor output values are reached, it is beneficial to analyse the attack surface generated by those sensors. If an adversary knows what actions there will be if a certain sensor or group of sensors reaches a certain value, these sensors become targets themselves. This means that using the information from those sensors will require additional security measures.

State actors, private companies and criminals may, for different reasons, be interested in this kind of information. States that infiltrate the 'sensor system' below the threshold of detection and use AI on real-time data will learn a great deal and can eventually also cause a lot of damage by using this information.

5.3 Insider threats

An insider threat, also mentioned in *Modelling Human Behaviour to Anticipate Insider Attacks* (Greitzer and Hohimer 2016) is a very difficult threat to mitigate. Analysing human behaviour with the use of sensors will increase the probability of detecting this behaviour, although it will probably be necessary to combine other sources of information.

6. Conclusions

Host- and network-based events drive traditional monitoring and detection systems. The amount of generated log data is tremendous and manually defined rules are reaching the limits of what is possible to define. To address this, researchers and solution vendors are applying AI. In this research, we explore what could be the next step after the currently used defence systems augmented with AI reach their limits.

The next logical step would be to introduce additional data sources that are relevant in the security context. We have presented our exploration of sensors that could be used as these sources. We have defined two distinct approaches; system-centric and human-centric sensing. We covered each of the approaches by developing a proof of concept using it.

System-centric cyber sensing has already been shown to be feasible today without additional hardware, using sensors that have been installed as part of different systems, primarily building automation. The security rules can still be defined manually as many relevant sensor statuses can be reduced to a binary value. In this approach human presence is inferred, meaning activities of humans cannot be measured directly, which is a major drawback. When a sophisticated adversary is aware that such a defence exists and of its relevant rules, they can craft scenarios by-passing the rules. We have implemented both the defence and by-passing scenario for the simulated power grid.

Human-centric sensing has proved to be a much more complex topic. The possibilities for collecting data are endless. We expect that these multi-sensor units, covering a wide range of inputs, will gain traction. We developed this type of universal multi-sensor prototype and are gathering data in cyber exercises as it is clear that this approach requires AI. Sensors generate too much data to be reduced to simple manual rules handling a few binary values.

Cyber sensing is highly promising and could very likely be the next leap in cyber defence advancement. We predict that cyber sensing will follow the same pattern of development as current IT defence systems. Solutions will start from simple system-centric sensing, and as they develop and reach their limits, it will start to transition into human-centric sensing relying on AI.

References

Bass, T. (1999) Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems, Iris National Symposium.

Bhattacharya, S. Nurmi, P. Hammerla, N. and Plötz, T. (2014) Using unlabeled data in a sparse-coding framework for human activity recognition, *Pervasive and Mobile Computing.*

Bonetti, B. (2014) *How to Stress Less: Simple ways to stop worrying and take control of your future*, Capstone.

Candás, J.L. Peláez, V. López, G. Ángel, M. Álvarez, F.E. and Díaz, G. (2014) An automatic data mining method to detect abnormal human behaviour using physical activity measurements*, Pervasive and Mobile Computing.*

Chaaraoui, A.A. Padilla-López, J.R. Ferrández-Pastor F.J. Nieto-Hidalgo M., and Flórez-Revuelta F. (2014) *A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context*, Sensors, Switzerland.

Chen, C. Cook, D.J. and Crandall, A.S (2013) The user side of sustainability: Modeling behavior and energy usage in the home, *Pervasive and Mobile Computing.*

Fucic, A. Jalali, S. and Pacheco-Torgal F. (2012) *Toxicity of Building Materials*, Woodhead Publishing.

Gelenbe E., Gorbil G. and Wu F. (2012) Emergency Cyber-Physical-Human Systems, *21st International Conference on Computer Communications and Networks*, Munich.

Greitzer, F.L. and Hohimer, R.E. (2011) Modeling Human Behavior to Anticipate Insider Attacks*, Journal of Strategic Security.*

Gutzwiller, R.S Fugate, S. Sawyer, B.D and Hancock P.A. (2015) The human factors of cyber network defense, *Proceedings of the Human Factors and Ergonomics Society.*

Higson, S. (2012) *Biosensors for Medical Applications*, Woodhead Publishing.

Hu, L Chen, Y Wang, S. and Chen, Z. (2014) b- COELM: A fast, lightweight and accurate activity recognition model for mini-wearable devices. *Pervasive and Mobile Computing.*

Kont, M., Pihelgas, M., Maennel, K., Blumbergs, B. and Lepik, T. (2017) Frankenstack: Toward real-time Red Team feedback, IEEE Military Communications Conference (MILCOM), Baltimore, MD, pp. 400–405.

Le, D. N. Van Le, C. Tromp, J. G. and Nguyen, G.N. (2018) *Emerging Technologies for Health and Medicine*, Wiley-Scrivener.

Lee, R.M Assante, M.J. Conway, T. (2016) Analysis of the Cyber Attack on the Ukrainian Power Grid white paper e-isac, https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf, accessed on 16 Jan 2020.

Lin S.H., Mak M.W. and Kung S.Y. (2004) *Biometric Authentication: A Machine Learning Approach*, Prentice Hall.

MITRE Corporation (2020) Engineering Workstation Compromise, https://collaborate.mitre.org/attackics/index.php/Technique/T818, accessed on 16 Jan 2020.

Owens, A. and Efros, A.A. (2018) Audio-visual scene analysis with self-supervised multisensory features. *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*).

Rieping, K. Englebienne, G. and Ben Kröse (2014), Behavior analysis of elderly using topic models, *Pervasive and Mobile Computing.*

Sazonov, E. (2014) *Wearable Sensors*, Academic Press.

Scanaill, C.N. and McGrath, M.J. (2013) *Sensor Technologies: Healthcare, Wellness and Environmental Applications,* Apress.

Seiter, J. Amft, O. Rossi, M. and Tröster, G. (2014) Discovery of activity composites using topic models: An analysis of unsupervised methods, *Pervasive and Mobile Computing.*

Sowe, S. K., Zettsu, K., Simmon, E., de Vaulx, F., and Bojanova, I. (2016). Cyber-Physical Human Systems: Putting People in the Loop. *IT professional*.

Yus, R. Mena, E. Ilarri, S. and Illarramendi A. (2014) Sherlock: Semantic management of location-based services in wireless environments, *Pervasive and Mobile Computing*.

Cyber Security Capacity Building: Cyber Security Education in Finnish Universities

Martti Lehto
University of Jyväskylä, Finland
martti.lehto@jyu.fi

DOI: 10.34190/EWS.20.112

Abstract: Cyber security in Finland is part of other areas of comprehensive security, as digital solutions multiply in society and technologies advance. Society may be developed through examining different complex issues and services. Even today, a broad range of different actors in society participates in cyber security work, both from central government and the private sector. Cyber security is one of the primary national security and national defense concerns. Cyber security has quickly evolved from a technical discipline to a strategic concept. Since cyber security has evolved from a technical discipline to a strategic concept, and because cyber-attacks can affect national security at the strategic level, cyber security must be looked at beyond the tactical arena. Cyber self-sufficiency is seen as a challenge that requires improved higher education, more research and increased support from companies to solve. Every country needs well trained professionals working in cybersecurity roles. These professionals are critical in both private industry and the government for the security of individuals and the nation. Cyber security capacity building can be measured based on the existence and number of research and developments, education and training programs, and certified professionals and public sector agencies. By all accounts, the world faces a current and growing workforce shortage of qualified cybersecurity professionals and practitioners. A shortage in the global cybersecurity workforce continues to be a problem for companies in all industries and of all sizes. (ISC)² (2019) estimates that together in 10 major economies the workforce total is 2.8 million and the global gap estimates are 4.07 million, so the global workforce needs to grow by 145%. Pursuant to the Finland's Cyber Security Strategy (2019) "National cyber security competence will be ensured by identifying requirements and strengthening education and research." This paper analyzes the know-how demands and needs of the cyber security (world?) which the different national and international actors of the cyber world present. The know-how demands and needs are compared with the cyber security education which is offered in Finland's universities. The paper is based on a survey conducted on Finnish universities and universities of applied sciences in 2018 and 2019 and on the analysis of its results.

Keywords: cyber education, cyber competence, cyber strategy, security

1. Introduction

The picture of war has become more complex. In order to achieve political objectives, political, economic and military pressure, forms of information and cyber warfare, combinations of all of the above and other forms of hybrid influencing, among other things, are used in a coordinated fashion on top of the constantly developing military means. The capabilities required by cyber security will be strengthened, among others, in the European Union, with NATO, and bilaterally. In its international cooperation Finland aims at seizing some of the lucrative prospects offered by the cyber domain and digitalization. (PMO, 2016)

Finland strengthens its national defence and intensifies its international defence cooperation. Land, maritime and air defence, as well as joint capabilities, will be developed in line with the requirements of the operating environment. New capabilities will be created for the cyber domain. The cyber domain is becoming increasingly important. The use of cyber operations for the purpose of pursuing political objectives cannot be excluded. The digitalization of society, the dependency of technical systems on trans-border IT networks as well as the interdependencies and vulnerabilities of systems expose society's vital functions to cyber-attacks. (PMO, 2017)

The CSEC2017 JTF advances cybersecurity as a new computing discipline and positions the cybersecurity curricular guidance within the context of the current set of defined computing disciplines. The CSEC2017 Joint Task Force (JTF) defines cybersecurity as: "A computing-based discipline involving technology, people, information, and processes to enable assured operations in the context of adversaries. It involves the creation, operation, analysis, and testing of secure computer systems. It is an interdisciplinary course of study, including aspects of law, policy, human factors, ethics, and risk management. Cybersecurity is a computing-based discipline involving technology, people, information, and processes to enable assured operations in the context of adversaries. It draws from the foundational fields of information security and information assurance; and began with more narrowly focused field of computer security." (CSEC, 2017)

This paper presents cyber education at the universities where cyber can be studied in Bachelor's and Master's degrees, major or minor studies.

2. The Finnish education system

The Finnish education system is a mixture of state controlled or steered and relatively autonomous elements. The government determines the general objectives of education and the division of classroom

hours between different subjects. The Ministry of Education and Culture drafts legislation and government decisions pertaining to education. The Finnish National Agency for Education lays out the concrete objectives and core contents of instruction in the different subjects and is responsible for the national core curriculum with its directive norms for good achievement in each. Local authorities (generally municipalities) are responsible for the practical arrangement of schooling and for composing the municipal curriculum based on the national core curriculum. Each school, in turn, writes its own curriculum based on both the national core curriculum and the municipal document. (Kupiainen et.al, 2009)

The Finnish education system consists of:

- Early childhood education and care which is provided for children before the compulsory education begins,
- Pre-primary education which is provided for children in the year preceding the beginning of compulsory education,
- Nine-year basic education (comprehensive school), which is compulsory,
- Upper secondary education, which is either general upper secondary education or vocational education and training, and
- Higher education provided by universities and universities of applied sciences.
- Adult education is available at all levels.

2.1 Higher education system in Finland

The Finnish higher education system comprises universities and universities of applied sciences. The universities engage both in education and research and have the right to award doctorates. The universities of applied sciences are multi-field institutions of professional higher education.

There are 14 universities in the Ministry of Education and Culture sector. University-level education is also provided by the National Defence University, which is part of the Defence Forces. Altogether there were nearly 153 400 university students in 2018. From the beginning of 2010 universities have had the status of independent legal entities and have been separated from the state. However, the state continues to be the primary financier of the universities. Direct government funding covers about 64% of university budgets. (MEC, 2020)

At universities, students can study for Bachelor's and Master's degrees (science or arts) and also the licentiate and the doctorate degrees. In the two-cycle degree system, students first complete the Bachelor's degree, after which they may go for the Master's degree. As a rule, students are admitted to study for the Master’s degree. Figure 1 illustrates the education system in Finland.

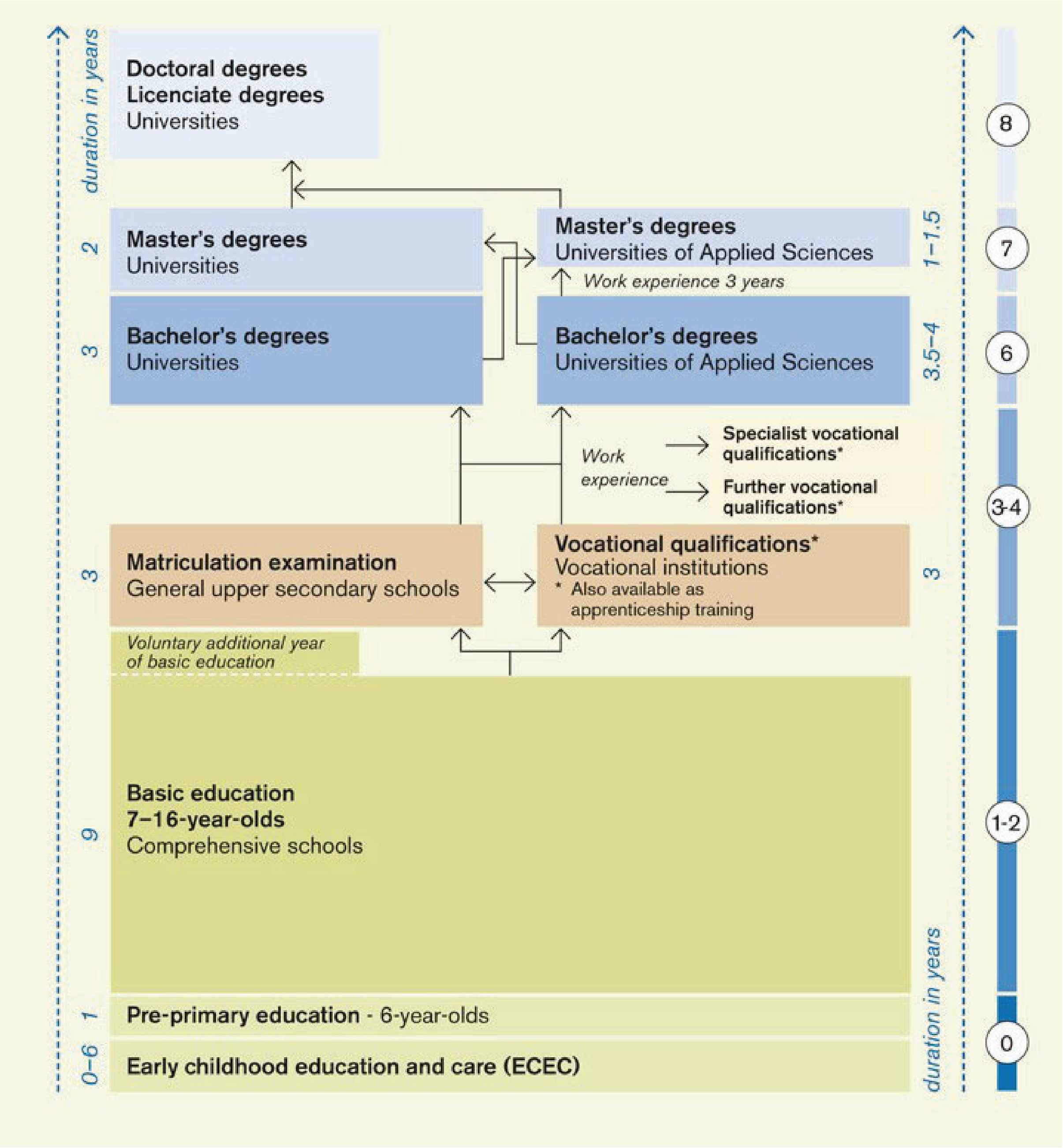

duration in years
Doctoral degrees
Licenciate degrees
Universities
2
Master's degrees
Universities
Master's degrees
Universities of Applied Sciences
Work experience 3 years
1–1.5
3
Bachelor's degrees
Universities
Bachelor's degrees
Universities of Applied Sciences
3.5–4
Work
experience
Specialist vocational qualifications*
Further vocational qualifications*
3
Matriculation examination
General upper secondary schools
Vocational qualifications*
Vocational institutions
* Also available as apprenticeship training
3
Voluntary additional year of basic education
9
Basic education
7–16-year-olds
Comprehensive schools
1
Pre-primary education - 6-year-olds
0–6
Early childhood education and care (ECEC)
duration in years
8
7
6
3-4
1-2
0

Figure 1: Education system in Finland[1]

3. The Guidelines for Cyber Security education

3.1 Cybersecurity Strategy of the European Union 2013 and 2017

According to the Cybersecurity Strategy of the European Union (2013) EU should safeguard an online environment providing the highest possible freedom and security for the benefit of everyone. The strategy proposes specific actions that can enhance the EU's overall performance. These actions are both short and long term, they include a variety of policy tools and involve different types of actors, be it the EU institutions or Member States or industry. The Cybersecurity Strategy adopted, set out strategic objectives and concrete actions to achieve resilience, reduce cybercrime, develop cyber defence policy and capabilities, develop industrial and technological resources and establish a coherent international cyberspace policy for the EU. (European Commission, 2013)

[1] The Finnish National Board on Research Integrity TENK and the Committee for Public Information TJNK, 2016

In September 2017 the EU updated its 2013 Cyber Security Strategy. According the proposal European Union has taken several actions to increase resilience and enhance its cybersecurity preparedness. The proposed regulation provides for a comprehensive set of measures that build on previous actions and fosters mutually reinforcing specific objectives:

- Increasing capabilities and preparedness of Member States and businesses
- Improving cooperation and coordination across Member States and EU institutions, agencies and bodies
- Increasing EU level capabilities to complement the action of Member States, in the case of cross-border cyber crises
- Increasing awareness of citizens and businesses on cybersecurity issues
- Increasing the overall transparency of cybersecurity assurance of ICT products and services to strengthen trust in the digital single market and in digital innovation
- Avoiding fragmentation of certification schemes in the EU and related security requirements and evaluation criteria across Member States and sectors.

(European Commission, 2017)

The new version is intended to improve the protection of Europe's critical infrastructure and boost the EU's cyber resilience in a world of digitalization.

3.2 Finland´s Cyber security strategy 2013 and 2019

According to Finland's cyber security strategy 2013, Finland has excellent chances of rising to the vanguard in cyber security. The implementation of cyber security R&D and education at different levels does not only strengthen national expertise, it also bolsters Finland as an information society. Cyber security development will heavily invest in cyber research and development as well as in education, employment and product development so that Finland can become one of the leading countries in cyber security. The strategic goal 7 states that "Inputs into R&D and education will be increased as well as action to improve cyber security know-how in the whole of society." (Security and Defence Committee, 2013)

According to Finland's new cyber security strategy 2019, national cyber security competence will be ensured by identifying requirements and strengthening education and research. Finnish society needs cyber security competence both in public administration and in the business community. At the national level, it must ensure that companies have both top-level experts and other competent personnel. Skills development is also emphasized in the cooperation between the business community and research. (Security and Defence Committee, 2019)

The new strategy and its implementation are also part of the implementation of the EU Cyber Security Strategy. The Finnish cyber security strategy (2019) presents the following strategic measures to promote cyber security competence:

- Training programs related to cyber and information security, software and application development, information networks and telecommunications in vocational education, universities of applied sciences and universities will be strengthened.
- The high level of education required by nationally critical cyber competence areas will be ensured. This is supported by both national and international training and exercises.
- The national system for training and exercising digital security will be strengthened as part of digital security training in the public administration.
- The public administration, the business community and private individuals will be made more aware of information security for new services and products.

3.3 Finland's cyber security report 2017

The report for the Prime Minister office states; "Cyber security is the enabler of digitalization. Finland has strong international trust capital and it is essential to utilize this trust and Finnish cyber expertise. Without credible private-sector business in this field Finland cannot be a forerunner in cyber security. In order to achieve the cyber security vision significant extra investments in resources are needed, for example to strengthen the activity of the National Cyber Security Centre." (Lehto et.al, 2017)

The report further states, that 'General awareness and knowledge of the basic elements of cyber and information security are considered to be basic civic competences in the present-day digitalized information society in Finland. General awareness of the basic elements of cyber security and incorporating them into everyone's daily routine must be actively improved in Finland. The training of experts requires better coordination between the fragmented education and research that exists among educational establishments as well as extending research even wider.' (Lehto et.al, 2017)

3.4 The implementation program for Finland's Cyber Security Strategy

The implementation program for 2017–2020 addresses the development of cyber security within the service complex comprising the state, counties, municipalities, the business sector and the third sector in which the individual citizen is the customer. The business community provides most digital services and their cyber security through international service complexes and networks. (Security and Defence Committee, 2017)
The implementation program set four goals in the area of cyber security competence:

- The competence of the public administration's information and cyber security personnel will be improved
- Citizens will have better information and cyber security skills
- A national information and cyber security week will be organized annually
- Basic skills in cyber security and the digital environment – general education and professional training will progress

So far, the implementation programs of the 2013 cyber security strategy have been based solely on proposals from actors committed to its development and the partly sectoral work of the competent authorities. This will be improved by a cyber security development program extending beyond government terms; this will replace the earlier implementation program. The program will concretize national cyber security policies and clarify the overall picture of cyber security projects, research and development programs. (Security and Defence Committee, 2019)

4. Cyber security as a field of education

Cyber security competence is not just another professional field of expertise. Rather, it ranges from civic skills all the way to international-level professions. Therefore, cyber security should be included in different educational levels. When it comes to comprehensive level education, the education must ensure that pupils after basic education sufficiently have the skills required by the cyber domain, that they understand its threats, and that they can protect themselves accordingly. In the upper-secondary level and vocational education these competencies are further deepened, creating the base for special expertise in higher-level education. Vocational education can include such cyber security education and training which provides basic professional skills and qualifications needed in working life. (Lehto & Niemelä, 2019)

Universities emphasize the scientific research of cyber security, and concomitant education. Universities of Applied Sciences provide practical-oriented cyber security education which corresponds to the needs of employment. It should be possible to receive Bachelor's and Master's degrees in cyber security. Correspondingly, universities should also make it possible to receive Bachelor's, Master's and Doctoral degrees in cyber security. Cyber security education provided across the spectrum of higher education generates top-level experts for the different tiers of society, whose skills and know-how meet the competence requirements determined for each task.

Cyber security must be included as a part of adult education. Such education can include basic degree level education, studies included in a degree, training that prepares for competence-based qualification, apprentice training, supplementary and continued education to update and expand a person's professional skills as well as social studies and leisure studies that improve civic and working skills.

The effective development of national cyber security expertise demands the definition of competence areas and their contents and create content of the cybersecurity curricular framework for education organizations, to make it possible for each level of education to provide the needed education.

There is not available today a globally accepted and standardized definition of cybersecurity and a clear identification of its domain of development and of application. Also, no common content of the cybersecurity curricular framework is not available. Many organizations however defined their own taxonomy tailored for their own specific needs. The table 1 describes cyber security taxonomies from four organizations. (JRC, 2019)

Table 1: Different cyber security taxonomies

CSEC, 2017	The Association for Computing Machinery (ACM), 2012	NIST Computer Security Resource Centre (CSRC)	The Institute of Electrical and Electronics Engineers (IEEE), 2017
• Data Security • Software Security • Component Security • Connection Security • System Security • Human Security • Organizational Security • Societal Security	• Cryptography • Security services • Network security • Security in hardware • Software and application security • Systems security • Formal methods and theory of security • Intrusion/anomaly detection and malware mitigation • Database and storage security • Human and societal aspects of security and privacy	• Security and privacy • Technologies • Applications • Laws and regulations • Sectors	• Access control • Computer security • Cryptography • Data security • Information security • Terrorism

5. Cyber Security Education in Finnish Universities

Security and the cyber domain embrace many walks of life and research subjects. Several disciplines address themes associated with cyber security. For a long time already, questions associated with cyber security have been studied in computer science and engineering, information systems science, information processing science, ICT technology as well as automation science and engineering.

Traditionally, universities have dedicated separate departments to information technology, the information processing sciences and to automation science and engineering. In addition to these, research has been intensified in big data, cloud services, usability and embedded systems, among others, which also include perspectives on cyber security.

In support of continuously improving the competence and awareness of the actors of society, inputs will be made to developing, utilizing and training common cyber security and information security instructions. Inputs into R&D and education will be increased as well as action to improve cyber security know-how in the whole of society.

5.1 Aalto University

In the Aalto University ICT and digitalization research relates to both theory and practice, computational and mathematical sciences, software and hardware technologies, secure communications, digital media, digital services, and digital engineering and management across technologies and industrial sectors. Aalto University teaches and conducts research on a wide range of technical and societal topics in the field of cyber security. (Aalto, 2020a)

At the School of Electrical Engineering, the natural sciences, engineering and information technology intertwine to form smart systems and innovations that save energy and promote well-being. The Department of Communications and Networking (Comnet), is the largest unit in Finland in its research area. The Department of Communications and Networking provides most of the courses of Communications Engineering (CE) major, which is one of the majors under the Master's Program in Computer, Communication and Information Sciences. The CE major gives a solid understanding of Internet technologies, wireless communications and communications ecosystems - from the perspective of concepts, technologies and methodologies. A Diploma in Digital Security program will provide up-to-date information and tools for the security management of a digitalizing business. The training will help you enable your company to function as part of the cyber environment. As well as helping participants to identify and prepare for threats and disruptions, it will open their eyes to new opportunities. (Aalto, 2020b; 2020c)

5.2 University of Helsinki

The Faculty of Science carries out top-notch research on an international level. The instruction is developed and taught by leading scientists. Data science combines computer science and statistics to solve exciting data-intensive problems in industry and many fields of science. In the interdisciplinary Master's Program in Data Science the focus is data-intensive areas of industry and science, with the skills and knowledge needed to construct solutions to complex data analysis problems. Students can specialize either in machine learning and algorithms, infrastructure and statistics or in its applications. (UH, 2020)

In the Master's Program in Computer Science, students get the skills that can lead to create new network solutions, build the future digital society, develop secure digital services, or be involved in a ground-breaking software project. The master's program has three study tracks: algorithms, networks and software. (Ibid.)

5.3 University of Jyväskylä

The University of Jyväskylä is a nationally and internationally significant research university and an expert on education that focuses on human and natural sciences. This profile is strengthened through the following core fields:

- Learning, teaching and interaction
- Basic natural phenomena and mathematical thinking
- Sustainable business and economics
- Language, culture and society
- Physical activity, health and wellbeing
- Information technology and the human in the knowledge society

The University focuses on cyber security, one of University’s profiling areas. University is the only cyber security university in Finland and the only one that has the Cyber Security Master of Science and Doctoral programs in Finland. (JYU, 2020a)

The Faculty of Information Technology plays a key role in developing one of the University’s core fields, information technology and the human in the knowledge society. One of the Faculty’s primary strengths is its ability to view IT broadly, integrating various perspectives and identifying the joint effects of different phenomena. This is combined with internationally recognized research in the strategic areas, as well as with active societal interaction. Cyber Security Master of Science program (120 ECTS) focuses on cyber security technology, cyber security management and cyber security in society. The aim of the Cyber Security Master of Science program is to provide solid skills in the kinds of demanding management and development tasks that require comprehensive awareness in cyber security. The studies comprise an entity which addresses the cyber world and its security from societal, functional, systemic and technological perspectives. (JYU, 2020b)

5.4 National Defence University

The National Defence University (NDU) is a military tertiary education institution that is part of the Finnish Defence Forces. NDU is responsible for educating the future leaders of Finland’s armed forces. The National Defence University is divided into divisions, departments and other administrative units. The university is led by a rector. The National Defence University offers undergraduate, masters and doctorate studies and research programs in the area of military science. The university has an under-graduate, graduate and post-graduate degree division. As well as providing a university education leading to an academic degree, the divisions are responsible for officer education. The National Defence University the major subjects in the undergraduate degree are leadership, military pedagogy, military theory and military technology. (NDU, 2020)

The officers graduating from the National Defence University are commissioned as officers in the Finnish Defence Forces and the Border Guard. There are four alternative study programs within a bachelor's program: The Army program, The Navy program, The Air Force program, Pilot Officer Program. (Ibid.)

5.5 University of Oulu

The University of Oulu trains experts and multifaceted professionals for large scale of specialized fields of study. Focus areas of the university are biosciences and health, information technology, cultural identity and interaction, environment, natural resources and materials. Focus areas are sustainable materials and systems, lifelong health, digitalization and smart society, changing climate and northern environment, understanding humans in change. (OY, 2020a)

Faculty of Information Technology and Electrical Engineering has the International Master’s Degree Program in Computer Science and Engineering (CSE). It is a two-year research-oriented program concentrating on intelligent digital solutions to real world problems. The program has three specialization options: Applied Computing, Artificial Intelligence and Computer Engineering. (OY, 2020b)

In the Faculty of Science are the degree programs of Bachelor of Science (BSc) and Master of Science (MSc) associated with the Research Unit of Mathematical Sciences. In the studies of computational mathematics and data technology is possible to study cryptography, data mining, big data and encryption techniques. (OY, 2020c)

5.6 Tampere University

Tampere University was created in January 2019 by the merger of the University of Tampere and Tampere University of Technology, which are joining forces to create a new foundation-based university. Cybersecurity is one of the areas of expertise in the Faculty of Information Technology and Communication Sciences. (TUNI, 2020a)

The Computing Sciences degree programs familiarize students with the broad scope of information technology ranging from circuit boards, electronics and digital processors to software and peripheral interfaces, software development, information networks, information security, and intelligent machines with the ability to learn. (TUNI, 2020b)

The Electrical Engineering degree program in electrical engineering provides students with the skills to put their knowledge of natural phenomena into practice. As graduates with a degree in electrical engineering understand the theoretical foundations underlying electrical phenomena, they have the ability to apply their knowledge to create new ideas and innovations that help promote the development of our society and the wider world. (Ibid.)

The Information Technology, Communication Systems and Networks degree programs focuses on the Internet of Things and the Industrial Internet, which are the forerunners of a connected and programmable world where things are also able to communicate. Digitalization will have an impact on a wide range of industries, such as energy, transportation, logistics and manufacturing. Another prevailing trend in society and industry is robotization which manifest in modern inventions, such as self-driving cars and drones, strongly rely on highly efficient and reliable wireless communications. (Ibid.)

5.7 University of Turku

The University of Turku is a science university whose operations are based on high-quality, multidisciplinary research. The research is profiled though the following thematic collaborations: Biofuture, Digital Futures, Cultural memory and social change, Children, young people and learning, Drug development and diagnostics, and Sea and maritime studies. The Strategy of the University of Turku has identified information security as a research area that are in an advanced stage of development. (UTU, 2020a)

The Master’s Degree Program in Information and Communication Technology (120 ECTS) provides versatile ICT education in selected fields of ICT, with an established reputation in innovative interdisciplinary and international education. The program offers four specialization tracks: Cyber Security, Smart Systems, Digital Health and Cryptography. (UTU, 2020b)

The Cryptography specialization track offers a solid background on classical and modern aspects of mathematical cryptography. The Master's thesis in the Cryptography specialization track focuses on the mathematical aspects of cryptosystems and cyber security in general. In what follows, some possible topics for the Master's thesis are listed:

- Public-key cryptosystems
- Homomorphic encryption

- Blockchains in cryptography
- Quantum cryptography

(UTU, 2020b)

5.8 Summary of the cyber security education in Finnish universities

Cyber security education is on the rise in Finland. Two models of education are being used: a curriculum that focuses on cyber security or one that integrates cyber security studies into other curricula. The previous model is used by the Jyväskylä and Turku Universities and the latter, integrated approach, by other universities. Both models are indispensable for improving competence in the field. The cyber security curriculum generates experts in the domain of cyber security who have become adept in some niche area. The integrated approach generates experts who master a specific field of technology (ICT technology, communications technology, automation science, etc.) as well as attendant cyber security issues. (Lehto & Niemelä, 2019)

Table 2 summarizes the number of cyber security courses and credits at the universities covered.

Table 2: Number of courses and ECTS in the universities

University	Number of the courses	Total ECTS
Aalto University	7	40
University of Helsinki	5	25
University of Jyväskylä	23	115
National Defence University	5	14
University of Oulu	5	25
Tampere University	9	44
University of Turku	19	92

Different cybersecurity curricular frameworks consist of different content in the cyber security courses. Usually the selection is tailored to meet the goals of the university's degree programs. Of course, universities with a full degree cyber security program have a wider range of courses. Table 3 summarizes the key cyber security courses available at Finnish universities.

Table 3: Cyber security course portfolio in the universities

	Aalto University	University of Helsinki	University of Jyväskylä	National Defence University	University of Oulu	Tampere University	University of Turku
Anomaly detection			X				
Computer security					X		
Cryptography	X	X			X	X	X
Cyber security basics		X	X	X	X		X
Cyber ethics			X				X
Cyber intelligence			X				
Cyber security and law			X				
Cyber security management			X		X	X	
Cyber security solutions	X		X				X
Cyber security in systems			X			X	X
Cyber warfare			X	X			
Human cyber security							X

	Aalto University	**University of Helsinki**	**University of Jyväskylä**	**National Defence University**	**University of Oulu**	**Tampere University**	**University of Turku**
Identity and access management						X	
Information operations			X				
Network cyber security	X	X	X			X	X
Operation security				X			
Privacy			X				X
Quality Assurance	X						
Secure programming						X	X
Societal cyber security			X				X

The basic elements of cybersecurity can be introduced in a course or two, but the programs should provide students with the opportunity to study cyber security in depth, just as they might choose specialties in cyber security management, risk management, critical infrastructure protection, cyber security technologies, resilience, or cyber warfare studies. Cybersecurity must take a multidisciplinary approach to offer both management and technology perspectives. With comprehensive cyber security education, we can provide students with a valuable skill set with which to address what might one of the most challenging national security and defense issues for the future. (Kessler & Ramsay, 2014)

6. Conclusion and discussion

Cybersecurity professionals are vital in today's world. They help develop new ways to combat cyber threats. A key element of the long-term national cyber strategy capacity building is formal education. This paper described cyber security education in Finland and the related principles and implementation models.

A shortage in the global cybersecurity workforce continues to be a problem for companies in all industries and of all sizes. The EU and Finland alike have set requirements for improving national cyber security competencies. The EU expects that network and information security (NIS) education be provided in schools and universities, and that supplementary training be given within the public administration.

IT skills, communication skills, media literacy and the use of social media constitute the foundation for the skills needed to use digital services. The proposed goals included the incorporation of ICT use as an integral part of the education in schools as well as in basic and supplementary teacher training. In addition, the investment in applied ICT know-how should be scaled up – with cyber security being one of its elements – and given a more prominent place in curriculum design throughout all tertiary education.

This research and the international benchmarking shows that cyber security is taught in different curriculum in Master of Science programs. Models are cybersecurity as an independent program, cyber security in different disciplines of IT, or cybersecurity within other disciplines.

The challenge regarding the way things are now is the lack of cyber security education objectives for the entire education system. The sought-after improvement of competencies requires defining the basic skills and competencies for the entire national education system. It is necessary to have an understanding of what each citizen needs to know about cyber security, the demands of working life, what kind of professional skills are needed and what kind of supplementary training needs to be provided to those already in working life. At the moment universities provide cyber security education from their own perspectives without clear information of national cyber security needs (skills and numbers).

National cyber security of supply requires enough motivated staff in terms of quantity and quality. The future development program based on the Finnish cyber security strategy should have clear definitions of critical areas of competence. A study of the private and public sector needs is needed. A long-term investment program is needed to implement cyber security education. The promotion and facilitation of national cyber security research and education requires a hub or center of excellence that brings together universities, industry and security authorities. Such co-operation is essential in a small country like Finland, where resources are limited

References

Aalto (2020a) Aalto University, https://www.aalto.fi/en/school-of-electrical-engineering

Aalto (2020b) Aalto University, https://www.aalto.fi/en/department-of-communications-and-networking

Aalto (2020c) Aalto University, https://www.aaltopro.fi/en/programs/diploma-in-digital-security

CSEC (2017) *Cybersecurity Curricula, Version 1.0*, Report 31 December 2017

European Commission (2013) *Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace,* JOIN(2013) 1 final, Brussels, 7.2.2013

European Commission (2017) *Cybersecurity Act*, COM(2017) 477 final, Brussels, 13.9.2017

(ISC)² (2019) *Cybersecurity Workforce Study*, 2019

JRC (2019) *A Proposal for a European Cybersecurity Taxonomy*, Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service.

JYU (2020a) University of Jyväskylä https://www.jyu.fi/en

JYU (2020b) University of Jyväskylä https://www.jyu.fi/it/en

Kessler G.C. and Ramsay J.D. (2014) A Proposed Curriculum in Cybersecurity Education Targeting Homeland Security Students, 47th Hawaii International Conference on System Science

Kupiainen S., Hautamäki J., Karjalainen T. (2009) The Finnish Education System and Pisa, Ministry of Education Publications, Finland 2009:46

Lehto M., Limnéll J., Innola E., Pöyhönen J., Rusi T., Salminen M. (2017) *Finland's cyber security: the present state, vision and the actions needed to achieve the vision*, Publications of the Government´s analysis, assessment and research activities 30/2017

Lehto M., Niemelä J. (2019) *Kyberturvallisuuden kansallinen osaaminen*, University of Jyväskylä, Faculty of Information Technology, research paper, 83/2019

MEC (2020) *Education System in Finland*, Ministry of Education and Culture,

NDU (2020) National Defence University https://maanpuolustuskorkeakoulu.fi/en/studies

OY (2020a) University of Oulu https://www.oulu.fi/university/

OY (2020b) University of Oulu https://www.oulu.fi/itee/

OY (2020c) University of Oulu https://www.oulu.fi/science/

PMO (2016) *Government Report on Finnish Foreign and Security Policy*, Prime Minister's Office Publications 9/2016

PMO (2017) *Government's Defence Report*, Prime Minister's Office Publications 7/2017

Security and Defence Committee (2013) *Finland's Cyber Security Strategy*, Government Resolution, 24 January 2013, http://www.yhteiskunnanturvallisuus.fi/en/materials

Security and Defence Committee (2017) *Implementation Programme for Finland's Cyber Security Strategy for 2017–2020*

Security and Defence Committee (2019) *Finland's cyber security strategy 2019*, Government Resolution 3.10.2019

TUNI (2020a) University of Tampere https://www.tuni.fi/en/about-us/tampere-university

TUNI (2020b) University of Tampere https://www.tuni.fi/en/about-us/faculty-information-technology-and-communication-sciences

UH (2020) University of Helsinki, https://www.helsinki.fi/en/faculty-of-science

UTU (2020a) University of Turku https://www.utu.fi/en

UTU (2020b) University of Turku https://www.utu.fi/en/study-at-utu/masters-degree-programme-in-information-and-communication-technology-smart-systems

How to Secure the Communication and Authentication in the IIoT: A SRAM-based Hybrid Cryptosystem

Christoph Lipps, Pascal Ahr and Hans Dieter Schotten
German Research Center for Artificial Intelligence, Kaiserslautern, Germany
Christoph.Lipps@dfki.de
Pasacal.Ahr@dfki.de
Hans_Dieter.Schotten@dfki.de

DOI: 10.34190/EWS.20.061

Abstract: Currently, the developments of the Fourth Industrial Revolution are taking place, accompanied by the advancements of the Industrial Internet of Things (IIoT). This includes, among others, the interconnection of different industrial spheres, devices and use-cases up to Machine-to-Machine (M2M) and Machine-to-Service (M2S) communication. However, especially this communication is critical because of the partly sensitive content as well as the amount to data transmitted. Furthermore, the reliability and integrity of the data, in particular with regard to industrial applications, an important issue. But as the IIoT devices are designed for low energy consumption rather than to handle with complex cryptographic approaches, new lightweight but nevertheless sound and secure techniques are required. Furthermore, a strong authentication with a power optimized technique and an access control management is necessary. To guarantee both, a secure communication and a strong authentication, a Physical Layer Security (PhySec) based system, in particular a Static Random-Access Memory (SRAM) related approach is a promising opportunity. Especially because most Microcontroller Units (MCUs) are already equipped with SRAM, which requires no additional implementation effort. In this work the ability of SRAMs to use them as a Physical Unclonable Function (PUF) as well as the inherent given characteristics are examined. For instance, the start-up value - the hardware fingerprint of the device – is taken into account. Despite the reputation of PUFs to enable a hardware related deviation of cryptographic keys for secure communication and device authentication, the lack of practical usability if often criticized. To face this, in this work a practical application for PUFs with its potential with respect to the IIoT is presented. Therefore, within a M2M communication scenario the application of a SRAM-PUF driven hybrid cryptosystem is demonstrated. A secure asymmetric cryptosystem is applied to exchange synchronisation data, followed by the PUF-based cryptography. The individual key is calculated from their PUF sequence in conjunction with pre-transmitted helper data. As another benefit of the approach, there is no need to store any cryptographic credentials on the device itself, because the key is regenerated every time required. This enables not only completely new applications in IIoT environments but is also a resource saving, lightweight and powerful security primitive.

Keywords: Physical Layer Security; Physically Unclonable Functions; Authentication; Static Random-Access Memory; Secure Communication; Plug & Trust;

1. The Requirements of the Industrial Internet of Things and how to Face them

At present, the industrial landscape is in a state of upheaval, a jolt passes through the infrastructure and leads to changes in the way communication takes place. Furthermore, new application scenarios such as Machine-to-Machine (M2M) and Machine-to-Service (M2S) communication arise, motivated by the ever growing amount of inter-connected devices (Farooq, et al., 2015). The networks itself, the industiral manufactoring environments as well as the Industrial Automation and Control Systems (IACSs) are developing to an Industrial Internt of Things (IIoT) and Cyber-Physical Production Systems (CPPS). This new and altered production environments consist of a multitude of communicating devices such as Automated Guided Vehicles (AGVs), Robot Motion Control Systems (RMCSs), sensors and actuators.

The development is driven by the key enablers mobility, flexibility and scalability to overcome the rigid structures and low processing times within the networks. Complex and computational intensive security mechanisms are as counterproductive as unwieldy key management structures. Restrictions in form of bandwidth limitations and requirements such as integrity of messages and data transmitted do not make things easier. The ongoing challenges with regard to the IIoT object identification, location and authentication remain (Zhang, et al., 2014).

Besides that, in industrial use-cases the live cycles are much longer compared with Commercial-of-the-Shelf (COTS) hardware components, which makes it a crucial task to establish and integrate security mechanism into existing systems. In addition, most of the devices do not even have any kind of interface to enter *traditional* credentials such as username and passwords. In general, human participants are increasingly rare.

Nevertheless, the task of reliable and trustworthy communication including a doubtless authentication of participating entities remains. A lightweight, flexible but still sound and secure mechanism to enable trust and confidence into the existing as well as the upcoming networks must be the aim. This can be addressed by the application of new authentication schemes in combination with already existing and inherent given features of IIoT components: Static Random-Access Memory (SRAM). The concept of Physically Unclonable Functions (PUFs), utilizing the behaviour and manufacturing individual characteristics of SRAM-cells is a promising and worthwhile approach to secure the systems.

A benefit of the approach is the resource efficiency, in that case that no complex computations are required to calculate cryptographic credentials, the "secrets" are already given by the circuit, intrinsic. Furthermore, the keys do not need to be stored in Non-Volatile Memory (NVM) – key storage is also vulnerable to attacks and prone to reengineering – as the keys can be regenerated every time required.

The hybrid cryptosystem discussed in this work uses the SRAM-based PUF concept to derive individual cryptographic credentials, a binary PUF vector (Kusters & Willems, 2019), that can be used for encryption as well as for device authentication. In combination with a Master-Slave set-up, remains a symmetric cryptography system, including all benefits of Symmetric Key Cryptography (SKC).

The remainder of this work is organized as follows. In Section 2 an overall introduction to the Physical Layer Security and PUF concept is given, to provide a basic introduction to the topic. Section 3 describes the operating principle of the proposed hybrid cryptosystem. To demonstrate the general applicability of SRAM-cells with respect to cryptographic operations, Section 4 provides the evaluation and results of a SRAM-cell test scenario. Finally, Section 5 concludes the work and provides an outlook on future work and the next steps to the further enhancements of the concept.

2. The General Concept of Physical Layer Security and Network Security

The general concept of Physical Layer Security (PhySec) is based on various individual physical properties of mediums and electrical circuit components. The Secret Key Generation (SKG) (Ambekar, et al., 2012) also known as Channel Reciprocity based Key Generation (CRKC) (Zenger, et al., 2015) uses influences of a wireless radio channel to derive cryptographic credentials and establish symmetric keys on both sides of this channel (Weinand, et al., 2017).

Besides that, the PUF concept as electrical or silicon PhySec derivative is based on various hardware characteristics. Since the general introduction by Gassend, et al. (2002) meanwhile exist a lot of different approaches such as Ring-Oscillator PUFs (Gao & Qu, 2014), Field Programmable Gate Array (FPGA) intrinsic PUFs (Guajardo, et al., 2007) and Optial-PUFs (Rührmair, et al., 2013). Furthermore, SRAM-based approaches, as discussed in more detail in Section 4, are a cheap, easy to handle and promising alternative.

The significantly higher physical security (Suh & Devadas, 2007) that can be achieved by PUFs is ensured by the physical "secrets" and the complex structures of the used components. Slightest deviations and inaccuracies during the manufacturing of the components, for instance variations in the oxide layer thickness and impurities during the doping of semiconductor devices (Lipps, et al., 2018; Halak, 2018), add up to a chip individual behavior. However, these differences are so small and cannot be influenced and predicted that the overall behavior is unfeasible to predict (Kusters & Willems, 2019). Therefore, PUFs are often called as "an objects fingerprint" (Maes, 2013) and „a physical entity whose behavior is a function of its structure" (Halak, 2018).

Besides the PUF solutions there are different methods to secure communication in the IIoT sector. Plaga, et al. (2018) propose an Out-of Band Authentication (OoBA) protocol with One Time Passwords (OTPs) and Quick Response (QR) codes in an combination with Transport Layer Security (TLS). An IPsec extension for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPAN) in short range wireless networks is descirbed by Raza, et al. (2012). Another hardware related opportunity are Trusted Platform Modules, as they are used by Hoeller & Toegl (2018) to secure Cyber Physical Systems (CPSs), but with the difference to PUFs that the cryptographic credaentials are precaluculated and brought into the systems from outside. A certificate-based secure key management scheme related to Elliptic Curve Cryptography (ECC) is applied by Lee, et al. (2014), and the overall necessity to abolish and replace (traditional) passwords, also becouse human users are becoming increasingly rare in industrial workarounds, is requested by Bonneau, et al. (2012). Not only beacouse of this work, the concept of the SRAM-based Hybrid Cryptosystem, descibed in Section 3, can be derived. An excerpt and overview of the work of other is provided in Table 1.

Table 1: Overview of PhySec, PUF and IoT-Security work

Scope of the work	Work
General introduction of the PUF concept	Gassend, et al., 2002
Ring-Oscillator PUFs	Gao & Qu, 2014
PUFs and their applications	Herder, et al., 2014
SRAM as Device-Fingerprint and TRNG	Holcomb, et al., 2009
SRAM characteristics and suitability tests for PUFs	Lipps, et al., 2018
FPGA intrinsic PUFs	Guajardo, et al., 2007
Optical PUFs	Rührmair, et al., 2013
SRAM PUF Key Generation	Gao, et al., 2019
PhySec in Cellular Networks	Lipps, et al., 2019
Multiple Enrolment Scheme for SRAM-PUFs	Kusters & Willems, 2019
General PUF Introduction	Suh & Devadas, 2007
Combination of PhySec and SDN	Lipps, et al., 2019b
Demand to replace password	Bonneau, et al., 2012
Certificate-based approach with ECC	Lee, et al., 2014
One Time Password authentication protocol	Plaga, et al., 2018
TPMs in Cyber Physical Systems	Hoeller & Toegl, 2018

3. The General Concept of the Hybrid Cryptosystem

This chapter introduces the general concept of the hybrid cryptosystem, a *Master-Slave-System* to generate a secure communication environment. IIoT devices do not have high performance processors and they are not suitable and designed to consume much energy. Therefore, on those devices, complex computations like asymmetric cryptography, for example the decryption part of the Rivest–Shamir–Adleman (RSA) algorithm (Schneier, 2015), do not perform well and cannot be used reliably. The drawback of using a symmetric cryptography system, which performs much better with respect to energy consumption, it is necessary to exchange the symmetric keys anyhow. Thus, there are often hybrid methods implemented.

In a first step, they use the asymmetric system to exchange the keys. In the next stage they use the symmetric method with the exchanged keys. Another disadvantage of direct communication between IIoT devices with a symmetric key is a growing square number of keys with added number of devices. A *Master-Slave* system with hybrid cryptographic concept combines the advantages of symmetric methods with those of asymmetric ones by avoiding disadvantages of both.

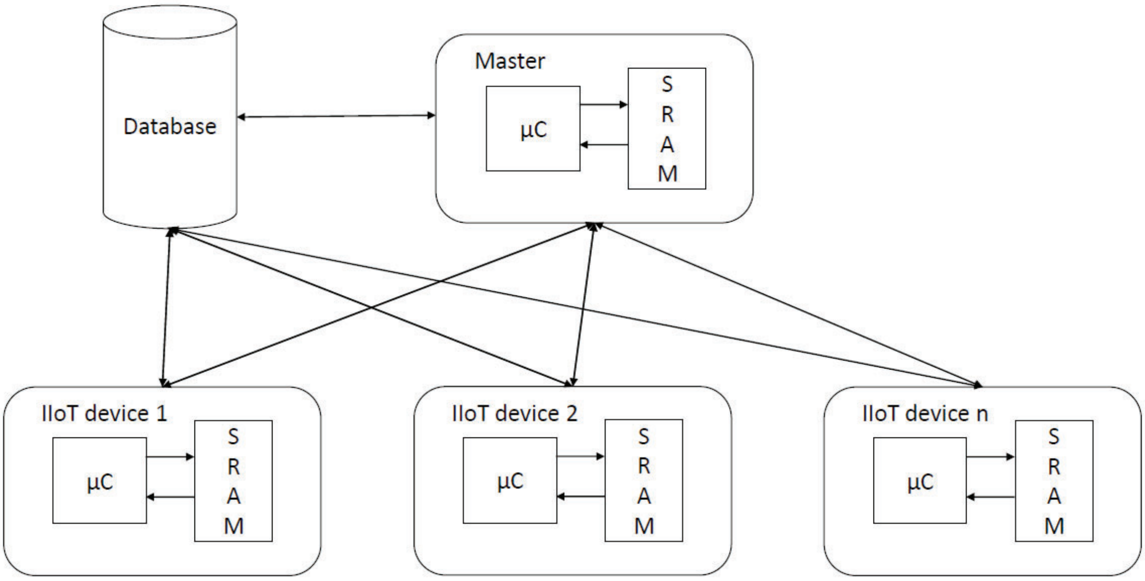


Figure 1: Master-Slave Set-Up

Such a structure consists of database, master and slaves. Every device is able to communicate with each group of devices bi-directionally. The database contains helper data for decryption. The master is the only one which does the complex parts of cryptographic computations as well as manage the communication between the IIoT devices. The master is the central part of the structure and does not have to follow the same conditions like the IIoT devices. Due to that, the master does not care about energy saving and can contain high performance processors. This structure is illustrated in Figure 1.

The first task to implement the Master-Slave Hybrid Cryptosystem is to generate some helper data. All IIoT devices and the master generate a unique key, a binary PUF vector, which have to be accessible for themselves every time. As next step it is necessary to establish an asymmetric cryptographic connection between the master and the slaves. The slaves encrypt their unique generated key with the public key from the master and then send it. The master decrypt the received encrypted key and perform a suitable function with it and its own generated unique key. Because the master does the asymmetric decryption and the IIoT device does only the simple encryption with the public key, it is possible to use an asymmetric method. The result of the suitable function is stored in the database as helper data. This key exchange is done for every IIoT device. The procedure is illustrated in Figure 2.

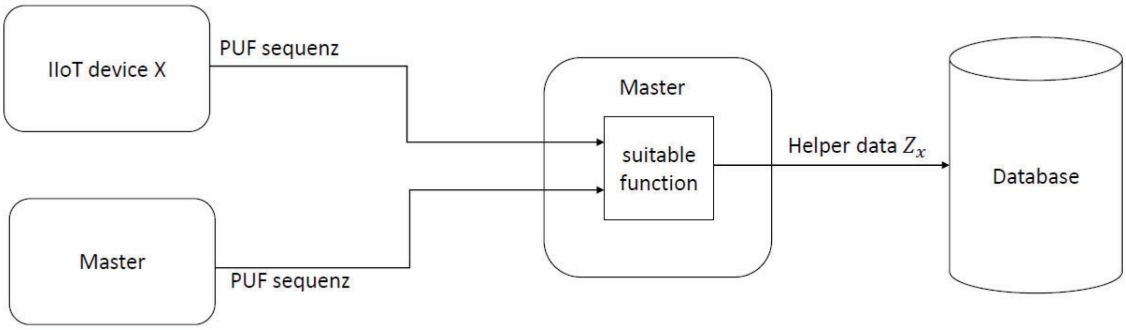


Figure 2: PUF Vector and Helper Data Concept

After the helper data has been created and transferred to the database, the asymmetric part of the cryptography communication is done. An IIoT device perform the same suitable function, like the master has done, with its unique generated all time accessible key and the data that it would like to send. The result of this function is an encrypted *ciphertext* of the massage which can be sent to the master. This construct is illustrated in Figure 3.

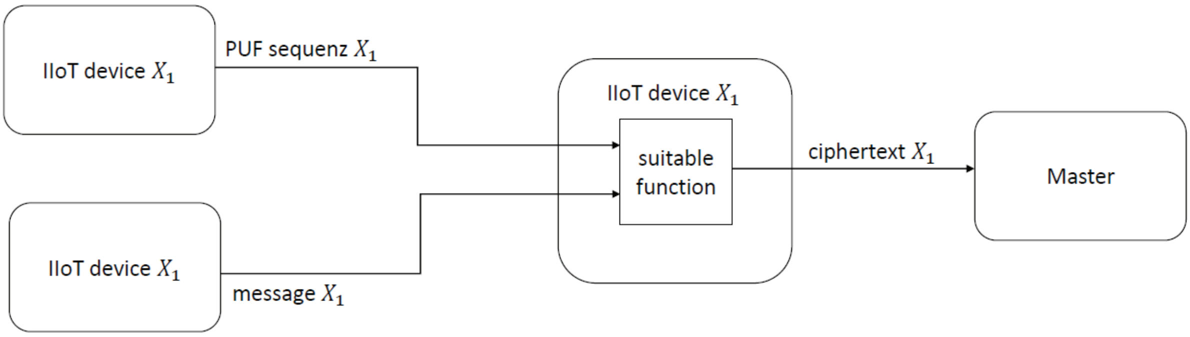


Figure 2: Message Encryption Process

As the next step, the master takes the helper data of the sending device from the database and perform with its generated unique key, the helper data and the send message, the inverse of the suitable function of step one. The result of this procedure is the encrypted message. The procedure is illustrated in Figure 4.

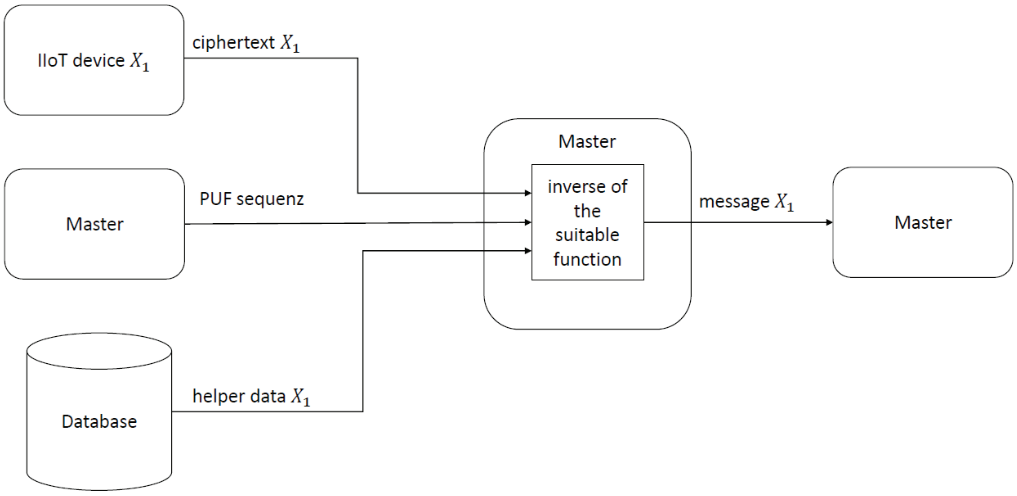


Figure 3: Message Decryption Process

To send the message from one device to another IIoT device, the master encrypts the received decrypted message by performing the same suitable function, like in step one, with the encrypt message and its own unique key. The result is another ciphertext of the message of the IIoT device at the very beginning. Afterwards the master sends it to the target IIoT device.

As a final step of the communication, the ciphertext must be decrypted on the target device. This takes effect by taking the helper data of the master/target device from the database and performing the inverse function of the suitable one, like in master, with its own unique key, the received ciphertext and the helper data. The target device of the sending device receives the message.

In this way, no IIoT device has to perform high computable tasks to encrypt and decrypt messages. All unique keys of the devices, which nowhere else have to be stored, except on them self. The helper data can be stored publicly, if the suitable function is strong enough and the length of the generated unique keys are long enough. Because no IIoT device has to perform the complex part of the asymmetric method it is simply possible to exchange the keys with a high frequent to ensure that it is not possible that it is not possible to get information out of the helper data or the ciphertext by simple Brute-Force-Attacks. An example for a suitable function is given in Equations 3.1 – 3.4.

$$S_1 \times S_2 = z_{1,2} \tag{3.1}$$

$$S_1 = \frac{z_{1,2}}{S_2} \tag{3.2}$$

$$x \cdot S_1 = y \tag{3.3}$$

$$x = \frac{y \times S_2}{z_{1,2}} \tag{3.4}$$

With S_1 as unique all-time accessible key from device 1, S_2 as unique all-time accessible key from device 2, $z_{1,2}$ as helper data, x as decrypted message and y as encrypted message.

4. SRAM – based Physically Unclonable Functions

To generate the unique all-time accessible key which is necessary to implement the Master-Slave System, described earlier, Static Random Access Memory - Physical Unclonable Function (SRAM-PUFs) is a worthwhile approach. Usually SRAMs are built of six Transistor (6T) Metal-Oxide Semiconductor Field-Effect Transistors (MOSFETs) (Lipps, et al., 2018) which are connected like two successive inverters. Because of this structure, SRAM cells are bi-stable circuits and always contain a stored ‘one’ or a stored ‘zero’.

By applying a source voltage on the SRAM, the SRAM cells take the state 'one' or 'zero'. This is the so-called start-up value. Most of the cells always take the same state by starting up. Due to production fluctuations MOSFETs are always different in their channel length, channel width and doping. This causes different threshold voltages, so that some MOSFET conduct earlier or better than others and the regenerative behaviour of two successive connected inverter let the cell takes one of the two states. But some cells change their state by new start-ups, so they do not have always the same start-up value. The reason for this behaviour is that, the two inverters of the cells have nearly the same threshold voltage. At some start-ups, one inverter conducts earlier at another time the other one does and the start-up value is changing. This behaviour is comparable to a ball on different inclined planes. As depicted in Figure 5, the probability of the ball rolling to the '1' or '0' side depends on the inclination of the plane. This corresponds to the to the threshold values of the MOSFETs, and results in the stable start-up states. The ball upon a flat plane, cells near to the individual threshold value, have the same probability of '1' and '0' and can therefore switch one state to another at a new start-up. The preference of state and cell is randomly due to physical characteristics by fabrication, which described earlier. Those physical characteristics generate a unique technical attribute, equivalent to a human fingerprint, the device fingerprint of every SRAM (Maes, 2013). By reading the stable start-up values it is possible to extract a PUF out of SRAMs. The PUF can be used as the unique all-time accessible key, because it is constant by using only stable cells as well as unpredictable. The absence of a necessity of storing the key is a further advantage of using SRAM PUFs. The key can always be regenerated every time needed by reading the start-up values. A detailed description about this can be found in Guajardo, et al. (2007), Lipps, et al. (2018) and Böhm & Hofer (2013).

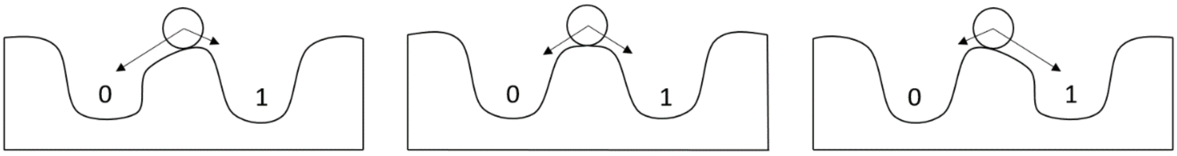


Figure 4: Start-Up behaviour of SRAM Cells

For detailed results, a test series to evaluate the SRAM as an PUF was done. The test includes four different SRAMs of different manufacturer with different sizes and all are assembled with six MOSFET SRAM cells in a stack of 8-Bit words. Table 2 provides an overview about the different SRAM manufacturers, sizes and types, used in the validation. The evaluation set-up itself is depicted in Figure 6. To measure the start-up values, an Arduino Mega is utilized. It serves as power supply of the SRAM and read all cells by incrementing the SRAM addresses. Before starting the next measurement, the SRAM will be restarted by power down and up all Control Signal (CS) Pins, Input/Output (I/O) Pins and the supply voltage, which are controlled by Arduino Mega. To guarantee that the SRAM is completely shut down, there is a delay between power down and power up, so that all discharging processes are completed.

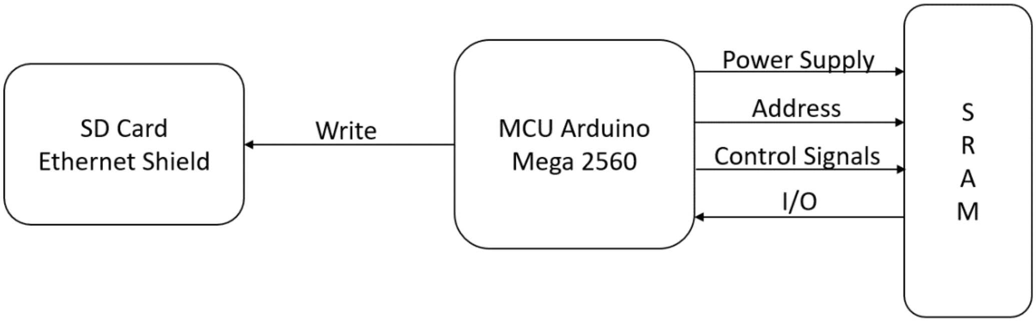


Figure 5: Evaluation Set-Up of MCU and SRAM

Table 2: SRAM Circuits Evaluated

Manufacturer	Size	Type
UTRON	8 KB	UT6264C
S@Tech	32 KB	HM62256
Alliance	32 KB	AS6C62256
Alliance	512 KB	AS6C4008

This process will return 500 times, because it is a compromise between needed time and accuracy. Only the 512 KB SRAM is measured for 250 times due to its size and thus time is needed. The measurement results are analysed by stability over all measurements based on the first one and the number of ones and zeros per measurement. Figure 6 illustrate the results of the Alliance 32 KB SRAM representatively. At Figure 6 the Y-axis shows the percentage of different cells based on the first one, and there is the measurement on the X-axis. The line above 6 percentage is the average and the points are the results of the measurement. Figure 6 presents at its Y-axis the percentage of ones and at the X-axis the measurement. In this figure, the points are the results and the line above 41.6 percentage, is the average as well. All measured SRAMS exhibit the same behaviour qualitatively, there are only some differences in quantity.

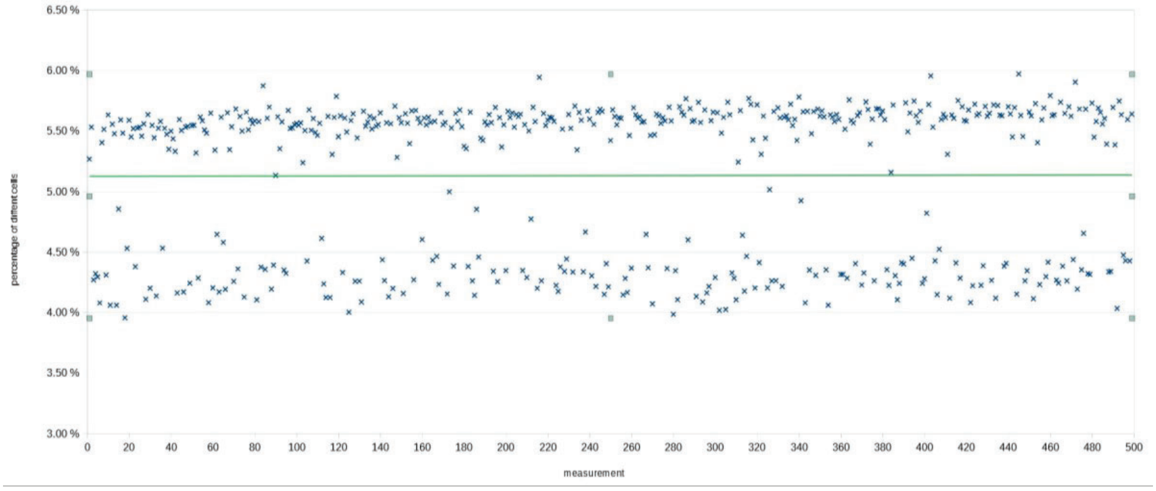


Figure 6: Percentage of Non-Equal Cells based on the First Measurement

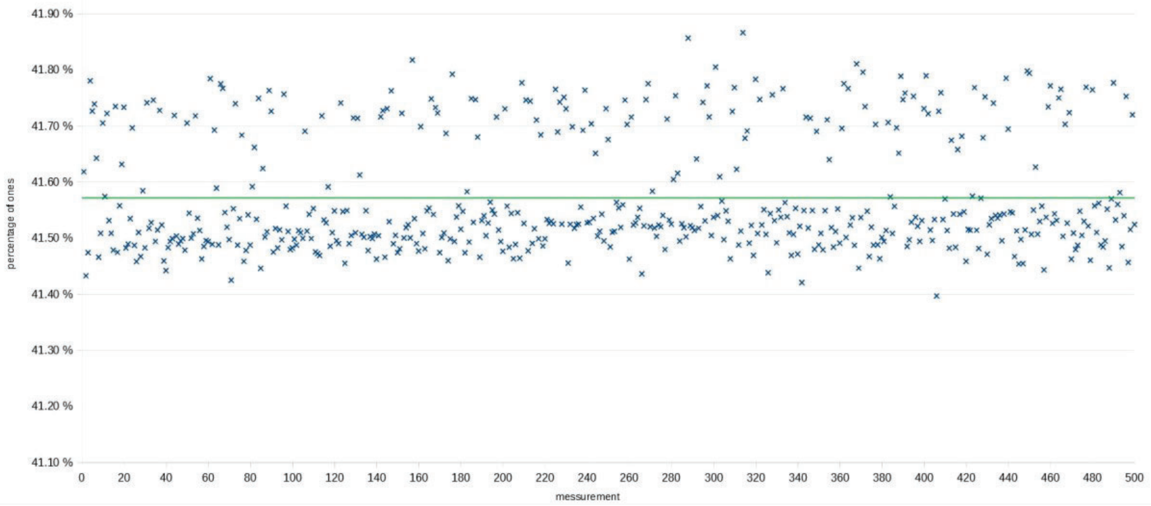


Figure 7: Percentage of Ones per Measurement

The maximum of differences based on the first one in this case is 6% by an average of 5.13% and a standard deviation of 0.0063%. Non-stable cells and environmental influences causes this behaviour for more information see (Lipps, et al., 2018). The average number of measured ones is 41.57% by a standard deviation of 0.00104%. In average it is a Shannon-Entropy of 0.9794 out of 1.0. This means it is very difficult to forecast the start-up value of the SRAM. Therefore, it is possible to use the SRAM as a key generator which can regenerate the same key every time due to its high number of stable cells over the whole measurement. This also means that it is not necessary to protect saved keys against attacker and against damage. Every time the SRAM starts up the key is regenerated. Thus, the SRAM only have to be activated if the key is needed and therefore, it saves energy. Furthermore, if the SRAM is active it does not consume a lot of energy due to its architecture. Other works, such as (Lipps, et al., 2018) have also shown a high resistance against environmental influences for the start-up values, which is important in industrials exercises.

5. Conclusion and future work

This work introduces a method to implement low power computing and therefore energy saving but nevertheless sound and secure communication between IIoT devices. The *Master-Slave* implementation is a well-known structure and this can be used to expand the presented cryptographic system. Due to this structure, it is possible to use a hybrid cryptographic system without their energy intensive key exchange and therefore change the keys with a high frequent. Without a growing square number of keys by devices, it is also not a big deal to integrate more devices. By using SRAMs to generate the symmetric likely key improve this method by its PUF characteristic further. No key has to be saved anywhere even not by the devices their self and therefore no key protection is needed, what improves the security and decrease the effort of the implementation and computing tasks. Furthermore, a big advantage of using SRAMs instead of some other PhySec types is that, nearly all IIoT devices included one and has not to be implemented extra. SRAM PUFs are therefore low-cost solutions and can also be implemented in existing IIoT tasks. The high Shannon-Entropy of 0.9794 out of 1.0 and the high stability greater than 94% of the Start-Up Values, are shown in this work and underline the appropriateness suitability of SRAMs as a PUF. There are also no requirements of the size or the fabricant, as long as the SRAM is a six Cell CMOS and an eight Bit Word architecture. To put it in a nutshell a hybrid master slave cryptographic method based on SRAM PUFs is an energy lightweight and cost-effective solution for secure communication in IIoT applications.

Future works have to examine how many energies this method saves in comparison to others and how safe the communication is. The exact implementation of the present cryptosystem is also a further challenge as well as design a suitable function to generate the helper data and the ciphertext. This function has to be invertible and optimized for IIoT devices but must confirm the requirements of security. Another important research is the effect on SRAMs by its aging for years to test the lifetime of such a system.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS0932), IUNO Insec. The authors alone are responsible for the content of the paper.

References

Ambekar, A., Schotten, H. D. & Kuruvatti, N., 2012. *Improved Method of Secret Key Generation Based on Variations in Wireless Channel.* Vienna, Austria.

Böhm, C. & Hofer, M., 2013. *Physical Unclonable Functions in theory and Practice.* 1sth Edition Hrsg.Springer.

Bonneau, J., Herley, C., van Oorschot, P. C. & Stajano, F., 2012. *The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes.* IEEE Symposium on Security and Privacy.

Chze, P. L. R. & Leong, K. S., 2014. *A secure multi-hop routing for IoT communication.*, IEEE World Forum on Internet of Things (WF-IoT).

Farooq, M., Waseem, M., Khairi, A. & Mazhar, S., 2015. A Critical Analysis on the Security Concerns of Internet of Things (IoT). *International Journal of Computer Applications*, 111(7), p. 1–6.

Gao, M. & Qu, G., 2014. *A highly flexible Ring Oscillator PUF.* 51st ACM/EDAC/IEEE Design Automation Conference (DAC).

Gao, Y. et al., 2019. *Building Secure SRAM PUF Key Generators on Resource Constrained Devices.* International Conference on Pervasive Computing and Communications Workshop.

Gassend, B., Dwaine, C., van Dijk, M. & Srinivas, D., 2002. Silicon physical random functions. *Proceedings of the 9th ACM conference on Computer and communications security*, pp. 148-160.

Guajardo, J., Kumar, S. S., Schrijen, G.-J. & Tuyls, P., 2007. *FPGA Intrinsic PUFs and Their Use for IP Protection.* International Workshop on Cryptographic Hardware and Embedded Systems.

Halak, B., 2018. *Physically Unclonable Functions - From Basic Designn Principle tp Advanced haradware Security Apllications.* 1st Edition Hrsg. Southampton: Springer.

Handschuh, H., Schrijen, G.-J. & Tuyls, P., 2010. Hardware Intrinsic Security from Physically Unclonable Functions. In: *Towards Hardware-Intrinsic Security.*Springer , pp. 39-53.

Herder, C., Yu, M.-D., Koushanfar, F. & Devadas, S., 2014. Physical Unclonable Functions and Applications. *Proceedings of the IEEE*, 102(8), p. 1126–1141.

Hoeller, A. & Toegl, R., 2018. *Trusted Platform Modules in Cbyer-Physical Systems: On the Interferences Between Security and Dependability.* London, UK , IEEE European Symposium on Security and Privacy Workshops (EuroS&PW).

Holcomb, D. E., Burleson, W. P. & Fu, K., 2009. *Power-Up SRAM State as an Identifying Fingerprint and Source of True Random Numbers.,*IEEE Transactions on Computers.

Katsikeas, S. et al., 2017. *Lightweight & secure industrial IoT communications via the MQ telemetry transport protocol.*, IEEE Symposium on Computers and Communications (ISCC).

Kusters, L. & Willems, F. M. J., 2019. *Secret-Key Capacity Regions for Multiple Enrollments With an SRAM-PUF.* , Transactions on Information Forensics and Security .

Lee, Y. s., Alasaarela, E. & Lee, H., 2014. *Secure Key management Scheme based on ECC algorithms of Patinet's Medical Information in Healthcare System.*, IEEE.

Lipps, C., Krummacker, D. & Schotten , H. D., 2019b. *Securing Industiral Wireless Networks: Enhancing SDN with PhySec.* Mauritius, Conference on Next Generation Computing Applications (NextComp).

Lipps, C., Strufe, M., Mallikarjun, S. B. & Schotten, H. D., 2019. *PhySec in Cellular Networks: Enhancing Security in the IIoT.* Coimbra Portugal, Proceedings of the 18th European Conference on Cyber Warfare and Security.

Lipps, C. et al., 2018. *Proof of Concept for IoT Device Authentication based on SRAM PUFs using ATMEGA 2560-MCU.* South Padre Island, Texas, USA, 1st International Conference on Data Intelligence and Security (ICDIS).

Maes , R. & Verbauwhede, I., 2010. Physically Unclonable Functions: A Study on the State of the Art and Future Research Directions. In: *Towards Hardware-Intrinsic Security.*:Springer, pp. 3-37.

Maes, R., 2013. *Physically Unclonable Functions - Constructions, Properties and Applications.* Berlin, Heidelberg: Springer-Verlag.

Nishimura, H., Omori, Y., Yamashita , T. & Furukawa, S., 2018. *Secure Authentication Key Sharing between Mobile Devices Based on Owner Identity.*

Plaga, S., Niethammer, M., Wiedermann, N. & Borisov, A., 2018. *Adding Channel Binding for an Out-of Band OTP Authentication Protocol in an INdustiral Use-Case.* South Padre Island, Texas, USA, 1st International Conference on Data Intelligence and Security (ICDIS).

Raza, S. et al., 2012. Secure communication for the Internet of Things—a comparison of link-layer security and IPsec for 6LoWPAN. *Security and Communication Networks,* 12(7), pp. 2654-2668.

Rührmair, U. et al., 2013. *Optical PUFs Reloded.*

Schneier, B., 2015. *Applied Cryptography: Protocols, Algorithms and Source Code in C.* 20th Anniversary Edition Hrsg.:John Wiley & Sons Inc.

Suh, G. E. & Devadas, S., 2007. Physical Unclonable Functions for Device Authentication and Secret Key Generation. *DAC* .

Weinand, A., Ambekar, A. & Schotten, H. D., 10.11.2017. *Providing Physical Layer Security for Mission Critical Machine Type Communication,*

Zenger, C. T. et al., 2015. *Exploiting the Pyhsical Environment for Securing the Internet of Things,* Proceedings of the 2015 New Security Paradigms Workshop:

Zhang, Z.-K.et al., 2014. *IoT Security: Ongoing Challenges and Research Opportunities.* s.l., 7th International Conference on Service-Oriented Computing and Applications.

Blockchain Based Voting Systems

Yash Soni, Leandros Maglaras and Mohamed Amine Ferrag
De Montfort University, School of Computer Science and Informatics, Leicester, UK
Guelma University, Department of Computer Science, Guelma, Algeria
p17237732@alumni365.dmu.ac.uk
leandros.maglaras@dmu.ac.uk
Mohamed.amine.ferrag@gmail.com

DOI: 10.34190/EWS.20.122

Abstract: Voting is an essential tool for any democratic government. It is the most important factor that makes the government for the people, by the people. It is surprising that many countries in the world haven't moved on from the traditional paper ballot system of voting. There are some countries that have taken a forward step by using Electronic voting machines. As simple as it may sound, there are a lot of issues with these systems. Between having to physically go to the venues and standing in a long queue at the polling station to counting the vote by hand the entire process is time-consuming at the same time expensive. Malicious actors have many avenues to influence paper ballot-based elections using intimidation at the polling booth by capturing the ballot box or stuffing the box by fake votes. Similarly, in Electronic voting machines, a highly motivated group can get hold of the machine and change the chip so whichever button is pressed, all the votes go to one political party. Another issue is whether it is impossible to keep track of the votes. How can we be sure that our vote was counted or perhaps in case of a ballot paper, not thrown away? The present article discusses ways of bringing voting systems into the 21st century by using Blockchain Technology.

Keywords: Cyber Security; Blockchain Technology; Electronic Voting; Digital Transformation.

1. Introduction

Today the world around us is rapidly changing. Artificial Intelligence has been making huge waves in today's industries. It has forced people to realign their priority. At the core of all of this buzz, there is data on which all the sophisticated algorithms train on. Since the advancement in AI, Data has become the new oil. Companies that consume public data like Facebook, Instagram, Twitter, Snapchat or e-commerce sites like Amazon, Flipkart and Alibaba control enough of market capital to influence the outcome of any event or perhaps shape people's' opinion. There are several cases where the data centers of these social media giants were breached and profile details of the millions of users were leaked Maglaras et al. (2018). Voting systems around the world have come under scrutiny where there are several cases of fraud reported. Because of this people tend to lose faith in the Electoral Process. Violence, corruption, riots are often things that go hand in hand during the Election process. Due to this, there is a social imbalance amongst society.

For decades we have been elected our leader by participating in the democratic festival of voting. Each country has its own way of electing the members of its parliament. Elections are held every 4 or 5 years. In countries like United States, Germany, Spain, Japan, and Israel elections are held every four years while in some democracies like the United Kingdom, India, South Korea, Canada, and Ireland there is a five-year term for the elected leaders. Various voting methods are deployed around the world. Some of the most common systems are,

- First Past the Post System
- Instant Run-off Method
- Two Round Run-off
- Proportional Representation

Every system has certain advantages and disadvantages attached to it. We will discuss all the systems in-depth and try to figure out which system should be preferred for a blockchain-based Voting system Hj´almarsson et al. (2018). We will discuss all the systems in-depth and try to figure out which system should be preferred for our blockchain-based Voting system. Before we dwell into these systems let's assume that there was an election in our local constituency and after counting the votes, our four candidates,

- Mary got 35
- Bob got 16
- Jane got 15
- Danny got 34

We will discuss each system and see how the elected candidates vary as we change the way leaders are elected.

1.1 First Past the Post System

In this system, the candidate that has received the most votes amongst all the candidates is declared the winner. There is no minimum cut-off that needs to be breached in order to be declared the winner. So, as per our example, Mary will be elected the leader of our constituency. However, this system is at times not fair to the people who voted in the process. In the 2015 Indian General Elections, the winning party got 31% percent votes. Similarly, in the 2016 United States Presidential Election, Donald Trump received 3 million fewer votes than his opponent Hillary Clinton. One can wonder how this system is fair if the majority of the people did not vote for the elected person. In our example, 65% of people did not have Mary as their first choice but now she is going to be their leader.

Another disadvantage of this system is that people would generally vote for the candidate that seems to be winning and not the candidate who they think deserves to win. The reason behind this is extensive media coverage, print interviews that everyone thinks that apart from the top two or three candidates, voting for any other would not make a difference and thus, be a waste of a vote. This discourages independent voters, the democratic voice within the societies to gain strength. There are people who want to improve the condition around them, but as they are not affiliated with the major parties, they get crushed due to the popularity of their rival. Moreover, voter's also voting not thinking of a candidate should win rather a candidate should not win. This can also be termed as negative voting. In our example, let's assume that supporters of Jane dislike Danny. So, to stop Danny from winning, seeing that Jane does not have a clear chance of winning, they might instead vote for Mary cause her chances of winning are more and thus forsake their choice. This can be termed as the voting of lesser of the two evil. Apart from the winning candidate, there are other candidates who came in third or later positions whose votes didn't make a major difference. All those votes cast were in a way wasted.

However, as we only must count the votes one and get with them, often it seems that the results are declared quickly in this process. There are more chances of a single party forming a government rather than multiple parties forming a coalition government. In terms of economic strength, a government with the full majority is considered much more stable rather than a coalition government. This method is used in major Democracies like United States, United Kingdom, India, etc.

1.2 Instant Run-off Method

In this method, instead of voting for their favorite candidate, the voters rank their candidates in terms of preference. This way the candidates choose the voters instead of popularity voting can vote for the leader they believe in. After counting, if no candidate gets more than 50 % voting, then the candidate who got the least number of votes, all the voters whose first choice was that candidate, now their votes are redistributed keeping their second preference in mind. In our example, we can see nobody got more than half of the votes. So, in this scenario, Jane's votes will be redistributed among the rest of the candidates based on voters' second preference and see if anyone has breached the half mark. If not, we repeated the process until someone does. This method has a higher correlation with the sentiment of people and the outcome of the Electoral process. The minimum barrier for a candidate to cross in order to be declared as a winner is decided depending on the size of continency and eligible voters. From the countries' perspective, the benefits of this system are that there will be fewer chances of negative voting. If we see from voter's perspective, then they don't have to vote in desperation rather vote the candidate whose ideology you believe in. In a scenario where you want a third candidate to lose, you can just put him down your ranking order. So, this way the disadvantages of the First-Past-The-Post system are eliminated. This type of voting system is used in many countries for electing the leader of their country like Australia, Sri Lanka, India (Election of President, where only the members of parliament vote), etc.

Many countries have modified this method to suit their system. In Sri Lanka, The President is elected through a variation of the Instant Run-Off Method called Contingent Voting. The voters only select their top three or top four candidates from the list. The winner is decided amongst those three of four candidates. Interestingly even in the most popular Hollywood Awards, The Academy Awards, the winner is decided through the Instant Run-Off Voting method.

1.3 Two Round Voting System

In most of the Presidential Countries like Russia, France, and South American Countries they employ a system called Two Round Voting System. As the name suggests, the Electoral voting process is divided into two rounds, if in the first round the candidate with majority vote has not crossed halfway mark. Then, upon completion of the first round of voting, the top two candidates are pitched against each other in the next round, which decides the eventual winner. This system often shares all the disadvantages of the First-Past-The-Post system. In fact, in some countries, only two candidates are selected from a constituency so to avoid going to the next round of voting. So, the regional parties or new parties are often crushed by the heavyweight National Parties. They never get to reach the top or provide a different, fresh narrative to the Nation. In the end, the voters find themselves voting for the candidate whose chances of winning are more. This way it encourages use of Parties to spend a lot of money marketing, advertising their candidate to popularize their candidate. This is often linked to parties receiving money from illicit sources and encourage corrupting in the process.

So, in our example, After the first round of voting, no candidate has received more than 50 percent of votes. The election goes to the second round for that constituency where people now only have two candidates to choose from, Mary and Danny. This way the Electoral process can go over several weeks.

1.4 Proportional Representation

In an ideal democracy, the percentage of votes a party gets should equate to the number of seats it wins in the parliament. This method is called the Proportional Representation System. It is a common system in the National Legislation of around 80 countries. Belgium was the first country where this system was first implemented in 1900 General Elections. There are multiple ways of implementing this system. The most common two are,

- Pure Proportional Representation.
- Open and Closed List System

In countries like Israel and the Netherlands, the entire country is one constituency and not split into different constituencies. People cast votes keeping Political parties in mind and the vote percentage it receives, gets proportionally converted to the number of seats in the Parliament. This called Pure Proportional Representation. In some countries, the parties publish a list of candidates and people cast their votes based on this candidate. This is a closed list system. Contrary to this, in some countries when the parties publish the list, people also rank these candidates and which type of list is published and what percentage of votes should be given to what candidate on the list. This is called an open list system. So, in the earlier example wherein the 2015 Indian General Election, The Bharatiya Janta Party won 31% votes and formed the government by winning 336 seats out of 545 seats in parliament, by using First-Past-The-Post system, they would have only won 169 seats if Proportional Representation system was considered instead.

It's not like there are no disadvantageous of this system. The critiques of this system say that the people lost connection from their leaders on the local level. When having a broader view from the National Perspective, it looks good, but from the ground level, people fail to associate with their local representatives. Further, as the seats are distributed as per vote percentage, there are very high chances of Coalition Government formation. This is not encouraged and often leads to ending the term of the government before it should have. Keeping all this in mind, in countries like Germany, instead of using any one system, they have implemented a mixture of the above methods. They use a combination of the First-Past-The-Post system and Proportional Representation. In their National Legislation, 50 percent of seats have Proportional Representational and rest have First-Past-The-Post system. This way the advantages of both the side is balanced perfectly. Any country is free to adjust the system as per their needs. For example, In Greece, the party that wins the majority, is awarded 50 additional seats just because they won the majority.

So, in conclusion, we can see that for different systems, the elected leader varies. Some systems have their advantages while they have little correlation to people's sentiment on the ground, while others have a better connection to the Electorate but the process could be really long. However, with Blockchain voting, we can digitalize the entire Electoral Process and thus can opt for longer election process as less effort will go for voters to cast their vote again and at the same time we can have a mandate the maximum people agree with the cause, in the end, Democracy is of the people, by the people!

2. Voting Methods

There are numerous ways in which a person can vote in a democracy. Over the years, many countries have upgraded from simple ballot paper-based elections to electronic machines to elect their representatives. The most common used methods are as follows,

- Ballot Paper
- Electronic voting Machine
- Vote over Internet
- Biometric voting Machine

2.1 Ballot Paper

One of the simplest forms of election mechanism is writing your preference on a piece of paper and submitting it. This is called voting by Ballot Paper. It has a table to the first column containing the name of the party contesting the election and name of the candidate and another column which is kept blank for the voter to mark their choices. The rules are pretty simple. You can put a cross sign in front of any candidate of your choice and fold your paper and submit it into Ballot Box. The Election Authority in charge of overseeing the election must take care of the Ballot Box and see that it isn't tampered with. The Advantages of using a Ballot are as follows.

- It is one of the oldest forms of voting. Takes very little effort to explain to people how to cast their vote and easy to understand. There are many people in rural areas who are not connected with technological advancement in the world. For them, even understanding basic electronic devices might be difficult and thus, prone to mistakes. Thus, keeping it simple pen and paper-based voting will be easy to explain and at the same time, the entire process is quick.
- The voter can have full faith that his vote is listed correctly as there is no scope to correct the vote once crossed.
- Once the vote is cast, it stays in the ballot box for as long as kept safe. Unlike software-based voting where it might lose data if there is a power loss. In many countries, the Electoral processes are held in the summer period. Electronic machines can lose more power as they heat up fast and might malfunction at times. So, the risk for data loss is high.
- There is little to no automation while counting the ballot paper votes, thus it can create employment opportunity for short term for many people in the voting constituency.

As much as the Ballot Paper are praised for its merits, there are certain limitation to this system as well. Few of the most common ones are,

- As there is little to no scope of automation, it might take hours to get a result out. There has to be a human to verify the votes at the end.
- In a country where the Electoral Authority is not independent and Strong, there are chances that someone can tamper with the ballot box by inserting bogus votes in the box. No way of knowing whether the votes were cast by a real person or not.
- As they are piece of paper, it is not environment friendly to have so many papers with little use to be implemented.
- Paper is at the end of the day, flammable and thus can be subject to a fire which might risk losing entire votes in the box. Moreover, guarding paper ballots is cumbersome as they can be easily stolen away.
-

2.2 Electronic Voting Machines

Citing all the negative reviews from ballot paper, many countries upgraded their voting mechanism with advancement in embedded technology with machines called Electronics Voting Machines.

2.2.1 Structure

EVMs are electronic devices which are programmed only once when it is manufactured. The chip is loaded with the software and then it can't be reprogrammed. EVMs consist of two units. There is a control unit where the voter is presented with options of all the candidates to for. There is a button corresponding to the name of the candidate. Connected to this device, by a 5-meter cable, is another embedded device that collects all the electronic votes is Electronic Balloting Unit. EVMs don't require an external power supply as it runs on a standalone battery. They are light in weight, portable and easy to set up in any condition.

2.2.2 Working

The Electoral Authority in charge of overseeing the election presses the Ballot button on the control unit which initializes the EVMs. Then, voter pressing any button corresponding to the candidate he wants to vote for, the machine will light a LED light corresponding choice for the voter to be assured that the machine has recorded the vote correctly. Subsequently, the machine locks itself. Now it can only be unlocked by a new ballot number which the person in charge will press again when a new person goes to vote. This will ensure that there are not multiple votes from one person.

2.2.3 Advantages

- Lightweight, portable and runs on 6-volt alkaline battery. So, it can be taken to place where there is no electricity supplies.
- The votes are stored electronically, so no wastage of paper which eventually becomes waste after the elections.
- Votes are stored in the Electronic Balloting unit, which means it will help speed up the counting process.
- Illiterate people find easy to press a button corresponding their choice rather than using pen and paper.

2.2.4 Disadvantages

- In ballot paper, the most criminal thing that can happen is that entire ballot box is captured. Those votes are anyway wasted. This can be easily be tackled with. However, in case of EVMs, if anyone can get his hand on the device, he can manipulate with the machine by replacing the chip and commit a higher degree of sophisticated crime which might not even be detected.
- As the votes are electronically stored, if anything is to happen to the balloting unit then all those votes could be in danger. There are many ways where the embedded device memory can be corrupted with.

2.3 Vote over Internet

In 2005, Estonia pilot voting via the internet in local elections. During those days, the citizen had a national ID card with was embedded with a special chip and pin which concealed the details of the voters. Voters login to the website using National ID and cast his votes from anywhere in the world. To keep the government from knowing who voted for which candidate, when the voting data is stored in a database, the identity of the voter is sealed and thus anonymity of the voter is fully taken care of.

Such a system provides voting hassle-free and in the comfort of our house. However, unlike any internet-based service, even this platform has severe critique to point our discrepancy in its working. The government has called a team of cybersecurity experts from all over the world to test its voting platform time and time again. Often the team conducting tests has come to the conclusion that not only they can change the vote tally but also remove any trace of them being there. No digital fingerprint will be left in the system that can be traced back to them. This is caused by a lot of rift within the people. The founding idea of democracy was justice for all and Voting is one of the key festivals of democracy. So, it is the governments' job to ensure that the citizen has full trust in the system and if any issue arises, they are taken care of.

2.4 Biometric Voting

There are often cases where a lot of bogus votes have been cast. This is because once the votes got through, to keep the identity of voters private, data of voter is removed. Instead, some African Countries tried a different way of tackling this problem of bogus voting Yinyeh & Gbolagade (2013). They introduced a fingerprint-based voting system where there is a biometric machine that helps identify the person. In many studies, it is proved that a person’s fingerprint is unique. This can be used to identify the person at the time of voting. The whole process gets over quickly with less hassle. The government has to set up a national voter registration program where they register everyone with their fingerprints and make a final list of all the legible voters.

The advantages of such systems are:

- Quick and hassle-free voting
- Can help reduce illicit voting Disadvantages
- As all of this machines will use enormous electricity to run continuous, it will be difficult to use in rural areas.

In Conclusion, while developing a Blockchain-based system, we might come across a similar problem of authentication as the voters can vote from anywhere in the world. Thus, we may use of Biometric sensors or even Face Recognition software to ensure the voters that cast their votes are genuine. Moreover, In the online-based system, there are often chances that the hacker might perform the cult so professionally that it won’t be easy to easily trace it back to the original culprit. Thus, all of these things need to be kept in mind while developing a blockchain voting system. Even though blockchain is a distributed system, we can’t let even one of our nodes be under attack as it might let people question the entire process and, in a democracy, such speculation shouldn't be there.

3. Blockchain Voting System

To have a voting system based on blockchain, there needs to be certain do's and don'ts that needs to be considered Kshetri & Voas (2018). After evaluating the online voting system in general, following consideration needs to be considered.

- Voter's identity, in any circumstances, should not be traced back from the votes. The system should provide a robust and secure way of only voters to check whether their correct votes have been registered or not.
- There should be complete transparency in the voting process. Any voter should be able to check for authenticity of the entire process without risking everyone's privacy.
- Stakeholder's who will form the nodes in the system, their sizes in terms of capital, needs to be restricted so vote tampering is avoided. This way, big conglomerates, corporations, business are not able to influence the results.
- System should make sure only eligible voters vote in the electoral process. System should consider that a voter registered in one constituency is not able to vote in multiple constituency. Otherwise, duplication of votes will result in inaccurate results.
- There needs to be an Authority formed within the stakeholders that oversees any disputes that might result during the Electoral process. The final decision of the authority should be taken as ultimate verdict and should be respected by all.

3.1 Voting System Flow and Voting Machinery

The voters, on successful registration, will log in to the system to cast their votes. They will use the vote token generated at the time of registration for verification at the time of voting. As they have enough network currency to vote only once, they will go to their respective constituency and cast their vote. To do so, they can use Computers, Mobile Phones, or Government approved devices. Once their votes are validated, their votes are added to the blockchain to inform a new block. Upon successful casting the vote, the system generates a transaction ID to the voter which can be used totally his vote anytime as long as the network is active.

One of the key points of Blockchain-based voting is that the tallying of votes is done at the time when the votes are cast Shahzad & Crowcroft (2019). As the entire process is digital, the votes are added to the respective candidates tally as they are cast, Unlike in the conventional voting system where the counting of votes takes several days. On the other hand, when having to count the votes physically, the Election Commission needs to set up a team, provide security to the ballots and to establish dedicated personnel at the counting of votes Ferrag et al. (2018). All of this can be avoided and the capital can be invested in developing the network and protocols Kumar (2018).

3.2 Security concerns

Any human-made system is prone to errors and in a situation where thousands of people are coming out to vote, some issues are bound to arise. In such a scenario, all the stakeholders in the system form a committee amongst themselves to address such concerns. The Election Commission can introduce a few nodes in the network whose sole purpose is to oversee the whole transaction validating process. They have the right to read the ledger but not write to it. This team can be formed from within the stakeholders or it can be a set of individuals who have no connection from the stakeholders. This way they will not be influenced to overlook any error. In spite of being a robust system, there are other several problems that might arise during voting.

3.2.1 Double Spending

In a digital world, it easy to duplicate anything. It's as easy as pressing two buttons at a time. So when it comes to currency, one of the concerns was what if the same thing happens. In the real world, we have physical cash, so when we hand it to the vendor, it is physically transferred. In the digital world, its just bits that are been transferred. So, it can be copied and rebroadcasted. This can be avoided by a procedure where we wait for a certain time before adding a block to the blockchain. In such a case when a sender tries to send the same coin two different addresses, both the transaction will be sent for verification, the one where the nodes verify the transaction first will go through and other transactions will be termed invalid. The same things can be used in or system where before adding a block to the final chain we wait around 10 mins.

3.2.2 51 % Attack

One of the main reasons why blockchain is trending recently is due to its nature. It is distributed and decentralized. So, every node in the network is independent of each other. If one of the nodes goes down, it will still perform. All the records in the blockchain are immutable as to add an incorrect record to blockchain the hacker would need control over half the nodes in the network to have consensus over the phony transaction. This is the practical world is nearly impossible. No single entity has enough computational power to pull this off. So, keeping this in mind, it is in our benefit to have as many stakeholders in our system as possible so to keep the network from coming under attack.

3.2.3 Authentication Flaw

As the election is happening digitally, the voters can vote from anywhere. Be it from the comfort of their homes, or at a polling booth organized by one of the stakeholders. One issue that might arise and needs to be taken seriously in rural areas with people less connected with the technology, they can be influenced by rigging the voting process. People can vote on their behalf once they have authenticated themselves. This needs to be taken into consideration by all the stakeholders and Election Authority and identify areas where something like this likely to happen.

4. Discussion and Conclusions

Various conventional voting systems around the world were studied. Most countries have common voting systems, like the First-Past-The-Post system, Proportional Representation, Instant-Runoff or Two round voting system Wu & Yang (2018). Every country modifies the base system as per their needs Carr III et al. (2018). Many researchers have been asking the question of whether Blockchain can provide an answer to challenges typically in the area which have suffered due to lack of efficient technologies for modern voting systems around the world. This article addresses this in three distinct phases.

There is a phase before voting was keeping track of who is registered and where who is allowed to vote is harder than it should have been. We have a lot of examples where the voter showed up on the day of the polling booth only to find out their names missing from the list. A lot of paperwork goes missing, people get removed from voting rolls, etc. A public ledger that tracks who is registered to vote and legitimate voters can register from their home machine and confirm that before they showed up at the pole would help a lot in reestablishing trust in the system and validating that whoever is eligible to vote is allowed to vote by providing independent verification of that.

The second phase is to actually vote. Now there is always an issue with digitizing government processes. We have come across stories where databases were compromised and private information was leaked or changed. This can be very challenging when it comes to voting, especially voting for the leader of the country. By introducing the Blockchain framework, not only we can reduce corruption in the Electoral process but actually help researchers, scientists, teachers empower so they can help bring net generation up to the challenges of the emerging world. We keep them as nodes in the network that will verify all the transactions. This way voters can trust the outcome of the result more and the system will be more transparent Zhang et al. (2018).

The Final phase is Counting the Votes. The Election Authority has taken the total that is coming from all the different sources like, postal ballots, vote by mail and the actual machines. There we have to keep faith in the system to count all the votes correctly and report it nationally with accuracy. In many countries, the bases of this trust are violated. There is not a lot of confidence that the total from a polling place is accurately summed up. So, using a distributed ledger, not to track the individual votes but to track the totals from each of the polling places will reassure the voter that the local polling didn’t overcount the votes. The voter can verify his votes from the transaction ID provided by the system to actually check whether his votes were counted correctly or not.

References

Carr III, L., Newtson, A. & Joshi, J. (2018), Towards modernizing the future of american voting, in ‘2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)’, IEEE, pp. 130–135.

Ferrag, M. A., Derdour, M., Mukherjee, M., Derhab, A., Maglaras, L. & Janicke, H. (2018), 'Blockchain technologies for the internet of things: Research issues and challenges', IEEE Internet of Things Journal 6(2), 2188–2204.

Hj´almarsson, F. Þ., Hreiðarsson, G. K., Hamdaqa, M. & Hj´almty`sson, G. (2018), Blockchain-based e-voting system, in '2018 IEEE 11th International Confer- ence on Cloud Computing (CLOUD)', IEEE, pp. 983–986.

Kshetri, N. & Voas, J. (2018), 'Blockchain-enabled e-voting', IEEE Software 35(4), 95–99.

Kumar, D. D. (2018), 'Secure electronic voting system using blockchain technol- ogy', International Journal of Advanced Science and Technology 118(1), 13– 22.

Maglaras, L. A., Kim, K.-H., Janicke, H., Ferrag, M. A., Rallis, S., Fragkou, P., Maglaras, A. & Cruz, T. J. (2018), 'Cyber security of critical infrastructures', Ict Express 4(1), 42–45.

Shahzad, B. & Crowcroft, J. (2019), 'Trustworthy electronic voting using ad- justed blockchain technology', IEEE Access 7, 24477–24488.

Wu, H.-T. & Yang, C.-Y. (2018), A blockchain-based network security mecha- nism for voting systems, in '2018 1st International Cognitive Cities Conference (IC3)', IEEE, pp. 227–230.

Yinyeh, M. & Gbolagade, K. (2013), 'Overview of biometric electronic voting system in ghana', International Journal of Advanced Research in Computer Science and Software Engineering 3(7).

Zhang, W., Yuan, Y., Hu, Y., Huang, S., Cao, S., Chopra, A. & Huang, S. (2018), A privacy-preserving voting protocol on blockchain, in '2018 IEEE 11th In- ternational Conference on Cloud Computing (CLOUD)', IEEE, pp. 401–408.

The Future Direction of Cybercrime and the Difficulties of Digital Investigations: A Rationale for a Review of Digital Investigation Specialist Education

Angela Mison, Gareth Davies and Peter Eden
University of South Wales, Treforest, Pontypridd, UK
angela.mison@southwales.ac.uk
gareth.davies@southwales.ac.uk
peter.eden@southwales.ac.uk

DOI: 10.34190/EWS.20.095

Abstract: The Context for this paper is the law of England and Wales. Historically, admissibility of evidence resulting from digital forensics has been seen as a major element in the failure of successful prosecution of cybercrimes and computer enabled crimes (UNODC, 2016). Having identified the common law relating to improperly obtained evidence and its admissibility, the search for the real reasons for such failures revealed that the struggle between individual right to privacy -v- investigatory access to digital information, jurisdiction, standardisation of methods, reproducibility, reliability of findings of digital forensics investigators and the lack of awareness among digital forensic practitioners of their place and responsibilities in the criminal justice system have major interlocking roles. As such, when examining data traces in the Cloud, consideration must be given to: the way the Cloud is owned, controlled, and regulated by companies and governments, by judges, to the demonstrable professionalisation of digital forensics investigation, and by digital forensics investigators, knowledge that potential jurisdictional, international law and Conflicts of law, issues permeate all investigations. The paper questions if current training offers suitable approaches for the training of future digital forensics investigators. The hypothesis is that specialist education must change to keep pace with developments and usage as exemplified by the rapid growth in digital service provision.

Keywords: Cybercrime, education, investigations, Cloud, OCG, CraaS

1. Introduction

The nature of crime and terrorism in cyberspace, the methods of investigation of same, and intelligence gathering are changing radically. The provision of digital services has fuelled a mass movement away from the localisation of data held and processed in desktop computers and laptops with ever increasing capacity hard drives. The provision and availability of high broadband speeds, the norm of app usage, identity obfuscators, and the acceptance that services, applications and storage are virtual, potentially limitless, remote from the individual user's location, ready to be accessed on tap, has contributed to an increasing need for a new type of digital investigators, forensic or otherwise, as organised crime groups, criminal individuals, nation states and terrorists follow, and in some cases lead, the trend. These investigations may be exacerbated for digital investigators by devices syncing with IoT devices on a seemingly random basis; the information an investigator is looking for may be held anywhere.

The hypothesis presented is that it is possible to divide the class of investigator needed for a particular crime according to the ‘business’ operational characteristics of the perpetrators. Naturally, there may be hybrid operations, thus exposure to a number of crime classes becomes necessary during education, permitting the investigator to specialise in the latter part of the course.

2. Digital investigations in the real world

The characteristics of information systems and devices are dictated by the way business is effected. The Locard exchange principle indicating that one cannot in a place without leaving a trace, and that place will also leave its own trace, underpins forensic science. This is even more relevant in the time of the Internet of Things (IoT).

2.1 Traditional crime

Not all digital crime is cybercrime. Computer-enabled crime is an instance where economies of scale can be achieved by means of digital technology. Taking the England and Wales Police Forces as an example, they are tasked with investigating traditional, volume crimes of murder, vehicle crimes, and theft.

Seemingly networking is ubiquitous, and the general population may not be as in control their digital devices as they believe. Prioritisation of investigation must take place. In discussion with one police force representative, it transpired that, although in nearly every crime there was an associated digital device which was required to be investigated, 85% of investigators' time was spent on investigating indecent images of children (IIOC), where the investigative imperative was to identify the victims and stop the abuse; the emphasis was not on prosecution.

Technologically, traditional crime is usually a fairly static operation, involving use of the Cloud, often without the knowledge or consideration of the user, deriving from the use of apps. The rewards to the individual may vary from low to extremely high, depending on the product, service, market demand or energy of the entrepreneur, but the identifying characteristic is the organisational and operational base.

The investigator skill set required for traditional crime includes an understanding of how the crime might be carried out. The investigational aspect involves an understanding of the operational model and discovering the matching artifacts to support it, without being so directed that exculpatory evidence was missed, and that bias or preconceptions obscures other possibilities.

There is a danger that such individual entrepreneurs may become enmeshed with organised crime groups who are looking to outsource aspects of their business. It is arguable that the County Lines drug operation, in which children are trafficked to sell drugs throughout the country, may be a subcontract operational model of a larger group looking to distance themselves from a distribution network.

Some 'sub-contractors' can lose their independence as operators and become coerced or compelled into continuing action. Examples of individuals becoming enmeshed with organised crime may be represented by wittingly or unwittingly in a supply chain, e.g. distribution of fake pharmaceuticals or hiring of a small haulage business or individual driver for movement of a cargo for a single leg of a journey. Having become involved once, it may not be possible to refuse further contracts.

There may not be continuing criminal intention. A compelled sub-contractor may remain silent through fear, and if willing to provide any information, may give rise to a substantial protection cost. The lack of a willingness to bear a protection cost will negate trust in law enforcement.

2.2 Organised Crime Groups

Organised Crime Groups (OCG) present as a comparatively known quantity and mode of operation; they are multinational organisations, availing themselves of IT personnel and sophisticated technology in the manner of any of the Fortune 500. Their mode of investigation is well established having both a financial and a communications aspect.

The communication across groups, as revealed by societal networking research, is rare at other than the top of the hierarchy where it is high value and uncommon, and at the bottom where the group interacts with the market for its services, and there is a degree of acceptable expendability. Middle level members of the OCG rarely have business communications traversing the boundaries of the organization and thus, as enabling functionaries, they have a high level of protection.

It is worth noting the scale of operation of Organised Crime Group which in 2018 was estimated globally to be $600 billion or 0.8% GDP. This is up from the 2014 figures of $328 billion and 0.5% of GDP. The top 20 to 30 OCG act on a nation state level. (McAfee LLC, 2018)

The annual revenues of the five largest crime groups in the world as at 2014 were $27.4 billion (Matthews, 2014) whilst total law enforcement spend for Interpol, Europol, FBI, England & Wales was $17.375 billion (Interpol, 2014), (Europol, 2014), (Full Fact, 2014), (US Dept of Justice, 2014)

Legalisation of some drugs represents a diminution in OCG business and with the rise of Crime as a Service, groups acting as disruptors to other business lines, the future for OCG is not as certain as it once was. There is discussion of partnerships being formed between establishment OCG and 3rd world. (Burbank, 2018).

OCG investigation entails international cooperation (Eurojust, 2019). The imperative in this type of crime is to 'follow the money'. Organised crime groups are long term investors in established businesses. The aim of investigators may be limited to inflicting economic damage on a part of their business. (Eurojust, 2018) (FBI, n.d.).

Cloud usage is anticipated to be private with hosting of associated groupings as necessary. The IT structure is that of a typical large-scale business. The OCG will use experts in their fields, effect risk assessments and business continuity plans and employ encryption as a norm. The infrastructure may be fairly static and thus traceable. With the development of technology, mobility of systems is a reality and identification of the location is imperative.

Much of the type of training required for the investigation of OCGs is found among Forensic Accountants as evidenced by investigation of money laundering and tax evasion programs. Investigations by Her Majesty's Revenue and Customs (HMRC) date back in style to Al Capone. They may also involve the investigation of the creation and use of tax avoidance schemes and shell companies. These complex investigations require knowledge of: international law, tax treaties, company formation tracking of data location, identification of societal networks, data science and analytics techniques.

2.3 OCG v Crime as a Service

Cybercrime as a Service (CraaS) is a disruptor to OCG. For a disruptor, consider traditional retailers who failed to innovate and lost market share to out of town big box stores, who in their turn failed to innovate from bricks and mortar to the on-line environment and failed to consider pop-up stores. A disruptor in any market can cause surprising and catastrophic failures among established regimes.

The original model for CraaS was that of a prime contractor, where a business is set up for a specific task or contract, typically by a small, cohesive, anarchic, technically sophisticated group with complementary skills. The prime contractor/entrepreneur/group of entrepreneurs outsourced or employed the skill sets they needed on a quasi-contractual, short term basis, to do specific tasks.

In the manner of terrorist cells, pseudonymous contractors, may never meet or connect to anyone other than the hiring function – which is often from the dark web. The contractor is unaware of the context of the task. The gig economy has arrived in the dark web

As an evolution of the prime contractor model, current CraaS is the development of a method for execution of a crime which, rather than being executed by the developers, is sold as a product. The sale model may be absolute, commissioned, franchised or licence based, and hence as a Service.

CraaS is likely to present as an unknown and potentially unique investigation in every case. On examination there may be a common mode of operation for a particular series of crime, indicating a successful sales campaign by the CraaS group to multiple other criminal groups.

2.4 Crime as a Service

Traditional organised crime groups may lose business to CraaS disruptors, or they may choose to innovate, a survival tactic, in particular areas or outsource the relevant crimes, currently perceived as mainly financially motivated and based, to CraaS as a partner/prime contractor. The ability of OCGs to enforce CraaS contracts and agreements may be limited due to the structure and organisation of the CraaS groups. The very skill set, anonymity and transience of the CraaS grouping are likely to be their protection.

The identification of the characteristics of CraaS or entrepreneurial crime groups are determined from the way they carry out their business. Modes of operation were determined from the examination of eDedix, Avalanche and GozNym crimes – it should be noted that reports of these are public domain, thus the current mode of operation may have evolved further. The model for CraaS has similarities with cyberwarfare and terrorism. Future serious cybercrime is likely to be based on an initial hacking event.

Table 1: Comparison of the characteristics of OCG v CraaS

Characteristic	OCG	CraaS
Comparative Model	Standing Army	Guerrilla Army
Strategic view	Long Term	Short term
Flexibility/Speed of change	Static	Dynamic
Hierarchy	Structured	Entrepreneurial - flat
Employee base	Large	Small
Technology use	Traditional	Innovative
Disruptor	Borders	Amazon
Business	Vertically integrated supply chain	Data driven - Exploitation of technology requiring initial malware

3. Cybercrime, the need for international cooperation and specialist investigatory teams

The emergence of cybercrime as a major force has caused Law Enforcement Agencies such as the FBI to acknowledge that they must rely increasingly on international collaboration, bi-lateral treaties, informal agreements, and the private sector. Including the statement:

Eurojust states:

> *The very nature of cyberspace means that cybercrime is borderless. Consequently, international measures are required to address the current challenges. (Eurojust, 2019)*

Cybercrime often requires an active initial event by the perpetrators which establishes a forward base for subsequent actions. Tracing the source is an imperative. For digital investigators, the potential mobility of users and their data across multiple digital service providers on a rotating basis, renders time of the essence in acquisition of location and data for examination.

The skill set required for the investigation of cybercrime is primarily based around identification of the crime, locating the original source, the executing source, and incident response rather than digital forensic investigation per se. Only when the primary servers are located can any traditional digital investigation take place, other than on a victim's device. It would appear that there is a great overlap between training of penetration testers/hackers, and the exponents of CraaS, an understanding of whose psychology is required as much as the traditional investigative training.

3.1 Recent-ish examples

If the example of GozNym is taken, the nature of law enforcement and investigation in cyberspace becomes apparent. Arguably, the digital expertise required needs to reside in the investigatory team.

GozNym, an operation created by 10 people, resulted in $100m theft from banks. The CraaS organisation behaved as a prime contractor, buying in the services it needed: running money mule networks, spammers, coders, and organisational and technical support. The investigation involved international cooperation from six Law Enforcement Agencies, Europol, and Eurojust. Searches were made in Bulgaria, Georgia, Moldova and Ukraine, and prosecutions brought in Georgia, Moldova Ukraine and the United States. (Bank Info Security, 2019) (Europol, 2019). The mindsets of digital investigators and digital forensic investigators are demonstrably different in this example.

'Avalanche', the hosting service used by the GozNym was itself a specialist criminal hosting company, reputedly bullet proof, operated by a key group comprising of 5 people, utilising approximately 39 servers. It also specialised in the deployment of botnets, malware and ransomware. It was taken down during 2016. (Eurojust, 2016)

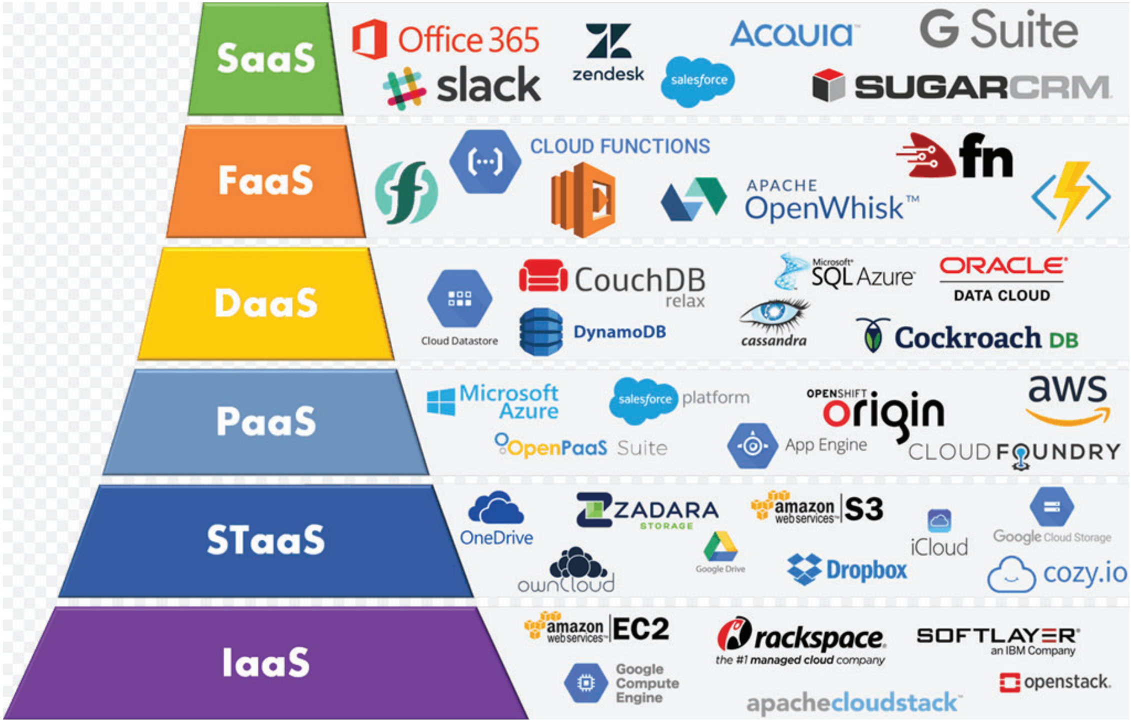
4. What is the Cloud and who owns it?

The Cloud is, now, too specific a term that should be expanded to Digital Service Providers (DSP). Statistics show that there has been a technological revolution resulting in the move from traditional models of computing to Cloud based computing, apps and IoT. Currently, investigation of the Cloud is taught, and forward-thinking Course Leaders already have included later technological developments in their teaching. The very nature of DSP presents investigatory problems to digital investigators.

When speed is of the essence, unknown digital structures and responsibilities are like a brick wall. Such is the Cloud. Discovery, or application for information, rejection, and lengthy fulfilment times delay the onset of the investigation proper. In cases where there is risk to life, it is not surprising that digital investigators sometimes act prematurely, thus commit crimes in the race to assist a victim. For the criminal, the Cloud can represent a dense layer of obfuscation which provides the luxury of time.

Rarely are the organisations and infrastructure underpinning the Cloud and service offerings documented sufficiently to permit a speedy, legally appropriate application, in the right jurisdiction, and to the right organisation which has access to the information required. That the Cloud infrastructure may be comprised of additional sub-contractors, who themselves sub-contract is not discernible from the service offering documentation. Unless there is experience of a particular service offering, the starting point can be guesswork, along the lines of ‘who owns it?’, ‘where are they located?’, ‘where do they operate?’, ‘what information do they hold?’, ‘who do I need to contact?’, ‘what paperwork do I need to put in place?, and ‘who needs to sign the paperwork?’.

An example of the complexity facing digital investigators is given in the following diagram.



©IMELGRAT.ME

5. Legal Education for Digital Investigators

Much more legal education for digital investigators is required. The need for legal education applies equally to digital investigators and digital forensic investigators. The bounds of what is technically possible far outstrip the bounds of what is legal.

5.1 Lawful Authority

In England and Wales, the police operate according to the doctrine of ultra vires. Simply put, if there is not a piece of legislation which permits what they are doing, legally, they cannot do it. Having lawful authority, the police may authorize digital investigators. If the police do not have lawful authority an investigator following through on a received instruction commits a criminal offence under the Computer Misuse Act 1990 (HMG, 1990).

5.2 Jurisdiction

Most of what an investigator will be instructed to effect will carry lawful authority and be legal, domestically in England and Wales, but where apps and the Cloud are involved, transborder issues arise and international cooperation must be sought and obtained. It is irrelevant what domestic legislation purports to permit and render legal. No statute, by itself, can authorize illegal actions in a foreign jurisdiction. A digital investigator may become criminally liable before even being aware of the transborder nature of his investigation.

There is current discussion that investigation may be legal when limited to local data on a device at hand. That examination may be legal. Consideration of syncing devices, without the express consent of the owners of the device and account holders, may represent a criminal action in a foreign jurisdiction. It is well to note that many countries consider access without consent of the device owner or account holder to be unauthorised and thus a criminal act.

5.3 Evidence

Statistically, the greatest reason for failure in any digital crime rests with inadmissibility of evidence, followed by difficulties in confirmation of attribution. These are areas that would benefit from greater emphasis in digital investigation courses. It is possible that there is some reluctance within the criminal justice system to recognize the professionalism of investigators practicing what is perceived to be an arcane art form.

5.3.1 Evidence -v- Intelligence

To define the distinction between evidence and intelligence, it is necessary to consider the different modes of operation between the digital investigator and the digital forensic investigator and to recognise their differing imperatives. With an increase in cybercrime and CraaS, it is likely that intelligence will be the output of on-going digital investigatory teams. That intelligence would act as circumstantial evidence to inform other, traditional investigators.

5.3.2 The digital investigator

A digital investigator may act under the pressure of time and not adhere to such rigour as is demanded by the criminal justice system for formal presentation of his findings in Court. Much of what is found may be circumstantial. The difficulty of attribution of activity to a specific end user or identified person remains. For the purposes of the investigation, attribution may be assumed.

Much of a digital investigator's talent may be employed in international investigations. In consideration of that, it is well to consider the job specifications and requirements applicable to potential colleagues, as although they may have the same job title in translation, empowerments may vary.

5.3.3 The digital forensic investigator

As with the digital investigator, additional concerns arise in an international investigation. Not only may empowerment be different, but processes and Standard Operating Procedures may be regulated by International Standards which it is intended enable free transfer of evidence without the need for re-examination to comply with domestic jurisdiction requirements. This compliance may not be achieved by the use of Standard Operating Procedures carried out against different International Standards in a different jurisdiction, even though such rigour may render outputs admissible as evidence in the domestic jurisdiction.

6. Law enforcement v Privacy

6.1 Privacy

People frequently give their personal data freely to unknown organisations on the internet. Those organisations should have privacy policies, but these important documents which advise what the organisation can and may do with that information go largely unread as the user ticks a box to obtain the functionality required. After much consideration, the Data Protection Act 2018 (DPA, 2018), based on the principles of informed consent, sought to control what organisations could do with personal information and the penalties for its misuse and abuse.

From the law enforcement perspective, the Data Protection Act (*ibid*) applies equally. Additional statutes also apply to items and data under investigation.

6.2 The effects of Human Rights Articles on forced consent

In England and Wales, legislation has attempted to simplify gaining access to devices and systems during a criminal investigation. Unfortunately, the law of unintended consequences applies.

Criminal law is vested in the presumption of innocence. Failure to provide information in response to a production order or disclosure requirement finds, without requirement for proof, the person who is being questioned, to be guilty of a second and unconnected offence.

The action of demanding answers under threat of automatic conviction of an unconnected offence may amount to oppression by the State, akin to the description in Article 3 Human Rights Act 1998 (HRA, 1998). The person provides the information under duress. This is not a purely academic argument. There is precedent (Emmerson & Ashworth A, 2001) 15-92 citing Heaney and McGuiness v Ireland. Injudicious use of the forced disclosure clauses may render international cooperation problematic.

7. Where should Digital investigation education go?

Digital investigation has become a discipline wherein technically aware problem solvers who, having received an investigatory instruction, promptly attempt to find what has been asked of them and present their findings coherently, are highly prized. Such capability and competence are rare talents in an occupation with a currently high and increasing demand for its skills.

Digital investigators must be self-monitoring, able to distinguish their permissible technical capabilities up to the boundaries of the legal and recognise their potential for criminal liability. Such boundaries may be stretched, but should never be compromised in the search for international cybercriminals or cyber terrorists if there is to be a successful prosecution. Investigators are bound by rules which do not constrain the behaviour or actions of criminals.

Although cybercrime, as identified here, is largely a product of OCG and CraaS, the two need very different approaches and mind sets. Cyberterrorism would seem closely related to CraaS.

- The OCG investigatory approach is a serious fraud/money laundering approach which can be lengthy, structured, and demonstrate a requires concentration of knowledge on anti-forensics and the high level of security employed by perpetrators.
- The CraaS investigatory approach owes more to hacking, can be short term, has sometimes, scant regard for territorial boundaries, and an imaginative approach to tracking down perpetrators. The investigators may become adept at recognising signatures, thus attribution of CraaS to particular groupings.
- Final year degree courses may present a range of options related to the specialisms required, which in themselves may change over time. Degree and/or other apprenticeships could be considered.

With the potential increase in international cooperation, courses may be designed which operate in the manner of International MBAs. This elite may be taught from a virtual campus via collaborating and cooperating experts in their fields, drawn from several universities, an example of which is Decamp – Open Distributed European Campus (Decamp, 2018), initially part-funded by the EU Erasmus+ programme. On-going professional development and research are required to ensure that investigators remain current in their skills and knowledge.

Those with lesser academic ambitions, but demonstrable skills should not be discouraged from participation. It should be recognised that some students who have the appropriate aptitude or mind set may not wish to undertake examinations. Having completed the practical courses necessary and demonstrated their capability and competence they could, if mentored, exhibit the same professionalism as their examined colleagues. Among these students may be a cadre of those who exhibit autistic-like syndromes which researchers have identified in the majority of those committing juvenile computer abuse offences.

Outreach programmes, talent spotting school children is underway. Tarian, the Southern Wales Regional Organised Crime Unit, operate a Prevent programme, through which they hope to prevent juvenile criminal offending by those with over-enthusiastic curiosity about what is possible with a computer, turning that curiosity to good purpose.

There is, however, a continuing problem in the digital investigation sphere. It is a time sensitive, results-based profession. Some UK graduates have identified difficulty in gaining their first employment without experience. The investigation system tends to cycle its employees almost as a closed user group rather than increase its pool from an available, qualified but semi-trained population. Organisations must accept an element of in-service training. In demanding qualifications and specific experience from applicants, the industry is its own worst enemy in the maintenance of skills shortages, and one must ask whether it is not from economic self-interest.

8. Conclusion

There are numerous choke points, a collection of interlocking factors related to the acquisition of data that include: the specific creation of offences complete with considered lawful authority clauses, legislation granting powers of investigation, location of the crime, location of the consequences of the crime, identification of relevant criminal legislation, identification of perpetrators, location of perpetrators, location of any infrastructure, determination of jurisdiction, international cooperation, contention between an investigator's needs and a user's privacy, Human Rights Conventions, foreign jurisdictions' laws and/or Constitutions….. Many are bureaucratic, but nonetheless, necessary to ensure both the avoidance of criminal liability for the investigator and ultimately the admissibility of evidence

In seeking international cooperation, there are issues concerning the conflict of laws, knowledge of international law, local law and constitutional law of relevant within a foreign jurisdiction. A particular issue relates to the structure and location of the elements which underpin a digital service provision. Much time may be spent requesting, from the wrong entity, information they do not hold. A detailed knowledge of the DSP would enable the consideration of applications to be made in more cooperative jurisdictions.

Domestic law cannot legitimise an investigator's transgressions in a foreign jurisdiction; neither can it compel a third party in a foreign jurisdiction to break his domestic law or breach his Constitution. An appreciation of this is particularly relevant with respect to digital service providers, many of whom are based in the United States, and who have both legal and Constitutional constraints on their actions.

With a necessary increase in international cooperative investigations, cultural sensitivities should be respected, and lessons learned. Networking across borders should be encouraged and periods of time in international organisations such as Interpol and Europol supported. Common International standards for the transfer of information and evidence should be developed to avoid the repetition of investigation in order to comply with domestic forensic requirements.

There is a requirement to revise and develop specialisms within the education of digital investigators. The final thought is left that the best investigator or teacher may possibly be a poacher turned gamekeeper, with or without a criminal record.

References

Bank Info Security, 2019. *FBI and Europol Disrupt GozNym Malware Attack Network.* [Online] Available at: ttps://www.bankinfosecurity.com/fbi-europol-disrupt-goznym-malware-attack-network-a-12493 [Accessed 10 August 2019].

Burbank, J., 2018. *The World's top five mob bosses.* [Online] Available at: https://themobmuseum.org/blog/worlds-top-five-mob-bosses/ [Accessed 11 August 2019].

CMA, 1990. *Computer Misuse Act.* London: Her Majesty's Government.

Decamp, 2018. *Decamp.* [Online] Available at: mydecamp.eu/partner-universities/ [Accessed 12 February 2020].

DPA, 2018. *Data Protection Act.* London: Her Majesty's Government.

Emmerson, B. Q. & Ashworth A, Q., 2001. *Human Rights and Criminal Justice.* 1st ed. London: Sweet and Maxwell.

Eurojust, 2016. *'Avalanche' dismantled in international cyber operation.* [Online] Available at: http://www.eurojust.europa.eu/press/PressReleases/Pages/2016/2016-12-01.aspx [Accessed 10 August 2019].

Eurojust, 2018. *Coordinated Crackdown on 'Ndanghetta Mafia in Europe.* [Online] Available at: http://eurojust.europa.eu/press/PressReleases/Pages/2018/2018-12-05b.aspx [Accessed 10 August 2019].

Eurojust, 2019. *Euto 24 Million cryptocurrency theft unravelled with Eurojust support.* [Online] Available at: http://eurojust.europa.eu/press/PressReleases/Pages/2019/2019-07-05.aspx [Accessed 10 August 2019].

Eurojust, 2019. *Setting the scene on cybercrime: trends and new challenges.* [Online] Available at: http://www.eurojust.europa.eu/press/PressReleases/Pages/2019/2019-07-05.aspx [Accessed 10 August 2019].

Europol, 2014. [Online] Available at: https://www.europol.europa.eu/publications-documents/europol-budge..t [Accessed 29 November 2019].

Europol, 2019. *GozNym malware:Cybercriminal network dismantled in international operation.* [Online] Available at: https://www.europol.europa.eu/newsroom/news/goznym-malware-cybercriminal-network-dismantled-in-international-operation [Accessed 25 July 2019].

FAS, 2000. *Chapter V Future of International Crime* [Online] Available at: https://fas.org/irp/threat/pub45270chap5.html [Accessed 10 Auguest 2019].

Full Fact, 2014. *Police funding in England and Wales.* [Online] Available at: https://fullfact.org/crime/police-funding-england-and-wales/ [Accessed 29 November 2019].

HRA, 1998. *Human Rights Act.* London: Her Majesty's Government.

Interpol, 2014. *Our Funding.* [Online] Available at: https://www.interpol.int/Who-we-are/Our-funding [Accessed 29 November 2019].

Matthews, C., 2014. *Fortune 5: Biggest organised crime groups in the world.* [Online] Available at: https://fortune.com/2014/09/14/biggest-organized-crime-groups-in-the-world/ [Accessed 11 August 2019].

McAfee LLC, 2018. *The economic impact of cybercrime - no slowing down.* [Online] Available at: https://www.mcafee.com/enterprise/en-us/assets/executive-summaries/es-economic-impact-cybercrime.pdf [Accessed 12 February 2020].

UNODC, 2014. *Five Largest Organised Crime Groups.* [Online] Available at: http://fortune.com/2014/09/14/biggest-organized-crime-groups-in-the-world/ [Accessed 17 December 2018].

UNODC, 2016. *Lessons learned UK.* [Online] Available at: https://sherloc.unodc.org/cld/lessons-learned/gbr/most_common_legal_and_practical_obstacles_to_the_successful_prosecution_of_cybercrime_acts.html?&tmpl=cyb[Accessed 7 July 2019].

US Dept of Justice, 2014. *Budget Fact Sheets.* [Online] Available at: https://www.justice.gov/about/fy17-budget-fact-sheets[Accessed 29 November 2019].

US Dept of Justice, 2016. Justice News 7 Dec 2016. [Online] Available at: https://www.justice.gov/opa/speech/asssistant-attorney-general-leslie-r-caldwell-delivers-remarks-highlighting-cybercrime [Accessed 25 July 2019].

Hybrid Warfare in Society's Information Environment: An Empirical Analysis Using the Grounded Theory

Erja Mustonen-Ollila[1], Martti J. Lehto[1]and Jukka Heikkonen[2]
[1]University of Jyväskylä, Finland
[2]University of Turku, Finland
erja.mustonen-ollila@quicknet.inet.fi
martti.lehto@jyu.fi
jukka.heikkonen@utu.fi

DOI: 10.34190/EWS.20.006

Abstract: This paper investigates hybrid warfare by analysing its dimensions, that is, hybrid influencing, information warfare, kinetic warfare and cyber warfare, in a society's information environment. The Grounded Theory approach was used to collect and analyse the data. Evidence of four thematic categories emerged, and each category was broken down into multiple items, derived from data and validated by past studies. The number of discovered item observations was 119, that is, 1–20 multiple item categories per thematic category, forming a total of 44 item categories. In this study, four thematic categories, their item categories and relationships between thematic categories defined the concept of hybrid warfare. Propositions regarding the thematic categories were then made on the basis of the inter-relationships between them. Seven higher-level abstractions of statements were found in the conceptual framework. Finally, six conclusions emerged from this study. First, modifying a society's IE for a foreign state's purposes is a coordinated political goal. Second, external pressure from foreign states is constant and seeks new forms of influencing. Third, a strategic goal exists to weaken society's IE through frequent cyber operations, cyber attacks, fake information, social influence, political influence and psychological influence in order to prevent the progress of the society. Fourth, the goal of military manoeuvres and constant exercises with new military weapons near borders is to threaten the society. Fifth, foreign actors use different types of warfare to prevent the society's decision-making. Sixth, hybrid influencing, information warfare, kinetic warfare and cyber warfare all overlap with each other in the society's IE.

Keywords: Society, Information Environment, Hybrid Warfare, Hybrid Influencing, Warfare, Grounded Theory

1. Introduction

Hybrid warfare means a violent conflict that combines a range of different dimensions of war (military, economic, information and cyber), tactics (regular and irregular) and actors (state and non-state) (Hoffman, 2009; Scheipers, 2016, p. 1) using information warfare, psychological operations and propaganda for strategic aims (Scheipers, 2016, p. 1; Renz, 2016; Bachmann and Gunneriusson, 2015).

Encountering hybrid warfare in the information environment (IE) makes identifying and analysing the different types of hybrid warfare within IEs essential. This study defines an IE as 'Information, aggregate of individuals, organizations and systems that receive, collect, process and convey/disseminate the information, or act on information, and the cognitive, virtual and physical space in which this occurs' (NATO 2012, p. 2). Its focus is on a society's IEs, society being defined as 'a group of individuals involved in persistent social interaction, or a large social group sharing the same geographical or social territory, typically subject to the same political authority and dominant cultural expectations' (Society, 2018).

Past studies (Lehto et al., 2017; NATO StratCom COE, 2016; Sartonen, Huhtinen and Lehto, 2015; Geers, 2015; Armistead, United States and Joint Forces Staff College, 2004, pp. 13-20; Hoffman, 2009) have shown that hybrid warfare consists of hybrid influencing and different types of warfare in an IE, but qualitative research on the origins of hybrid warfare and how a society becomes aware of it is lacking. In order to obtain a clear understanding of hybrid warfare's impact, it must be examined in a real society, and the nature of hybrid warfare must be described in detail by stydying the hybrid influencing, information, kinetic and cyber warfare that the society has faced, the extent to which influencing and different types of warfare are shaped by the IE context, and how the influencing and these different types of warfare are inter-related.

This qualitative in-depth case study applies past studies and empirical evidence (Yin, 2003) and identifies four thematic categories, that is hybrid influencing, information warfare, kinetic warfare and cyber warfare in a society's IE. The collected data were analysed using the Grounded Theory (GT) approach and a conceptual framework was developed using thematic categories, item categories and the relationships between these (Glaser and Strauss, 1967).

Each of the four thematic categories was further broken down into multiple item categories by deriving them from the data and finally validating them using past studies. The total number of item observations was 119, and they were categorised using GT analysis (Glaser and Strauss, 1967) and content analysis (Krippendorff 1985).

The hybrid influencing category, for example, economic and political influencing (Secretariat of Security Committee, 2018, p. 29) was the highest-level category, and consisted of 20 different item categories and 57 empirical item observations. The information warfare category, that is, influencing the state's social, political, psychological, economic, and military decision-making and performance, as well as citizens' opinions (NATO StratCom COE, 2016), consisted of nine different item categories and 17 empirical item observations. The cyber warfare category, involving cyber attacks on critical infrastructure and inflicting physical damage and casualties (Sigholm, 2013, p. 52, pp. 68-69), consisted of six different item categories and 12 empirical item observations. Finally, the kinetic warfare category, defined as the physical destruction of a target by military forces, terrorists, paid mercenaries and soldiers without patches or kinetic activities or intelligence operatives working for the government (Renz, 2016; Armistead, United States and Joint Forces Staff College, 2004, pp. 13-20), consisted of nine different item categories and 33 empirical item observations.

The 1–20 multiple item categories in each of the four thematic categories led to a total of 44 different item categories. We found evidence for the propositions regarding the thematic categories and the relationships between the thematic categories. Finally, we found seven higher-level abstractions of statements in the conceptual framework.

The rest of the paper is organised as follows: Section two discusses the related research, Section three presents the research method, Section four outlines data categorisation, and Section five presents the data analysis. Finally, Section six contains conclusions and the discussion.

2. Related Research

Past studies on hybrid influencing have included the corruption of politicians and security areas (Conley et al., 2016); pressure targeted towards a society's vital activities, economy systems, energy supply, and logistic functions (Hollis, 2011; Armistead, United States and Joint Forces Staff College, 2004, pp. 13-20); security, business, intelligence, political, contact and company networks that are guided, owned and funded by a foreign state; opaque foreign state networks influencing Russian-minded national politicians; extremism; capturing a state: changing laws in order to favour a foreign state; modifying the political environment for foreign state purposes; interfering in other states' affairs; and indirect military interfere (Conley et al., 2016).

Past research on information warfare has included hostile influence over the state's social, political, communicative, economic, public, psychological, and military decision-making, performance and citizens' opinions via IE (NATO StratCom COE, 2016); pressure through military power and deterrence (Sartonen, Huhtinen and Lehto, 2015) and Hollis (2011); and command-and-control, economic information and psychological, electronic, cyber, intelligence-based and hacker warfare (Libicki, 1995).

Cyber warfare studies have covered cyber penetration, cyber manipulation and cyber robbery, commonly recognised as war by the international community and defined as such by international legislation (Sigholm, 2013, p. 51; Lehto et al., 2017). Cyber warfare is primarily a military endeavour that includes doctrine, tactics and technologies – military cyber operations (not espionage, however) (Ottis, 2015, p. 89). According to Geers (2015), cyber warfare includes bringing down foreign states' electronic networks and net banks, crippling states and their critical infrastructures, identity theft, propaganda, and the destruction of control systems.

Finally, past studies on kinetic warfare have examined the annihilation of critical infrastructure, soldiers without patches, the creation of confused situations, strategic influence through war, inciting violence in other states (Conley et al., 2016; Hollis, 2011); foreign states' military performance, pressure and regional crises, military manoeuvres, extra forces near borders, increased military intelligence, and military and territorial disturbance of air, sea and land traffic (Lehto et al., 2017; Conley et al., 2016).

Thus, in spite of a large number of high-quality past studies on hybrid influencing, information warfare, kinetic warfare and cyber warfare, the literature has focused on their different types and largely neglected the relationships between these categories. Therefore, this study responds to the need for further research, and offers both practical and theoretical knowledge on influencing the different types of warfare in a society's IE and explores their relationships with each other, by formulating two research questions (RQs): 1) What are hybrid influencing, information warfare, kinetic warfare and cyber warfare in a society's IE?, and; 2) How are hybrid influencing, information, kinetic and cyber warfare in a society's IE related to each other?

3. Research Method

We chose a qualitative case study (Yin, 2003; Creswell, 2007) using the GT approach (Eisenhardt, 1989; Glaser and Strauss, 1967) to enable us to answer our two research questions. The unit of analysis in this case study was limited to one IE of a society, because the goal was to gain a deep understanding of the selected IE and to identify hybrid influencing and warfare issues in this specific site. Nine audio-recorded unstructured and semi-structured interviews between January and May 2018 (Table 1) were conducted that investigated the experiences of the four thematic categories. The interviewees were or had been involved in several civil and military hybrid influencing and warfare issues in their own fields of expertise during their working careers, which extended over a period of 6 to over 30 years in different positions and organisations in Finland and abroad. The interviewees represented eight different organisations in Finland. We also studied archival material, representing a secondary source of data, which included public news and past scientific studies of hybrid influencing and information, and kinetic and cyber warfare in Finland or abroad in general. Triangulation (Yin, 2003) was used to simultaneously combine different data sources to improve the reliability and validity of the data. When the qualitative data reached saturation point, data collection ended. The saturation point means achieving a theoretical saturation point, i.e. even if new data are collected they offer no new knowledge about the studied phenomena (Glaser and Strauss, 1967). In our study, this point was reached when the same categories started to be repeated in the data.

The interview questions were improved after each interview to better suit the next interview, and sometimes due to the schedule of the interviewee, had to be shortened. The longest interview consisted of 70 questions, and the 'shortest' of 30 questions. The interviewees recommended new interviewees, as they knew who should be interviewed. We analysed each interview transcript and identified four major emergent themes and concepts (Myers and Avison, 2002) in order to form thematic categories (Myers and Avison, 2002). The interviewees received the questions before the interviews in order to become familiar with them (Creswell, 2007), and were able to check their content to minimise mistakes.

Themes can also form the main categories or concepts in data. This is a selective way of finding the concepts and categories in the data and is based on the researcher's own intuition or knowledge. The concepts must be categorised according to relevant terminology and theories that form the most refereed work in categorising concepts in the research area. After the categories have been discovered, the number of categories must be decided on. The problem with the categories is whether enough proof can be found in the data to make them and the concepts valid and reliable, and whether the concepts and categories discovered are the correct ones. Some other concepts and categories may emerge from the data later. If the concepts and categories are not correct, the researcher must return to the data and discover new concepts. After the abstract concepts are found, they can be coded according to the instructions of Glaser and Strauss (1967), using selective coding to search the data categories. The abstract concepts can also be found using the content analysis approach (Krippendorff, 1985), which is a text analysis method. This approach requires the researcher to construct a category system, code the data, and calculate the frequencies or percentages that are used to test the hypotheses on the relationships among the variables of interest. It is assumed that the meaning of a text is objective, in the sense that a text corresponds to an objective reality.

The GT approach follows different phases of data analysis and uses content analysis as part of its categorisation method as follows: 1. Identification of thematic categories in the empirical data using content analysis. 2. Definition of the thematic category based on the empirical data. 3. Search for appropriate literature to be used as evidence for the identified thematic category. 4. Search for similar thematic categories in the empirical evidence to enable mutual exclusion (it is not wise to use thematic categories that use the same definition but are labelled (titled) differently). 5. Search for relationships between the thematic categories. 6. Determination of higher level of abstraction of statements about the relationships between the thematic categories, and propositions for the categories. The statements are based on empirical evidence. 7. Creation of a conceptual framework of thematic categories and their relationships in order to visualise results. The final product resulting from creating a theory from the case studies may be a concept, a conceptual framework or propositions, or possibly a mid-range theory (Eisenhardt. 1989; Mustonen-Ollila and Heikkonen, 2009; Mustonen-Ollila et al., 2020). The text is interpreted and understood without extraneous contextual knowledge. In case studies such as this study, the concepts are sharpened by building evidence that describes them. The data and concepts are constantly compared so that accumulating the evidence converges on simple and well-defined concepts, i.e. categories or constructs. In this study, the constant comparison between data and the concepts in past studies, in order to accumulate evidence converged on simple, well-defined thematic categories, led to a higher level of abstraction of statements about the relationships between the thematic categories. This theorising was in line with Pawluch and Neiterman's (2010) suggestions of creating a GT using Glaser and Strauss's (1967) approach. The higher level of abstraction of statements is presented in the conclusions and discussion section. Intuition and knowledge is also used in determining the categories, and a chain of evidence is created: the thematic categories are derived from the empirical data and then validated using past studies. In this study, Pawluch and Neiterman's (2010) GT analysis instructions, together with those of Glaser and Strauss (1967), support the finding of categories from data and their being based on the researchers' own intuition and knowledge.

Table 1: Interviewee details.

Interviewee number	Role of interviewee	Length of interview in minutes	Group or individual interview
1	Chief of Cyber Division	215	Individual interview
2	Military Professor	235	Individual interview
3	Civilian Security Officer	151	Individual interview
4	Civilian Official & Senior Adviser of the Security Committee	135	Group interview
5	University Teacher	31	Individual interview
6	Expert of the Security Committee	66	Individual interview
7	Military Professor	117	Individual interview
8	Officer of the Defence Command Finland	200	Individual interview
9	Researcher (Digitalisation, Cyber Security)	181	Individual interview
Total:		1341	

4. Data Categorisation

Because data should be categorised under several identifiable themes, when collecting the data in this study, we followed the following themes: hybrid influencing, cyber warfare, and information warfare and kinetic warfare. Each of the themes had specific questions. We used past studies to understand the complexity of the thematic categories and reflected them through the data. The definitions of these categories have changed over time and we think that only a highly experienced person in the field would understand them thoroughly and understand their change. We, however, cannot change the content of the thematic categories because they are based on empirical data and verified in past studies.

The audio-recorded interviews included frequent elaboration and clarification of meanings and terms, and the recordings were transcribed, yielding over 240 pages of transcriptions. After the transcription of the interviews, a qualitative research method based on GT (Glaser and Strauss, 1967) and content analysis (Krippendorff, 1985) was applied in order to categorise data under thematic categories, that is, hybrid influencing, information warfare, kinetic warfare and cyber warfare, according to relevant terminology and theories in the studied research area. In this study, hybrid influencing and three types of warfare were denoted as thematic categories (Glaser and Strauss, 1967).

In this study, the sampling was theoretical (Pare and Elam, 1997) which is not the same as probabilistic sampling. Our goal was to obtain a solid understanding of this case study in a society's IE and to discover categories and their relationships (Eisenhardt, 1989). The interviewers were given the opportunity to correct the mistakes or omissions in their own interviews and to contribute to the validity of this research by correcting the definitions of the thematic categories. This data interpretation also improved the reliability of the study (Klein and Myers, 1999). We thus created a chain of evidence in data categorisation and found the hierarchy of the four thematic categories. We then further broke down each thematic category into multiple item categories using content analysis. We derived these items from the data and validated them using past studies. In this specific analysis phase, we used the definitions of the four thematic categories and past studies to determine which items belonged to each specific thematic category. Table 2 below presents an example of an item observation concerning the 'Hybrid influencing' category. In Table 2, the first column contains the definition of the hybrid influencing item derived from the empirical evidence; the second column contains the empirical evidence on the item; the third column contains the evidence from the literature on the item in the second column; the fourth column contains the name of the literature source of the third column; and finally the fifth column contains the transcript number of the empirical evidence.

Table 2: Example of item observation concerning hybrid influencing category.

Item definition derived from empirical evidence: item category	Empirical evidence on item	Evidence from literature on item in second column	Literature source	Transcript number (TCn)
Influencing from outside society	Influencing can be carried out by external actors who are not inside a society. Their goal is to change the state's internal or external politics or status.	Changing laws in order to favour a foreign state, information networks funded by foreign states, changing the political environment to better suit to foreign state and destroying the internal unity of the state.	Conley et al., 2016	TC5

Table 3 shows the four thematic categories and their item categories. The first column consists of item categories for each thematic category; the second column shows the number of the item observations; and the third column shows the item number. Each item category contains a different number of observations.

Table 3: Thematic category of 'Hybrid influencing' and its item categories.

Hybrid influencing consisting of item categories	Number of item observations	Item number
Economic sanctions and intimidation	7	1
Intimidation to overthrow Finnish banking systems	2	2
Pursuit of political power in civil service and ministries	4	3
Political extremists' funding	2	4
Pursuit of power by politicians	2	5
Political and close circle influencing	2	6
Weakening society's coherence	1	7
Influencing citizens and organisations	1	8
Critical comments of Russian minority concerning Finnish society	3	9
Influencing from outside society	3	10
Weakening the credibility of Finnish society through internal changes	2	11
Weakening trust in society	1	12
Law and doctrines	1	13
Attempts to overthrow society and ministry	5	14
Using different identities in hoaxes	1	15
Narratives against Finland	5	16
Own narratives of Finnish council of state	5	17
Influencing the Finnish press	4	18
Influencing the environment and secret strategic influencing	4	19
Public communications	2	20
Total:	57	
Total number of item categories:		20

Tables 4–6 follow the same structure as Table 3. A total of 119 different empirical observations under 44 item categories were found using Glaser's and Strauss's (1967) GT analysis and Krippendorff's (1985) content analysis method.

Table 4: Thematic category of 'Kinetic warfare' and its item categories.

Kinetic warfare consisting of item categories	Number of item observations	Item number
Destroying critical infrastructure	5	1
Spreading radiological substances	1	2
Operative coordination	3	3
Inciting violence in other states	2	4
Strategic influencing using war	4	5
Atypical information on operations and inciting internal riots	7	6
Physical strikes, attacks and special operations by special forces	6	7
Activities of foreign state in Finland as precursor of war	3	8
Physical destruction and its propaganda	2	9
Total:	33	
Total number of item categories:		9

Table 5: Thematic category of 'Information warfare' and its item categories.

Information warfare consisting of item categories	Number of item observations	Item number
Influencing decision-making	1	1
Influencing warfare	3	2
Influencing people	1	3
Influencing with message contents	3	4
Preventing flow of information between networks	1	5
Disseminating false information between networks	1	6
Creating disagreements between networks	1	7
Political influencing	2	8
Modifying the IE	4	9
Total:	17	
Total number of item categories:		9

Table 6: Thematic category of 'Cyber warfare' and its item categories.

Cyber warfare consisting of item categories	Number of item observations	Item number
Cyber operations	2	1
Architecture attacks	3	2
Virus attacks	1	3
Infrastructure attacks	4	4
Cyber intelligence	1	5
Telecommunications attacks	1	6
Total:	12	
Total number of item categories:		6

5. Data Analysis

After data categorisation and data collection, the conceptual framework of the discovered categories was formulated on the basis of the empirical evidence and theories reflecting the findings in the field (Glaser and Strauss, 1967). We followed the suggestions of Glaser and Strauss (1967) to involve fragmentation and reassemble the data into thematic categories by trying to capture a broader conceptual foundation based on the interviewees' experiences and practical knowledge of the four thematic categories.

Figure 1 shows the thematic categories as ellipses, and the relationships between the thematic categories are marked by solid lines with letters (A to F). The small arrows with numbered circles pointing to the thematic categories are the multiple item categories composed by content analysis.

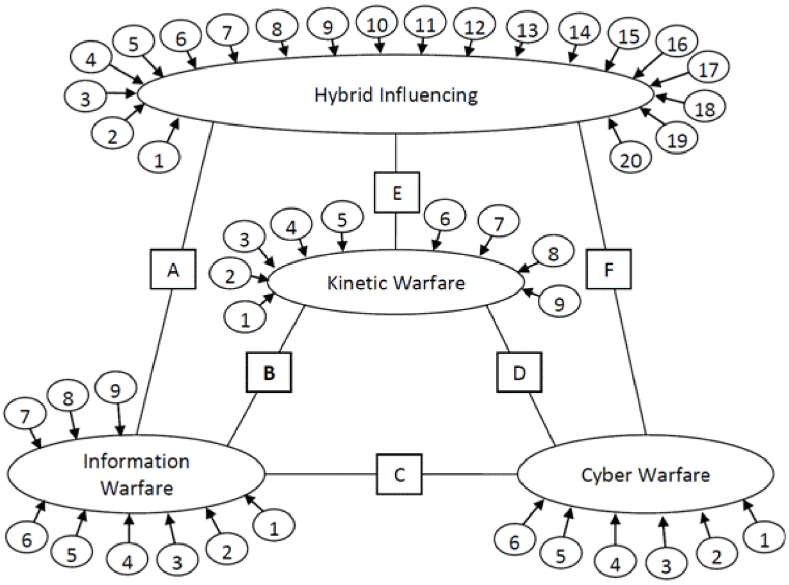


Figure 1: Conceptual framework of four thematic categories and their item categories.

After finding the thematic categories and their relationships, we determined the properties of the thematic categories and propositions (hypotheses) regarding how the categories were related to each other on the basis of the data (Glaser and Strauss, 1967). Table 7 presents examples of the relationships based on the empirical data.

Table 7: Properties of thematic categories and propositions (hypotheses) regarding how hybrid influencing, information warfare, kinetic warfare and cyber warfare categories are related on the basis of the data.

Thematic category/categories	**Properties of thematic categories and propositions regarding how the categories are inter-related (lines marked with letters A to F in Figure 1) on the basis of the data.**	**Lines marked with letters in Figure 1**
Hybrid influencing, information warfare	In principal, all these communication operations are atypical (influence) and may never have been used before (at this level or current scope) and can become a new method.	A
Information warfare, kinetic warfare	Information warfare is when a foreign country influences another country's decision-making by threatening to or actually destroying i.e. a factory.	B
Information warfare, cyber warfare	The goal of information warfare is to destroy the functionality of a factory, not by shooting bullets, but by intimidating with cyber warfare.	C
Cyber warfare, kinetic warfare	We think a lot about the cyber dimension or handling the electric magnetic spectrum. But if it is network warfare, then it is quite close to cyber warfare.	D
Cyber warfare, kinetic warfare, information warfare, hybrid influencing	Hybrid warfare means the following: power proofs, green men, refugees, economic sanctions, propaganda, cyber weapons.	A, B, C, D, E, F
Hybrid influencing, cyber warfare	The company's responsibility is to secure its own information and network environment (information and cyber security) and their boundaries, and to contribute to society's overall security.	F

The data and thematic categories and their item categories were constantly compared with each other, and this comparison led to seven higher-level abstractions of statements about the relationships between the categories. The higher-level abstraction statements are presented in the discussion and conclusions section.

6. Discussion and Conclusions

Based on eight individual and one two-person in-depth group interviews, this study tackled the four thematic categories (hybrid influencing, information warfare, kinetic warfare and cyber warfare) in a society's IE. The GT analysis (Glaser and Strauss, 1967) found their evidence in the society's IE, using the inductive approach. These four thematic categories were improved by evidence which discovered 119 item observations in the collected empirical data. The identified 119 item observations were categorised into different item categories. Glaser and

Strauss' (1967) constant comparison of concepts and data showed evidence of new item categories. The decomposition of each thematic category into multiple item categories through content analysis helped us derive new item categories and validate them using past studies and respected definitions. After determining the 44 different item categories, we defined the properties of the thematic categories and propositions (hypotheses) regarding how the thematic categories were related. Finally, we developed a conceptual framework of the thematic categories and item categories, on the basis of empirical evidence and past studies reflecting the findings of the field. In this study, four thematic categories, their item categories and the relationships between thematic categories defined the concept of hybrid warfare. The comparison with past studies led us to seven higher-level abstractions of statements about the relationships between the thematic categories. This theorising was in line with the recommendations of Glaser and Strauss (1967) on how to create a GT by deriving it from empirical data and letting it emerge from the data.

These seven higher-level abstractions of statements were as follows. 1) A society can constantly be the target of hybrid influencing. 2) Foreign actors have their own political goals to weaken society. 3) Foreign actors want to modify society's IE to exert political influence by carrying out cyber and information operations. 4) Foreign actors give funding to left- or right-wing extremists to weaken society. 5) Foreign actors carry out cyber intelligence actions aimed at Finnish citizens despite this being illegal in Finland. 6) Narratives against Finland are constant and structured, and funding is reserved and other resources available for carrying out information campaigns. 7) Some politicians or other citizens support foreign actors' policies against their own society.

The findings also support the claims of Conley et al. (2016) regarding the corruption of politicians and security; as well as those of Hollis (2011) and Armistead, United States and the Joint Forces Staff College (2004, pp. 13-20) that hybrid influencing is pressure targeted towards a society's vital activities; the activities of a society's economy systems, energy supply, and logistics' functions; and the activities of the security, business, intelligence, political, contact and company networks that are guided, owned and funded by foreign states. The foreign state networks are opaque and influence Russian-minded national politicians, active Russian networks, extremism, and capturing a state. Laws are changed in order to favour a foreign state and information networks (public relations) and are funded by foreign states, modifying the political environment for foreign state purposes, destroying the internal unity of the state, and interfering in other states' affairs and indirect military interference (Conley et al., 2016). The results were also in line with those of NATO StratCom COE (2016), Sartonen, Huhtinen and Lehto (2015), Hollis (2011) and Libicki (1995) in that information warfare takes place at all levels of warfare, and means hostile influence over a state's social, political, communicative, economic, public, psychological, and military decision-making and performance, as well as influence over citizens' opinions via the IE. Our results are also in line with the findings of Sartonen, Huhtinen and Lehto (2015) and Hollis (2011) regarding pressure through military power and deterrence belonging to information warfare. The claims of Sigholm (2013, p. 51) and Lehto et al. (2017) support our findings that cyber warfare includes cyber penetration, cyber manipulation and cyber robbery reaching the threshold of hostilities commonly recognised as war by the international community and as defined by international legislation. They also agree with Ottis (2015, p. 89) that cyber warfare is primarily a military endeavour that comprises doctrine, tactics, technologies, and military cyber operations (not espionage, however) carried out by intelligence operatives and information terrorism to bring down foreign states' electronic networks and net banks and to cripple states; identity theft; propaganda; and the destruction of critical infrastructure and control systems (Geers, 2015).

Finally, past studies on kinetic warfare have shown that annihilation of the critical infrastructure, special soldiers and forces, soldiers without patches, creating confused situations, strategic influence through war, spreading violence from one's own state to other states, military pressure and using military force (Conley et al., 2016; Hollis, 2011) are ways in which to carry out kinetic warfare. Our evidence also confirms the claims of Lehto et al. (2017) and Conley et al. (2016) that one state threat is Russia's military preparedness, the goal of which is to raise its kinetic performance and increase regional crises; to use strategic strikes or strikes that begin as an attack to conquer regions; to make territorial violations via land, sea and air; to make military manoeuvres and use extra forces near borders; to increase military intelligence; and to cause military disturbance of air and sea traffic (Lehto et al., 2017; Conley et al., 2016).

However, this study has some limitations. First, we had limited knowledge on the thematic categories and item categories because we only interviewed 10 people due to research resource limitations. Second, the use of only one society and one IE affected the findings, making it difficult to generalise the results.

Six conclusions emerge from this study. First, modifying a society's IE for a foreign state's purposes is a coordinated political goal. Second, external pressure from foreign states is constant and seeks new forms of influencing. Third, a strategic goal exists to weaken society's IE through frequent cyber operations, cyber attacks, fake information, social influence, political influence and psychological influence in order to prevent the progress of a society. Fourth, the goal of military manoeuvres and constant exercises with new military weapons near borders is to threaten the society. Fifth, foreign actors use different types of warfare to prevent the society's decision-making. Sixth, hybrid influencing, information warfare, kinetic warfare and cyber warfare overlap with each other in a society's IE.

The most important practical benefit of this study was the finding of a vast number of items in practice in society. The managerial contribution lies in making decision-makers in society aware of these categories and items before they make decisions that affect their society's IE. The methodological contribution is the way in which diverse qualitative research methods, such as GT (Glaser and Strauss, 1967), content analysis (Krippendorff, 1985), and rigorously applied methods can be used together to conduct a high-quality literature study (Wolfswinkel, Furtmueller and Wilderom, 2013).

We selected the cases for this study from a society's IE so that they would either predict similar outcomes or produce contradictory results for predictable reasons (Yin, 2003). We applied theory triangulation by interpreting a single data set from multiple perspectives to understand the research problems (Yin, 2003). The categories and their relationships were validated using the GT approach (Glaser and Strauss, 1967; Eisenhardt, 1989). According to Eisenhardt (1989), combining a case study with the GT approach produces a novel, testable theory; an empirically valid theory that emerges from the data, as it did in this study. The operatives and their motives behind hybrid warfare in a society's IE would be an interesting subject for future research.

References

Armistead, E. L. and Joint Forces Staff College. (2004) *Information Operations: Warfare and the Hard Reality of Soft Power,* Washington, D.C.: Potomac Books Inc.

Bachmann, S-D. and Gunneriusson, H. (2015) Russia's Hybrid Warfare in the East: The Integral Nature of the Information Sphere, *Georgetown Journal of International Affairs*, pp 198-211.

Conley, H.A., Mina, J., Stefanov, R. and Vladimirov, M. (2016) *The Kremlin Playbook. Understanding the Russian Influence in Central and Eastern Europe.* A Report of the Center for Strategic International Studies Europe Program and the CSD Economics Program, New York: Rowman & Littelefield.

Creswell, J.W. (2007) *Qualitative Inquiry and Research Design: Choosing Among Five Approache*s, California, U.S.: Sage Publications.

Eisenhardt, K.M. (1989) Building Theories from Case Study Research*, Academy of Management Review*, 14(4), pp 532-550.

Geers, K. (2015) Coder, Hacker, Soldier, Spy In: M. Lehto and P. Neittaanmäki, ed., *Cyber Security: Analytics*, *Technology and Automation,* 1st ed. Dortdrecht: Springer-Verlag, pp 73-87.

Glaser, B. and Strauss, A.L. (1967) *The Discovery of the Grounded Theory: Strategies for Qualitative Research,* Chicago, IL: Aldine.

Creswell, J.W. (2007) *Qualitative Inquiry and Research Design: Choosing Among Five Approaches,* California, US: Sage Publications.

Hoffman, F.G. (2009) Hybrid Warfare and Challenges, *Joint Forces Quarterly* 52/1[online], Available at: https://smallwarsjournal.com/documents/jfqhoffman.pdf Accessed 16 Aug. 2019.

Hollis, D. (2011) Cyber War Case Study: Georgia 2008. *Small Wars Journal* [online], Available at: http://smallwarsjournal.com/jrnl/art/cyberwarcase-study-georgia-2008 Accessed 30 Jan. 2019.

Klein, H. K. and Myers, M. D. (1999) A set of principles for conducting and evaluating interpretive field studies in information systems, *MIS Quarterly*, 23(1), pp 67-94.

Krippendorff, K. (1985). *Content analysis. An Introduction to its Methodology,* California, CA**:** Sage Publications Inc.

Libicki, M.C. (1995) *What Is Information Warfare?* Washington DC.: National Defence University.

Lehto, M., Limnéll, J., Innola, E., Pöyhönen, J., Rusi, T. and Salminen, M. (2017) *Finland's cyber security: the present state, vision and the actions needed to achieve the vision,* Publications of the Government´s analysis, assessment and research activities, Series 30, Helsinki: Prime Minister's Office.

Markus, K.L. and Robey, D. (1988) Information technology and organizational change: Causal structure in theory and research, *Management Science*, 34(5), 583-589.

Mustonen-Ollila, E. and Heikkonen, J. (2009) Historical research in information system field: from data collection to theory creation. In: A. Cater-Steel and L. Al-Hakim, ed., *Information Systems Research Methods, Epistemology, and Applications.* Hersey, New York: Information Science reference (an imprint of IGI Global), 140-160.

Mustonen-Ollila, E., Lehto, M. and Heikkonen, J. (2020). Components of Defence Strategies in Society's Information Environment: A Case Study Based on the Grounded Theory. *Security and Defence Quarterly. Forthcoming.*

Myers, M.D. and Avison, D.E. (Eds.) (2002) *Qualitative Research in Information Systems*: *Review*, London: Sage Publications.

NATO, (2012) *NATO military policy on information operations*. Available at: https://info.publicintelligence.net/NATO-IO-Policy.pdf Accessed 24 Jan. 2017.

NATO StratCom COE, (2016) *Social Media as a Tool of Hybrid Warfare*, Riga: NATO StratCom COE. Available at: http://www.stratcomcoe.org/social-media-tool-hybrid-warfare Accessed 30 May 2017.

Ottis, R. (2015) Cyber Warfare. In: M. Lehto and P. Neittaanmäki, ed., *Cyber Security: Analytics, Technology and Automation*, 1s t ed. Dordrecht**:** Springer- Verlag, pp 89-96.

Pare, G. and Elam, J.J. (1997) Using Case Study Research to Build Theories of IT Implementation, In: *Proceedings of The IFIP TC8 WG 8.2 Conference on Information Systems and Qualitative Research*, Philadelphia, Pennsylvania, pp 542-569.

Pawluch, D. and Neiterman, E. (2010) What is Grounded Theory and Where Does is Come from, In: A. Bourgeault, R. Dingwall and R. De Vries, ed., *The Sage Handbook of Qualitative Methods in Health Research*. London: Sage Publications, pp 174-192.

Renz, B. (2016) 'Hybrid Warfare' as a QuasiI-Theory of Russian Foreign Policy? *Aleksanteri Insight* [online], 22(2), Available at: www.helsinki.fi/aleksanteri/insight Accessed 15 Jan. 2017.

Sartonen, M., Huhtinen, A-M. and Lehto, M. (2015) From influencee to influencer: the rhizomatic target audience of the cyber domain, In: *ECCWS 2015 Proceedings of the 14th European Conference on Cyber Warfare & Security,* Hatfield, UK: Academic Conferences and Publishing International Limited, pp 249-256.

Scheipers, S. (2016) Moral forces in war: Clausewitz and hybrid warfare. In: B. Renz and H. Smith, ed., *After 'hybrid warfare', what next? Understanding and responding to contemporary Russia*, Government's analysis, assessment and research activities, Series 44, Helsinki: Prime Minister's Office, pp 11–13.

Secretariat of Security Committee (2018) *Vocabulary of Cyber Security*, Helsinki: The National Emergency Supply Agency.

Sigholm, J. (2013) Non-State Actors in Cyberspace Operations, In: J. Vankka, ed., *Cyber warfare,* National Defence University, Series 1, No. 34, Tampere: Juvenes Print, pp 47-76.

Society (2018) In: *Wikipedia* [online], Available at: https://en.wikipedia.org/wiki/Society Accessed 18 May. 2018.

Wolfswinkel, J.F., Furtmueller, E. and Wilderom, C.P.M. (2013) Using Grounded Theory as a Method for Rigorously Reviewing Literature, *European Journal of Information Systems*, 22(1), pp 45-55.

Yin, R.K. (2003) *Case study research: design and methods*, California: Sage Publications.

Actors in Society's Hybrid Information Environment: Grounded Theory Analysis

Erja Mustonen-Ollila[1], Martti J. Lehto[1] and Jukka Heikkonen[2]
[1]University of Jyväskylä, Finland
[2]University of Turku, Finland
erja.mustonen-ollila@quicknet.inet.fi
martti.lehto@jyu.fi
jukka.heikkonen@utu.fi

DOI: 10.34190/EWS.20.121

Abstract: This paper outlines the actors carrying out information and cyber operations and cyber attacks in society's hybrid information environment. It conceptualizes these actors using content analysis, as well as a qualitative and Grounded Theory approach by studying primary open sources, such as past scientific articles, government reports, theses, public news, non-scientific (popular) studies, books, websites, TV programmes, blogs, social media outlets, and personal observations of contemporary daily events. The total sum of actor observations was 189 among where information operations, cyber operations and cyber attacks were carried out by 49, 22, and 20 different actors, respectively. The discovered 91 actors are verified and validated. The conceptual model of these actors was developed using past studies and the Grounded Theory approach. This conceptual model acted as a theory on the actors carrying out these operations and attacks in society's hybrid information environment, and is the scientific result of this study. The practical, and managerial result is the revelation of who these actors are. Being aware of which actors are capable of such operations and attacks enables better defence of society's hybrid information environment and improves its security.

Keywords: hybrid information environment, society, actor, information operation, cyber operation, cyber attack, Grounded Theory

1. Introduction

Any study of information operations, cyber operations and cyber attacks must pay careful attention to the actors who *actually* conduct these operations and attacks in society's hybrid information environment. From the perspective of defending and securing this environment, these operations and attacks occur in a real world. They can harm targeted individuals and organizations and test their ability to defend themselves and their security (Giles, 2017). Some of the actors consciously intend to carry out these operations and attacks. Their actions can provide them with guidelines to determine how well the operations and attacks went, what their impact was, and how to improve their abilities to operate and attack again. The actors can serve as triggers to begin operations and attacks in society's other hybrid information environments in a similar way.

In this study, an **actor** is defined as an individual person with specific skills; a formal organizational unit; an organizational group with several different people sharing the same motive to carry out operations and attacks; a state or non-state actor; programs whose implementers are hidden; movements; associations; or religious actors capable of carrying out information operations, cyber operations and cyber attacks. An actor determines the types of these operations and attacks, and their place and time, and must have motive, control, funding and resources to carry them out. An actor can act in specific roles, such a hacker (Holger and Kilger, 2012), terrorist or terrorist group (NATO StratCom COE, 2016a), botnet (Jaitner, 2013), organizations or states (Sillanpää, 2016; Giles, 2017), non-states (Giles, 2017), criminal groups (Hyppönen, 2013), or digital persons or groups born autonomously on the internet (Wihersaari, 2015). The focus of this study when dealing with actors is society's hybrid information environment, in which a **society** is defined as "*a group of individuals involved in persistent social interaction, or a large social group sharing the same geographical or social territory, typically subject to the same political authority and dominant cultural expectations*" (Society 2018).

Past studies have defined the hybrid and information environment as two separate concepts. In Latin, the word hybrid (Hybrid, 2017), as referred to in this study, is a mixture of two very different things. The information environment (IE) referred to in this study is a country's political, economic, cultural, social, scientific, technological, and military structure; the environment of which is deliberately undermined by

various actions in order to achieve a contextual change in the society's political debate and by doing so to influence decision-making (Nissen, 2012). The IE is further claimed to include both information operations (IO) and information warfare (IW) (Armistead et al., 2004, pp. 13-20). This study uses the definition of information operations (IO) presented by Armistead et al. (2004) and Information Operations (2012, p. vii): "*The integrated military operations in which information is used in operations to influence own decision-making and get desired effects on the will or defend itself from enemies', potential enemies', decision-makers', cultural groups', and international communities' information operations and impact*". In this study, civil IO take place in the minds of human beings and are targeted towards the performance of adversaries, for example, communication systems and propaganda (Sirén, Huhtinen and Toivettula, 2011). The social media (SM) environment is defined as forms of electronic communication through which users create online communities to share information, ideas, personal messages, and other content (Social media, 2017), and it can also include military IO (Huhtinen and Rantapelkonen, 2016; Sartonen, 2015).

In this study, cyber attacks can be 'soft', 'hard' and social cyber attacks. Soft cyber attacks are targeted at economic and military information, and hard cyber attacks can be, for example, denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks (Jaitner, 2013; Geers, 2011), damaging government websites and preventing their services, hacking emails and mobile phones, trolling, botnets, manipulation of audio visual material, espionage, identity thefts, propaganda, and destruction of critical infrastructure and control systems. They can be used to cause a diplomatic or political crisis (Nissen, 2012). Social cyber attacks, on the other hand, are targeted towards people (Goolsby, 2013).

Although the actors in a society's hybrid IE (Lehto et al. 2017, pp. 20-21; Winter, 2017; Giles, 2017; NATO StratCom COE, 2016b; Sillanpää, 2016; Lange-Ionatamishvili and Svetoka, 2015; Mäntylä, 2014, p. 16; Hyppönen, 2013, p. 31; Jaitner, 2013, p. 66; Jantunen, 2013; Ottis, 2013, p. 191; Holt and Kilger, 2012; Hollis, 2011, pp. 3-6; Geers, 2011, pp. 41-42) are noted in highly qualified literature, we must refocus on them. There is a lack of qualitative research on how a society becomes aware of who the main actors are who carry out IO, cyber operations and cyber attacks in society's hybrid IE. This is because only then we can improve understanding of actors' abilities to carry out operations and attacks, and help determine what kind of operations and attacks they are capable of carrying out. This study developed a conceptual model of the actors in society's hybrid IE by analysing past studies using the Grounded Theory (GT) approach (Glaser and Strauss, 1967; Wolfswinkel, Furtmueller and Wilderom, 2013). The same actors carried out IO or both cyber operations and cyber attacks, or alternatively IO, cyber operations and cyber attacks. In this study, IO actors are denoted as IOA, cyber operation actors as COA and cyber attack actors are denoted as CAA. IO were carried out by 49, cyber operations by 22, and cyber attacks by 20 different actors. The total amount of different actors was thus 91.

Seven conclusions emerge from this study. First, the majority of IOA are states, politicians, political or non-political organizations, nationalists, patriots, terrorists, media, religious actors, and different movements. Second, the majority of COA and CAA are hackers or hacker groups, criminal groups, actors connected to cyber operations or programs and program armies. Third, the majority of IOA, COA and CAA are intelligence actors, hackers and different organizations. Fourth, politicians and citizens can themselves act as trolls, spreading false information with or without being aware of it themselves. Fifth, society's cyber networks, infrastructures and information systems can be attacked and crippled by programs that are autonomous and think independently. Sixth, operations and attacks can be sold to an outsider, which makes finding the real actor behind them, called an attribution problem, very difficult. Seventh, the actors carrying out the operations and attacks are dangerous due to their international and national connections, relationships and networks. A global network of actors carrying out operations and attacks is an even more severe threat to society's hybrid IE. The rest of the paper is structured as follows: Section two describes the research methodology. Section three presents the data categorization and analysis, and Section four contains the research findings. Finally, Section five contains the discussion and conclusions.

2. Research methodology

This study is a bibliographical, explorative and contextual qualitative case study (Creswell, 2007; Yin, 1993; Pettigrew, 1985) and its goal is to answer the research question: **Who are the actors carrying out information operations, cyber operations and cyber attacks in society's hybrid information environment?**

Yin (2003) claims that triangulation involves checking different data sources simultaneously to improve the reliability and validity of data, which enables the use of multiple data sources to provide a stronger substantiation of concepts. Collecting different types of data from different sources produces a wider scope of coverage and may result in a fuller picture of the phenomenon under study. Eisenhardt (1989) suggests that '*theory development can be carried out by analyzing literature and extending existing theoretical models and using common sense or experience*' (Wolfswinkel, Furtmueller and Wilderom, 2013, p. 8).

The first literature study was carried out using the Google search engine and by defining the following keywords, which are based on the research question: IO actors, cyber operation actors, cyber attack actors, IO in society, cyber operations in society, cyber attacks in society, IO in the IE, cyber operations in the IE, cyber attacks in the IE, IO in the hybrid IE, cyber operations in the hybrid IE and cyber attacks in the hybrid IE. This led to different literature sources, such as newspapers, blogs, internet news, books published in the field, theses, past scientific studies, and government reports. The literature sources abstracts/main points/introductions were read through. Furthermore, daily TV programmes and news were listened every day, and daily text news was read, because it showed relevant information on contemporary events. If the information read and listened to was suitable for the study, it was added to the file of literature source material.

The second literature search was carried out in scientific databases or well-known other sources, such as Doria, Springer Link, StratComCOE, and universities' databases, using the same keywords as those in the first literature review. After this second literature review, the literature sources' abstracts/main points/introductions were read through, more TV programmes and news were listened to, and more text news was read about contemporary events. This information was added to the same file of literature source material as that in the first literature study. After this, it was decided which of the collected literature source material was suitable for this research. 35 studies were chosen, which were read carefully from beginning to end, and these studies covered the literature base of IOA, COA and CAA in society's hybrid IE.

The collected data set consisted of two different types of literature sources. The first included past scientific studies, theses and government reports; and the second included all the other sources mentioned above. The latter literature source was collected between October 2016 and October 2019, and the former between January 2017 October 2019. Both literature source collections occurred almost in parallel, because many issues concerning operations and attacks were happening at the same time, in practice and in science. Using these two literature sources, the recognized operations and attacks were arranged into a chronological table containing one row for each operation event or attack event. Each row included the name of an actor, the definition of the actor based on the literature, the literature source, and actor type (IOA, COA or CAA). Each event was then added to the table in a chronological order based on the information on the IOA, COA and CAA. Similarly, the IOA, COA and CAA were treated as separate if they had carried out different operation and attack events. A condensed version of this table, consisting of 189 rows of data enumerating all operations and attacks (189 separate events) was finally created. Table 1 shows examples of actors, definitions of the actors based on the literature, type of actor (IOA, COA or CAA) and the literature source.

Table 1: Examples of actors, actor definition, actor type (IOA, COA or CAA) and literature source

Actor	Definition of actor based on literature source	Type of actor	Literature source
Influential religious leaders	An important part of the terrorist 'recruitment support strategy' are the speeches of influential religious leaders.	IOA	NATO StratCom COE, 2016a, p. 11
Semantic botnets (botnet army)	A botnet army could be used to a certain target via blogs, posting comments on news articles, sending targeted email, etc. The challenge for the attacker would be to make the communications as realistic as possible while making identity verification a complex and time-consuming challenge.	COA, CAA	Geers, 2011, p. 40
Implementers of blackmail programs	About 10 000 cyber attack trials were prevented (WannaCrypOr-blackmail program). The attackers were able to access an attack code which was leaked from National Security Agency (NSA) due to a Windows operating system's security vulnerability.	IOA, COA, CAA	HS Metrolehti, 2017

3. Data categorization and analysis

In this study, it was vital to recognize the IOAs, COAs and CAAs in society's hybrid IE, because the success of their operations and attacks depends on the maturity of society's IE. A chronological table was created of actors carrying out IO, cyber operations or cyber attacks, which contained 91 different actors. Table 2a, 2b and 2c show the IOA, COA and CAA, and IOA, COA and CAA, respectively, together with their literature source.

Table 2a. Information operation actor (IOA), and literature source

Actor (IOA)	Literature Source
Foreign states	Giles, 2017, p. 2
Foreign leaders	Jackson, 2015, p. 5
Politicians	Sillanpää, 2016
Secret terrorist supporters	NATO StratCom COE, 2016b, pp. 1-2
Hostile rulers	Jackson, 2015, p. 5
Local affiliates	Sillanpää, 2016
Terrorists and terrorist groups	Lehto et al., 2017; Winter, 2017, p. 6; NATO StratCom COE, 2016a; NATO StratCom COE, 2016b; Huovinen et al., 2014
Orthodox church and priests	Sillanpää, 2016; NATO StratCom COE, 2014
Domestic audience and public	Jackson, 2015, pp. 5-7
Trolls	Huhtinen and Rantapelkonen, 2016; NATO StratCom COE, 2016b
Hybrid trolls	Giles, 2017
Radical organizations' supporters	NATO StratCom COE, 2016a
Independent actors	NATO StratCom COE, 2016b
Local media owners	Sillanpää, 2016
Opinion leaders	Sillanpää, 2016
Ideological networks	Sillanpää, 2016
Trolling channels	NATO StratCom COE, 2016b
Media activists	Winter, 2017
Media strategists	Winter, 2017, p. 19
Propagandists	Winter, 2017, pp. 15-19
Media outlets	Sillanpää, 2016
Communist functionaries	Sillanpää, 2016
Sympathizers	Winter, 2017, p. 17
Imams and scholars	NATO StratCom COE, 2016a
Extremist groups	Pomerantsev, 2015, p. 34; NATO StratCom COE, 2016a
Political parties	Jackson, 2015, p. 6
Military elite	Jackson, 2015, p. 6
Micro-communities	NATO StratCom COE, 2016a
Actors in foundations	NATO StratCom COE, 2016a
Governing elite	NATO StratCom COE, 2014
Trolling factories	Lange-Ionatamishvili and Svetoka, 2015
Religious extremists	NATO StratCom COE, 2016a
Nationalistic movements	Laine, 2015
Provocateurs as trolls	Jaitner, 2015
Multi-ethnic nationalists	Laine, 2015
Compatriots abroad	NATO StratCom COE, 2014
Ersatz movements	Laine, 2015
Non-government organizations	Pomerantsev, 2015, p. 30
Non-political organizations	NATO StratCom COE, 2014
Activists in blogs	Jaitner, 2013
Lobbyists	Conley et al., 2016
Religious leaders	NATO StratCom COE, 2016a
Radical nationalist groups	Laine, 2015
Corrupted businessmen	Conley et al., 2016
Corrupted politicians	Conley et al., 2016

Actor (IOA)	Literature Source
Corrupted officials	Conley et al., 2016
Patriotic organizations	Laine, 2015
News agencies	NATO StratCom COE, 2016a
Foreign state-supported media	Giles, 2017
Total:	**49**

Table 2b: Cyber operation actor (COA) and cyber attack actor (CAA), and literature source

Actor (COA and CAA)	Literature Source
Foreign states' hackers	Holt and Kilger, 2012, p. 3, p. 12; Mäntylä, 2014; Geers, 2011
Programmer groups	Holt and Kilger, 2012
Hacker groups and networks	Holt and Kilger, 2012; Geers, 2011
Cyber-criminal networks	Huovinen et al., 2014
Non-state terrorist hackers	Holt and Kilger, 2012
Civil 'cyber warriors'	Holt and Kilger, 2012
Cyber spies (states/companies)	Holt and Kilger, 2012; Lehto et al., 2017
Automated computer programs	Geers, 2011, p. 40
Cyber criminals	Holt and Kilger, 2012; Huovinen et al., 2014
Hacker communities	Holt and Kilger, 2012; Geers, 2011
Criminal groups belonging to mafia	Uusi Suomi, 2017
Cyber terrorists	Lehto et al., 2017
Botnets (bot networks)	Jaitner, 2013; Geers, 2011
Hackers	Huovinen et al., 2014; Lehto et al., 2017; Geers, 2011
Semantic botnets (botnet army)	Geers, 2011
Disaffected individuals	Holt and Kilger, 2012
Hactivist movements	Hyppönen, 2013
Cyber commanders	Geers, 2011
Organized criminals (gangs)	Hyppönen, 2013
Criminal hackers and groups	Holt and Kilger, 2012; HS Metrolehti, 2017
Cyber tribes	Ottis, 2013
Cyber soldiers	Jantunen, 2010; Lehto et al., 2017; Huovinen et al., 2014
Total:	**22**

Table 2c. Information operation actor (IOA), cyber operation actor (COA) and cyber attack actor (CAA), and literature source.

Actor (IOA, COA and CAA)	Literature Source
Non-state actors and hackers	Holt and Kilger, 2012: Geers, 2011; Giles, 2017, p. 2
Non-legal organizations	Huovinen et al., 2014
Undemocratic regimes	NATO StratCom COE, 2016b
Foreign intelligence agents	Huhtinen and Rantapelkonen, 2016
Cyborg and bot program trolling	Jaitner, 2013; Nissen, 2012
Botnet controller	Geers, 2011
Governments	Hyppönen, 2013; Geers, 2011
Hacker associations	Holt and Kilger, 2012
Disguised agents of foreign countries	NATO StratCom COE, 2014
Online relationship hackers	Holt and Kilger, 2012
Patriot hackers	Lehto et al., 2017; Hollis, 2011; Mäntylä, 2014; Geers, 2011
Foreign civil intelligence	Huhtinen and Rantapelkonen, 2016
Information special forces	NATO StratCom COE, 2016b
Foreign intelligence domestic contacts	Rusi, 2017
Offline relationship hackers	Holt and Kilger, 2012
Implementers of blackmail programs	HS Metrolehti, 2017
Terrorist spies in Western societies	NATO StratCom COE, 2016a
Foreign intelligence's domestic contributors	Rusi, 2017

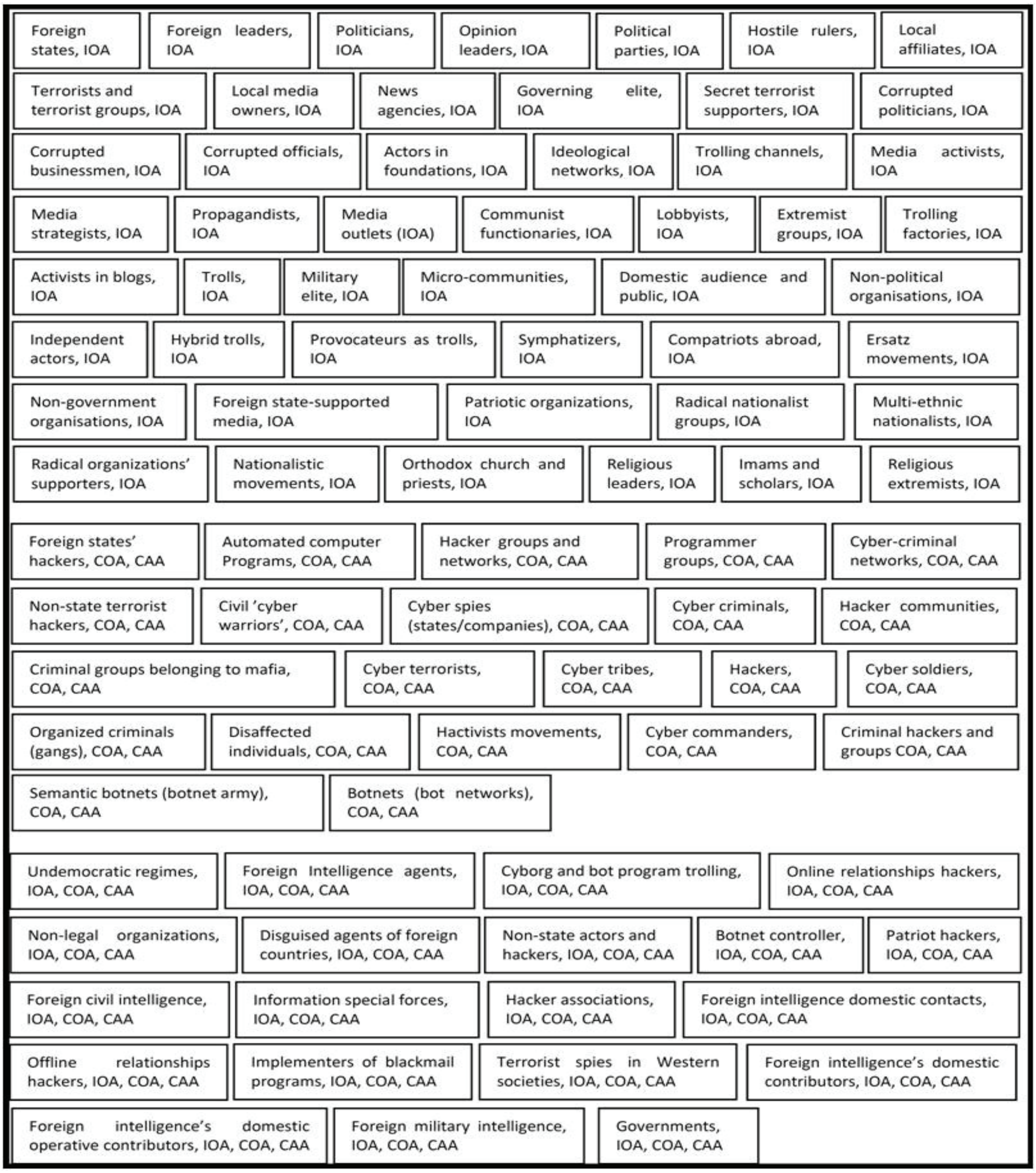
Actor (IOA, COA and CAA)	Literature Source
Foreign intelligence's domestic operative contributors	Rusi, 2017
Foreign military intelligence	Geers, 2011; Huhtinen and Rantapelkonen, 2016
Total:	**20**

The case study approach (Yin, 1993) was chosen for this study, and the discovered actors were repeatedly verified and validated by studying open primary literature sources in order to determine who carried out the operations and attacks in society's hybrid IE (Wolfswinkel, Furtmueller and Wilderom, 2013). In actor validation, content analysis (Krippendorff, 1985) and a qualitative research approach (Grounded Theory) (Glaser and Strauss, 1967) were applied to discover the actors and their literature sources of evidence. The IOA, COA or CAA were validated using the Grounded Theory approach (Glaser and Strauss, 1967; Eisenhardt, 1989; Wolfswinkel, Furtmueller and Wilderom, 2013). Eisenhardt (1989) claims that combining a case study with the Grounded Theory approach produces a new formal theory which can be tested later using empirical data. The actors carried out IO; or both cyber operations and cyber attacks; or IO, cyber operations and cyber attacks. After a saturation point of past studies was achieved, 49 observations IOA, 22 observations of COA and CAA, and 20 observations of IOA, COA and CAA were made. Thus the total sum of observations of IOA was 69; of actors carrying out cyber operations 42; and of actors carrying out cyber attacks 42.

Because the goal of the research was to create a new theory, a conceptual model (See Figure 1) of IOA, COA and CAA in society's hybrid IE was developed (Myers and Avison, 2002; Wolfswinkel, Furtmueller and Wilderom, 2013). The literature sources were examined until no new actors emerged from them and the saturation point was achieved.

4. Research findings

The found actors can be seen as small boxes in Figure 1, in no specific order. Even though Figure 1 explains the IOA, COA and CAA in society's hybrid IE, it lacks the relationships between the actors based on past studies. Furthermore, some actors may be hidden and secret. In this study, the IOA, COA and CAA are denoted as thematic categories (Glaser and Strauss, 1967).



Foreign states, IOA
Foreign leaders, IOA
Politicians, IOA
Opinion leaders, IOA
Political parties, IOA
Hostile rulers, IOA
Local affiliates, IOA
Terrorists and terrorist groups, IOA
Local media owners, IOA
News agencies, IOA
Governing elite, IOA
Secret terrorist supporters, IOA
Corrupted politicians, IOA
Corrupted businessmen, IOA
Corrupted officials, IOA
Actors in foundations, IOA
Ideological networks, IOA
Trolling channels, IOA
Media activists, IOA
Media strategists, IOA
Propagandists, IOA
Media outlets (IOA)
Communist functionaries, IOA
Lobbyists, IOA
Extremist groups, IOA
Trolling factories, IOA
Activists in blogs, IOA
Trolls, IOA
Military elite, IOA
Micro-communities, IOA
Domestic audience and public, IOA
Non-political organisations, IOA
Independent actors, IOA
Hybrid trolls, IOA
Provocateurs as trolls, IOA
Symphatizers, IOA
Compatriots abroad, IOA
Ersatz movements, IOA
Non-government organisations, IOA
Foreign state-supported media, IOA
Patriotic organizations, IOA
Radical nationalist groups, IOA
Multi-ethnic nationalists, IOA
Radical organizations' supporters, IOA
Nationalistic movements, IOA
Orthodox church and priests, IOA
Religious leaders, IOA
Imams and scholars, IOA
Religious extremists, IOA

Foreign states' hackers, COA, CAA
Automated computer Programs, COA, CAA
Hacker groups and networks, COA, CAA
Programmer groups, COA, CAA
Cyber-criminal networks, COA, CAA
Non-state terrorist hackers, COA, CAA
Civil 'cyber warriors', COA, CAA
Cyber spies (states/companies), COA, CAA
Cyber criminals, COA, CAA
Hacker communities, COA, CAA
Criminal groups belonging to mafia, COA, CAA
Cyber terrorists, COA, CAA
Cyber tribes, COA, CAA
Hackers, COA, CAA
Cyber soldiers, COA, CAA
Organized criminals (gangs), COA, CAA
Disaffected individuals, COA, CAA
Hactivists movements, COA, CAA
Cyber commanders, COA, CAA
Criminal hackers and groups COA, CAA
Semantic botnets (botnet army), COA, CAA
Botnets (bot networks), COA, CAA

Undemocratic regimes, IOA, COA, CAA
Foreign Intelligence agents, IOA, COA, CAA
Cyborg and bot program trolling, IOA, COA, CAA
Online relationships hackers, IOA, COA, CAA
Non-legal organizations, IOA, COA, CAA
Disguised agents of foreign countries, IOA, COA, CAA
Non-state actors and hackers, IOA, COA, CAA
Botnet controller, IOA, COA, CAA
Patriot hackers, IOA, COA, CAA
Foreign civil intelligence, IOA, COA, CAA
Information special forces, IOA, COA, CAA
Hacker associations, IOA, COA, CAA
Foreign intelligence domestic contacts, IOA, COA, CAA
Offline relationships hackers, IOA, COA, CAA
Implementers of blackmail programs, IOA, COA, CAA
Terrorist spies in Western societies, IOA, COA, CAA
Foreign intelligence's domestic contributors, IOA, COA, CAA
Foreign intelligence's domestic operative contributors, IOA, COA, CAA
Foreign military intelligence, IOA, COA, CAA
Governments, IOA, COA, CAA

Figure 1: Conceptual model of IOA, COA and CAA in society's hybrid IE.

5. Conclusions and discussion

A solid theory on IOA, COA and CAA in society's hybrid IE is important, because so many different actors carry out operations and attacks in hybrid IE, as can be seen in Figure 1. This study used past scientific research and popular news etc. to find the actors who carry out these operations and attacks, and developed a new theory – a conceptual model of the actors in society's hybrid IE (See Figure 1). Figure 1 shows a total of 91 different actors who carry out operations and attacks: 49 IOA; 22 COA and CAA; and 20 IO, COA and CAA were found in past studies. The 91 actors are denoted in thematic categories. This study is in line with the existing research and past news, as the discovered theory called a conceptual model in Figure 1 is based on it.

Seven conclusions emerge from this study. First, the majority of IOA are states, politicians, political or non-political organizations, nationalists, patriots, terrorists, media, religious actors, and different movements. Second, the majority of COA and CAA are hackers or hacker groups, criminal groups, actors connected to cyber operations or programs and program armies. Third, the majority of IOA, COA and CAA are intelligence actors, hackers and different organizations. Fourth, politicians and citizens can themselves act as trolls, spreading false information with or without being aware of it themselves. Fifth, society's cyber networks, infrastructures and information systems can be attacked and crippled by programs that are autonomous and think independently. Sixth, operations and attacks can be sold to an outsider, which makes finding the real actor behind them, called an attribution problem, very difficult. Seventh, the IOA, COA and CAA are dangerous due to their international and national connections, relationships and networks. A global network of actors carrying out operations and attacks is an even more severe threat to society's hybrid IE.

This study has several implications for the actors defending and securing society's hybrid IE. First, the 91 actors carrying out operations and attacks force defenders to be alert around the clock and to develop increasingly sophisticated tools, defence tactics, technologies and analysis procedures in order to respond. Increasing amount of expertise is needed to discover operations and attacks to information systems and networks and to recognize them in society's hybrid IE. Furthermore, defenders should think carefully about how to inform the citizens in order to keep them on their side. The research followed the Grounded Theory (Glaser and Strauss, 1967) approach. The theory, the conceptual framework of the IOA, COA and CAA, the scientific contribution of this study, emerged from the existing research in an inductive manner and in accordance with the theory of "*salient concepts arising from the literature"* (Wolfswinkel, Furtmueller and Wilderom, 2013, p. 2). The methodological contribution was the use of multiple different types of past studies together in the manner of qualitative research (Wolfswinkel, Furtmueller and Wilderom, 2013). As noted above, this study focused solely on IOA, COA and CAA in society's hybrid IE. This literature study shows that the hybrid IE is constantly undergoing undergoes dynamic changes. As a consequence, the defenders must make careful security choices about the future, that is, anticipate the possible ways in which an actor may damage society's IE.

This study has naturally some limitations. First, the attribution problem of identifying the real actors behind the operations and attacks causing harm to society, which can lead to a conflict that damages society's critical infrastructures or causes loss of lives if retaliatory actions against the actors carrying out the operations and attacks become cyber conflict situations or cyber warfare actions. Kinetic attacks and operations against the wrong actors may also begin if they are not really known (Holt and Kilger, 2012). The second limitation concerns the conceptual model in Figure 1, which lacks the relationships between the actors. The third limitation is that the evidence from past studies possibly may be insufficient to determine valid and reliable actors, or lack some important actors.

In the future, a case study could be carried out to determine the relationships between the actors, because information defence, cyber defence and cyber security need new ways to find the actors behind operations and attacks in society's hybrid IE, especially if these have already built up a global network.

References

Armistead, E. L., United States and Joint Forces Staff College. (2004) *Information Operations: Warfare and the Hard Reality of Soft Power*, Washington, D.C.: Potomac Books.

Conley, H.A., Mina, J., Stefanov, R. and Vladimirov, M. (2016) *The Kremlin Playbook. Understanding the Russian Influence in Central and Eastern Europe*. A Report of the Center for Strategic International Studies Europe Program and the CSD Economics Program, New York: Rowman & Littelefield.

Creswell, J.W. (2007) *Qualitative Inquiry and Research Design: Choosing Among Five Approaches,* California, US: Sage Publications.

Eisenhardt, K.M. (1989) Building Theories from Case Study Research*. Academy of Management Review*, 14(4), pp 532-550.

Geers, K. (2011) *Strategic Cyber Security.* NATO Cooperative Cyber Defence Centre of Excellence, Tallin: CCD COE Publication. Available at: https://www.scribd.com/document/62478319/Strategic-Cyber-Security-K-Geers Accessed 15 Jun. 2017.

Giles, K. (2017) *The Next Phase of Russian Information Warfare,* Riga: NATO StratCom COE.

Glaser, B. and Strauss, A.L. (1967). *The Discovery of the Grounded Theory: Strategies for Qualitative Research,* Chicago, IL: Aldine.

Goolsby, R. (2013) On Cybersecurity, Crowdsourcing, and Social Cyber-Attack, *Policy Memo Series*, 1, pp 1-8.

Hollis, D. (2011) Cyber War Case Study: Georgia 2008. *Small Wars Journal* [online]. Available at: http://smallwarsjournal.com/jrnl/art/cyberwarcase-study-georgia-2008 Accessed 30 Jan. 2017.

Holt, T.J. and Kilger, M. (2012) Know Your Enemy: The Social Dynamics of Hacking. Honeynet Project, *KYE-Paper* (Draft).

HS Metrolehti. (2017) 'Origin of cyber attack revealed'. Available at: https://www.pressreader.com/finland/hs-metro/20170629/281629600284751 Accessed 29 Jun. 2017.

Huhtinen, A-M. and Rantapelkonen, J. (2016) Junk Information in Hybrid Warfare: The Rhizomatic Speed of Social Media in the Spamosphere*.* In: *European Conference on Cyber Warfare and Security*. Munich: Academic Conference International Limited, pp 136-144.

Huovinen, P., Kärkkäinen, A., Lehto, M., Noronen, L., Pispa, K. and Viita, V. (2014) 'Cyber performances in 2030', Helsinki: National Defence University.

Hybrid (2017) In: *Dictionary.com* [online] California. Available at: http://dictionary.com/browse/hybrid Accessed 31 Mar. 2017.

Hyppönen, M. (2013) The Exploit Marketplace. In: *The Fog of Cyber Defence*, J. Rantapelkonen and M. Salminen, ed., Tampere: Juvenes Print Oy, pp 231-234.

Information (2017) In: *Merriam-Webster* [online] Springfield: Merriam-Webster, Inc. Available at: https://www.merriam-webster.com/dictionary/information Accessed 22 Aug. 2017.

Information Operations. (2012) *Joint Publication 3-13.* Available at: http://www.dtic,mil/doctrine/new_pubs/jp3_13.pdf Accessed 24 Jan. 2017.

Jackson, L. (2015) Revisions of Reality: The Three Warfares - China's New Way of War*.* In: *Beyond Propaganda. Information at War: From China's Three Warfares to NATO's Narratives,* London**:** Legatum Institute: Transitions Forum, pp 5-15.

Jaitner, M. (2013) Exercising Power in Social Media. In: *The Fog of Cyber Defence*, J. Rantapelkonen and M. Salminen, ed., Tampere: Juvenes Print Oy, pp 57-77.

Jantunen, S. (2013) *Strategic Communication: practice, ideology and dissonance,* Doctoral Dissertation, Tampere: Juvenes Print.

Jantunen, S. (2010) *Creating Modern Threat Scenarios: A Case Study on the Narrative of Chinese Cyber-Warriors*, Helsinki: Edita Prima Oy.

Krippendorff, K. (1985). *Content analysis. An Introduction to its Methodology.* California, CA**:** Sage Publications.

Laine, V. (2015) Managed Nationalism. Contemporary Russian Nationalistic Movements and Their Relationship to the Government, *FIIA Working Paper*.

Lange-Ionatamishvili, E. and Svetoka, S. (2015) Strategic Communications and Social Media in the Russia Ukraine Conflict. Chapter 12. In: *Cyber War in Perspective: Russian Aggression against Ukraine*, Kenneth Geers, ed., Tallinn: NATO CCD COE Publications.

Lehto, M., Limnéll, J., Innola, E., Pöyhönen, J., Rusi, T. and Salminen, M. (2017) *Finland's cyber security: the present state, vision and the actions needed to achieve the vision.* Publications of the Government´s analysis, assessment and research activities 30, Helsinki: Prime Minister´s Office.

Myers, M.D. and Avison, D.E. (Eds.) (2002) *Qualitative Research in Information Systems*: Review, London: Sage Publications.

Mäntylä, J. (2014) 'Cyber weapons' impact on information systems of critical infrastructure'. *Thesis*. Helsinki: National Defence University.

NATO. (2012) *NATO military policy on information operations*. Available at: https://info.publicintelligence.net/NATO-IO-Policy.pdf Accessed 24 Jan. 2017.

NATO StratCom COE. (2014) Analysis of Russia's Information Campaign Against Ukraine. Examining non-military aspects of the crisis in Ukraine from a strategic communications perspectives. *Executive Summary,* Riga: NATO StratCom COE.

NATO StratCom COE. (2016a) Daesh Recruitment: How the Group Attracts Supporters, Riga: NATO StratCom COE.

NATO StratCom COE. (2016b) *Social Media as a Tool of Hybrid Warfare*, Riga: NATO StratCom COE. Available at: http://www.stratcomcoe.org/social-media-tool-hybrid-warfare Accessed 30 May. 2017.

Nissen, T.E. (2012) *Social Media's Role in 'Hybrid Strategies',* Riga: NATO StratCom COE. Available at: http://www.stratcomcoe.org/social-medias-role-hybrid-strategies-author-thomas-elkjer-nissen Accessed 12 Jan. 2017.

Ottis, R. (2013) Theoretical Offensive Cyber Militia Models. In: *The Fog of Cyber Defence,* J. Rantapelkonen and M. Salminen, ed., Tampere: Juvenes Print Oy, pp 190-199.

Pettigrew, A.M. (1985) Contextualist research: A Natural way to link theory and practice. In: E. Lawler, A. Mohrman, S. Mohrman, G. Ledford and T. Cummings, ed., *Doing research that is useful for theory and practice*, San Franscisco: Jossey-Bass, pp 222-274.

Pomerantsev, P. (2015) Introduction. In: *Beyond Propaganda. Information at War: From China's Three Warfares to NATO's Narratives*. Legatum Istitute: Transitions Forum. Available from: www.li.com www.prosperity.com Accessed 30 May. 2017.

Rusi, A. (2017) 'Alpo Rusi asks: Did Russia's intelligence torpedo Finnish NATO discussions at the beginning of 2000?'. *Twitter/Kotimaa*, 16/17 July 2017.

Sartonen, M., Huhtinen, A-M. and Lehto, M. (2015) From influencee to influencer: the rhizomatic target audience of the cyber domain. In: *ECCWS 2015 Proceedings of the 14th European Conference on Cyber Warfare & Security,* Hatfield, UK: Academic Conferences and Publishing International Ltd., pp 249-256.

Sillanpää, A. (2016) The Moldovan Information Environment: Hostile Narratives, and their Ramifications, *Executive Summary*, Riga: NATO StratCom COE.

Sirén, T., Huhtinen, A-M. and Toivettula, M. (2011) 'From information warfare to comprehensive strategic communication'. In: T. Sirén, ed., 'Strategic Communications and Information Operations in 2030', Helsinki: Juvenes Print Oy, pp 3–21.

Social media (2017) In: *Merriam-Webster* [online] Springfield: Merriam-Webster, Inc. Available at: https://www.merriam-webster.com/dictionary/social%20media Accessed 1 Sep. 2017.

Society (2018) In: Wikipedia [online]. Available at: https://en.wikipedia.org/wiki/Society Accessed 18 May. 2018.

Uusi Suomi (2017) 'Russian mafia operating in Europe by order of Russia'. Available at: https://www.uusisuomi.fi/uutiset/ulkomaat Accessed 1 Feb. 2017.

Wihersaari, K. (2015) *Intelligence Acquisition Methods in Cyber Domain: Examining the Circumstantial Applicability of Cyber Intelligence Acquisition Methods Using a Hierarchical Model*, Thesis, Helsinki: National Defence University.

Winter, C. (2017) Media Jihad. The Islamic State's Doctrine for Information Warfare. Institute of Strategic Dialogue, The International Centre for the Study of Radicalisation and Political Violence.

Wolfswinkel, J.F., Furtmueller, E. and Wilderom, C.P.M. (2013) Using Grounded Theory as a Method for Rigorously Reviewing Literature, *European Journal of Information Systems*, 22(1), pp 45-55.

Yin, R.Y. (1993) Applications of Case Study Research, *Applied Social Research Methods Series,* 34, Sage Publications.

Young People and the Dark Side of Social Media: Possible Threats to National Security

Teija Norri-Sederholm[1], Reetta Riikonen[1], Panu Moilanen[2] and Aki-Mauri Huhtinen[1]
[1]Department of Leadership and Military Pedagogy, National Defence University, Helsinki, Finland
[2]University of Jyväskylä, Finland
teija.norri-sederholm@mil.fi
reettariikonen1@gmail.com
panu.moilanen@jyu.fi
aki.huhtinen@mil.fi
DOI: 10.34190/EWS.20.064

Abstract: Social media is increasingly becoming a forum for criminality, misuse, and hate speech, as there are no filters or other controlling mechanisms to filter user-generated content.Furthermore, disinformation and propaganda are becoming more sophisticated and harder to track. Hence, this dark side of social media can pose a viable threat to national security. Future generations will be born into an environment of polluted and polarised online information networks. Consequently, young people, many of whom use social media on a daily basis, will have to find ways to survive in these circumstances, often without the help, knowledge, or experience of earlier generations. Thus, young people are at risk of becoming predisposed to all kinds of harmful material, which, in turn, can affect their thinking and behaviour. This can lead to many new threats to national security. This study focuses on the observations of police officers on the current trends and threats youngsters face on the dark side of social media. The aim was to examine possible threats to national security related to young people's social media use based on data from three semi-structured interviews with police officers working in Preventive Measures Units. The analysis was done using inductive content analysis. Based on the analysis, there are three main threats to national security concerning young people's social media use: the amount of false information available online, the glorification of violence and crime, and large-scale gatherings and swarming. The results indicate that although most young people's social media use is harmless, social media platforms can also be used in a way that threatens national security. Many of the threats posed by young people's social media use have not yet been realised. However, it is important to be aware of the risks and be prepared for any possible negative outcomes in order to maintain national security.

Keywords: information influence, national security, police, social media, young people

1. Introduction

Social media is a core part of young people's everyday life. For example, in Finland three quarters of all 16–24 year–olds use social media on a daily basis (Statistics Finland, 2020). The core elements of social media consist of interactive services, user-generated content, user-specific profiles, and online social networks with other individuals and groups (Obar & Wildman, 2015). The dark side of social media can be defined as a "collection of 'negative' phenomena that are associated with the use of IT that has the potential to infringe on the well-being of individuals, organisations and societies" (Tarafdar et al., 2015, 161). Social media has multiplied the risks of misinformation, disinformation, propaganda, and hoaxes (Posetti & Matthews, 2018), and it is increasingly a hub for criminality, misuse like trolling and bullying, hate speech, and harassment (Jutterström 2019, 207; López-Fuentes 2018). Thus, social media does pose threats to national security.

One of the threats to national security by social media is information influencing. Social media can be used to alter and manipulate the social, cognitive, and psychological dynamics of a given social network or community. This is often referred to as influence or psychological operations. (Chandramouli, 2011.) These operations are often based on spreading disinformation. Disinformation can be defined as false information disseminated with the intent to deceive people, whereas misinformation does not include any malicious intent (Fetzer, 2004, 228; Lazer et. al. 2018, 1094). The ease with which disinformation can be spread is one of the properties linked to the essence of social media: it is based mainly on user-generated content and has none of the editorial controls of traditional media (Allcott and Gentzkow 2017, 211; Jensen et. al. 2010; Livingstone 2019, 6). In addition, it is easy to gain information of users and influence them in social media by using programmed algorithms, bots, which can also create filter bubbles because they present users with information that is similar to their previous behaviour in social media (Burkhardt 2017, 12; Spohr 2017, 152–153). This can lead to ideological polarization among users as the content they encounter aligns with their pre-existing beliefs (Spohr 2017, 153). The success of disinformation is based on an understanding of human psychology: it appeals to people's minds and their beliefs. Therefore, it has been successfully used for creating rifts between citizens and polarising society. (McGeehan, 2018.) Because of the nature of rhizomatic information networks, we cannot eliminate disinformation on social media. Hence, there will always be new applications to attract audiences into distributing disinformation. In addition, applications are developed to be more sophisticated and challenging to control by institutional regulations. (Napoli, 2019.) Overall, information influencing activities can be described as having four characteristics: they are deceptive, have bad intentions, "seek to disrupt constructive debate", and "interfere in debates or issues in which foreign actors play no legitimate role" (Pamment & Agardh-Twetman, 2018, 6-7).

Social media can also be used to create incidents and events. For example, swarming, unexpected gatherings of large numbers of people in particular places, is a phenomenon strongly linked to communication on social media (White 2006). Probably the best-known form of swarming is a flash-mob, in which a group of individuals normally unknown to each other meet in a certain place to do something silly and then disperse within ten minutes (Solecki & Goldschmidt 2011; Seo et. al. 2014). However, there also exists more serious forms of swarming, such as gatecrashings, riots and mobs (White 2006; Wasik 2011). Whereas traditional flash-mobs are positive and funny, these other forms of swarming might pose serious security threats. In addition, social media can be used to catalyse incidents, which have their origins in the virtual world, but which then spill over into the real world, possibly having serious consequences (Chandramouli, 2011; Irwin-Rogers & Pinkney 2017). Typically, the ultimate goal of influence or psychological operations is to cause effects in the real world. In addition, it is not uncommon that minor real-world incidents become catalysed in social media and result in more drastic real-world events.

Another threat to national security is the availability of violent content in social media. The link between social media content and real-world violence was originally identified as being related to gang violence in particular (Irwin-Rogers, Densley & Pinkney 2018), but it is actually a more common phenomenon. Among young people, spreading and consuming violent content is not rare: this can be in the form of extremely graphic videos and pictures for the purpose of thrill-seeking and entertainment, for example. In some cases, this kind of behaviour can be linked to a phenomenon called appetitive violence, in which violence and other forms of delinquency are admired and, in some cases, actually carried out. (Ching, Daffern & Thomas, 2012.)

In this paper, we focus on police officers' observations regarding current trends and threats associated with young people and the dark side of social media. Our aim is to examine possible threats to national security related to young people's social media use. In addition to the more traditional threats, it is important to understand these new threats related to social mediain order to enhance national security comprehensively.The results will serve as a pilot for our future studies concerning young people, social media, and national security

2. Material and Methods

This paper is based on a qualitative study design. The data was collected using semi-structured interviews. The interview themes relating to social media were present and future threats, new technology, the relation between the real world and the virtual world, and the communication environment. The empirical data included three interviews from three different police districts representing different geographical areas in Finland and different sized organisations in order to obtain sample diversity. The voluntary interviewees were chosen based on their role and knowledge in the Preventive Measures Unit. The purpose of preventive measures undertaken by the police is to improve public security and combat crime. Two interviews were one-to-one interviews and the other was a group interview with three police officers. Before the interview, all participants were informed about the study and signed an Informed Consent Form that included the description of the study, its intended use, the confidentiality of the study, and the rights of the participant. The interviews were conducted by the first, second, and third author, and were audio-recorded. The mean duration of each interview was approximately 74 minutes.

Interview data was transcribed verbatim and analysed (Figure 1) using a qualitative inductive content analysis (Krippendorff, 2013). In the first stage of the analysis, data was coded using Atlas.ti 8.4.22 qualitative data analysis software. Eight code groups were created: disinformation and information influencing, swarming, polarisation and strong language, transitions between social media and the real world, the generation gap, social media as an information source, glorification of violence and crime, and a channel to the dark side. In the second phase of the analysis, data not related to national security was excluded and a mind map used to regroup the remaining data. Following this, three groups were formed: false information in social media, the glorification of violence and crime, and large-scale gatherings and swarming.

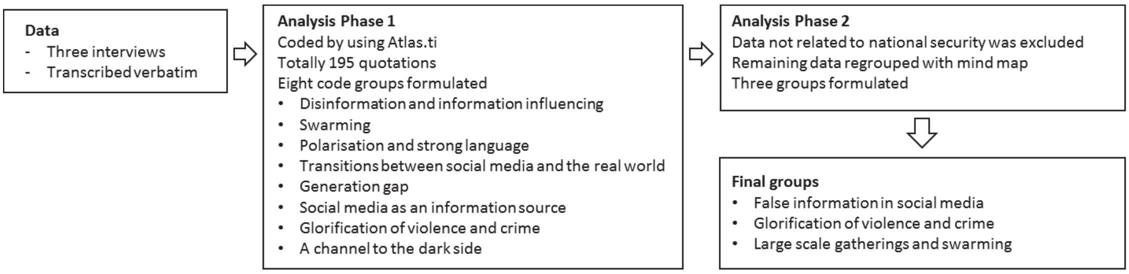

Data
- Three interviews
- Transcribed verbatim
Analysis Phase 1
Coded by using Atlas.ti
Totally 195 quotations
Eight code groups formulated
• Disinformation and information influencing
• Swarming
• Polarisation and strong language
• Transitions between social media and the real world
• Generation gap
• Social media as an information source
• Glorification of violence and crime
• A channel to the dark side
Analysis Phase 2
Data not related to national security was excluded
Remaining data regrouped with mind map
Three groups formulated
Final groups
• False information in social media
• Glorification of violence and crime
• Large scale gatherings and swarming

Figure 1: Analysis process

3. Findings

According to the police officers, the main threats to national security in the context of young people and the dark side of social media are the amount of false information in social media, the glorification of violence and crime in social media, and large-scale gatherings and swarming. Young people, sometimes just out of curiosity, can be exposed to things that have a bad influence on them. Social media offers everything for everyone and even so-called decent youngsters can stray onto the dark side of social media. The ease of buying drugs, surfing on the Tor-net, and joining groups with polarised thinking have been seen in police officers' work. The problem with the internet and social media is that a much larger group of youngsters is exposed to bad influences than before. For example, it is obvious that social media has had a great influence in the process of radicalisation and the number of radicalised people.

Spreading mis- and disinformation is very easy on social media as the information flow is enormous. This poses daily challenges to young people. According to the interviewed police officers, young people do not always have the ability to critically evaluate the information, the source, and the trustworthiness of the information they come across in social media. The interviewees noted that if the information comes from a reliable source from the youngsters' point of view, such as a friend or relative, they seem to assume that the information is true, even though it may not necessarily be so. This "blind" acceptance might be due to the mental immaturity of youngsters. Another challenge in the context of disinformation is the simplification of messages and the lack of reasoning behind statements and arguments in social media. This makes information influencing easier, as you only need the right timing for your provocative intervention.

When there is an incident, such as a school threat, it is typical that rumours, misinformation, and disinformation fill social media. As a result, the police must intervene and try to correct the information. However, the vast amount of rumours and disinformation makes it impossible for police to respond to them all. When such an incident happens, it is important that the police delivers the true information as quickly as possible in social media before someone fills the communication vacuum with rumours and misinformation. In recent years, police officers have detected a new phenomenon related to disinformation. As an attack happens, fabricated stories begin to spread on social media sometimes even within ten minutes of the first official news of the attack. The speed and the quality of the false material is such that, in addition to curious citizens, there has to be an organised group behind the disinformation. The look and feel of the fake news can be so similar to the official news channels that it is very hard to notice the difference and realise that these news reports are not true.

From the point of view of national security, the interviewees emphasised that what is more concerning is long-term influences and opinion formation. Extremist and other players know how to use incidents to their advantage and affect public opinion by creating negative impressions related to society and public authorities. In general, their messaging includes systematic communication and material. Target groups, sometimes including youngsters, are planned in advance, focusing on having more supporters and visibility in the media. It is notable that some young people seek this kind of action and they look for groups where people share, sometimes even quite polarised, ideas.

All the police officers in the interviews noted that lack of consideration and attention seeking are noticeable phenomena within youngsters on social media. Youngsters may create rather provocative pictures, texts, videos, or memes, for example, to gain more followers or to emphasise their belonging to a group. Lack of consideration can also be seen in the messaging between youngsters, as the messages sometimes even include threats of violence. Occasionally it is just a rhetorical way of messaging. However, sometimes the threat transforms into an act of violence in real life. These kinds of incidents happen as social media enables youngsters to indulge in faceless forms of behaviour they would never engage in when face-to-face with others. Nevertheless, it is possible that the changes in the youngsters' messaging and behaviour are due to some sort of turning point in society. The increase of polarisation and threatening messaging on social media are phenomena that concern youngsters and adults alike. This is also reflected in the level of security in society and in the citizens' feeling of safety.

The interviewees also highlighted the role of the internet in radicalisation processes. Young people spend a lot of time there looking for like-minded friends and people. In case young people are looking for these kinds of polarised groups to strengthen their own identity, it may cause cognitive distortion and increase a black and white worldview. Moreover, idolising violence and criminality has been identified as a phenomenon insome youngsters. Young people can watch exceptionally violent videos, including brute force being used on real people and this has become everyday entertainment. An example of this is the mosque shooting video in New Zealand, which the following day, children were watching on their phones as it was freely available on social media.

According to the police officers in the interviews, there are opinion leaders and influential individuals among young people in social media. Having thousands of followers, some of these youngsters post pictures and videos where, for example, they are carrying weapons or doing something criminal, such as stealing. The problem is that the followers of such influential individuals, ordinary young people, admire them and might want to copy this behaviour. Thus, this phenomenon of idolising crime can create a feeling of insecurity among young people and affect their mind set. Social media has also made swarming and organising gatherings easy and opinion leaders in particular can easily organise gatherings or mass brawls with many of their followers as bystanders out of curiosity. There can be tens or even hundreds of young people participating in such gatherings. Often these gatherings are harmless as the aim is simply to attract many people in the same place at the same time, but on other occasions the aim is to organise a mass brawl, for example. This pattern of behaviour and the way of influencing includes the traditional elements of being young wanting to be part of a group and to be accepted and recognised. One positive aspect is that for most youngsters there are no sinister intentions behind this kind of behaviour, but it is may be down to thoughtlessness or their will to gain visibility. On the other hand, this may cause young people to look for acceptance in such a way that they may actually become victims.

4. Discussion

The aim of this study was to examine possible threats to national security related to young people's social media use. This was done through a series of interviews with police officers from the Preventive Measures Unit. The results indicated that the main threats to national security concerning young people and the dark side of social media were the amount of false information, the glorification of violence and crime, and large-scale gatherings and swarming.

False information, whether it is mis- or disinformation, can have many negative outcomes for national security. In the case of sudden attacks and incidents, rumours as such can affect people's feeling of safety. In addition, it is possible that false information even leads to violent action in real life. For example, a manipulated picture or video shared at the right time can lead to a clash between groups with pre-existing contrary worldviews. This is one way of doing information influence. (Pamment & Agardh-Twetman, 2018.) However, the biggest risk to national security is the long-term information influencing on social media, which can diminish youngsters' trust in public authorities and the government. Local communities are also affected by false information as it can cause tension and communal dissolution. Information influence has many serious consequences in society, such as radicalisation and polarisation (see McGeehan, 2018).

Social media plays an essential role in young people's everyday life and mostly its usage is harmless. However, the same features, increasing communication and joy, are used for harmful purposes (Jutterström, 2019; López-Fuentes, 2018) posing threats to society. With the vast amount of false information in social media (Posetti & Matthews, 2018), the problem is that it is impossible for the police to detect and correct it all. The task is becoming even more difficult as new technologies and applications are constantly being developed (Napoli, 2019) and typically youngsters are the first ones to use new applications. Hence, it is hard for the police to keep up with the constant changes in social media platforms. This requires active participation and constant training from police in the use of social media. Furthermore, the role of specialised police teams and the use of artificial intelligence in data mining will increase in the near future. Therefore, it is crucial that the police adapt to this new environment. This may require changes in the selection and training of new police officers because the traditional qualities of a good police officer may not be enough in combatting crime in the virtual world.

Harmful content, available daily on social media, poses another threat to national security, as more and more youngsters become predisposed to detrimental material, such as strong language, violent content, drugs, and radicalised groups. In the context of national security, the glorification of violence and crime is an especially worrying phenomenon, as the admiration and consumption of violent content can turn into concrete action (Ching, Daffern & Thomas, 2012). The problem is that it is impossible to erase all the harmful content in social media. Hence, there will always be bad influences in social media regardless of the action of authorities. Another problem is that as more people than before become exposed to this kind of material there is the possibility that the content gradually becomes more common, accepted, and normal. Thus, the role and responsibility of parents in monitoring their children's social media use and educating their children in social media are more important than ever. Youngsters should learn at home where the line between normal and abnormal is. However, it is a fact that many parents are ignorant of the material their children are watching and sharing on social media.

The police officers' examples of swarming and other organised gatherings by young people showed how easy it is to organise large-scale gatherings on social media (see White 2006). This leads to at least two risks. First, it would be possible to organise massive gatherings or mass brawls of even thousands of people on social media. These events could lead to actual danger towards bystanders and threaten the sense of security of citizens. Secondly, although the gatherings organised by youngsters in Finland have not had outside actors or influencers so far, it is possible that some malicious actors would try to exploit the gatherings for their own purposes in the future.

The immensity of social media is, in itself, a threat to national security, as the total amount of information it contains is so enormous that it is impossible for the police to follow it all. For this reason, international police cooperation and citizens' awareness and willingness to share their knowledge are crucial elements in enabling successful preventive work and operations. This will be even more important in the future. In order to obtain information from citizens, the police must gain their trust and maintain a good relationship with them. In the case of young people, the police must adopt new ways of communication to reach them. For example, there should be more police officers working and chatting to youngsters on social media as this is a natural way of communication for youngsters. To reach youngsters, the police must also have a presence on the newest social media platforms the youngsters use. This requires a familiarity with the latest technological developments and applications and a willingness to learn how to use them. In addition, the police need an efficient way to organise all the information they acquire because not all citizens can accurately evaluate the importance and meaning of the information they share with the police.

Issues concerning the limitations and the validity of the study must also be considered. In particular, the study is small-scale with only three interviews. For this reason, the findings cannot be generalised. Having few more interviews would have ensured a better saturation in the data and thus increased the validity of the results. However, during the coding, data from each interview were part of each code group. This means that the themes of all three final groups came up in every interview. To increase the validity, the analysis was conducted in two phases by two of the authors. It should also be noted that the interviews were made in different parts of Finland to ensure a wider perspective. In addition, each interview represented a slightly different aspect of the Police Preventive Measure Units and all the interviewees were experienced police officers. Notwithstanding these limitations, the interviews provided data that describe current trends and threats of young people's social media use well. The study also served its purpose of forming a background for our future studies.

To conclude, the results of this study indicated that there are risks to national security that concern young people in social media. Nevertheless, further studies are required to discover what kind of experiences young people themselves have of the dark side of social media. The diversity of the social media platforms young people use is one of many challenges the public safety authorities face. To be able to secure national security in a constantly changing and transiting virtual environment, new kinds of skills and abilities are needed. Furthermore, there must be a willingness and a capability to adapt to these changes in order to be prepared for new threats arising from the dark side of social media.

Acknowledgements

This study is part of a research project funded by the Academy of Finland.

References

Allcott, H. and Gentzkow, M. (2017) "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 211–236.

Burkhardt, J. M. (2017) "Combating Fake News in the Digital Age", *Library Technology Reports*, Vol. 53, No. 8, pp.1–33.

Chandramouli, R. (2011) "Emerging social media threats: Technology and policy perspectives", *Proceeding of the 2011 Second Worldwide Cybersecurity Summit (WCS 2011)*, pp. 70–73.

Ching, H., Daffern, M., and Thomas, S. (2012) "Appetitive Violence: A New Phenomenon?", *Psychiatry, Psychology and Law*, Vol. 19, No. 5, pp. 745–763.

Fetzer, J. H. (2004) "Information: Does it Have To Be True?", *Minds and Machines*, Vol. 14, No. 2, pp. 223–229.

Irwin-Rogers, K., and Pinkney, C. (2017) *Social Media as a Catalyst and Trigger for Youth Violence*. Catch 22 / University College Birmingham, London.

Jensen, M. L., Burgoon, J. K., and Nunamaker Jr, J. F. (2010) "Judging the credibility of information gathered from face-to-face interactions" *Journal of Data and Information Quality*, Vol. 2, No. 1, pp. 3:1-3:20.

Jutterström, M. (2019) "Problematic Outcomes of Organization Hybridity: The Case of Samhall", in S. Alexius and S. Furusten (eds.) *Managing Hybrid Organizations Governance, Professionalism and Regulation,* Palgrave MacMillian Cham, pp. 199–214.

Krippendorf, K. (2013) "Content Analysis: An Introduction to Its Methodology", 3rd edition, SAGE, Los Angeles

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018) "The science of fake news", *Science*, Vol. *359,* No. 6380, pp. 1094–1096.

Napoli, P. M. (2019) *Social Media and the Public Interest. Media Regulation in the Disinformation Age*. Columbia University Press, New York.

Livingstone, S. (2019) "EU Kids Online", The International Encyclopedia of Media Literacy, Wiley Online Library.

López-Fuentes, F. de A. (2018) "Decentralized Online Social Network Architectures", in T. Özyer, S. Bakshi and R. Alhajj (eds.) *Social Networks and Surveillance for Society,* Springer*,* Lecture Notes in Social Networks. pp. 85–100.

McGeehan, T.P. (2018) "Countering Russian Disinformation", *Parameters*, Vol. 48, No. 1, pp- 49–57.

Obar, J. A., and Wildman, S. (2015) "Social media definition and the governance challenge: An introduction to the special issue" *Telecommunications Policy*, Vol. 39, No. 9, pp. 745–750.

Pamment, J. and Agardh-Twetman, H. (2018) *The role of communicators in countering the malicious use of social media,* NATO Strategic Communications Centre of Excellence.

Posetti, J. and Matthews, A. (2018) "A short guide to the history of ´fake news´ and disinformation", A learning module for journalists and journalism educators. ICFJ International Center for Journalist.

Seo, H., Houston, J. B., Knight, L. A. T., Kennedy, E. J., and Inglish, A. B. (2014) "Teens' social media use and collective action", *New Media & Society*, Vol. *16,* No. 6, pp. 883–902.Solecki, S. and Goldschmidt, K. (2011) "Adolescents texting and twittering: the flash mob phenomena", *Journal of Pediatric Nursing* Vol. 26, No. 2, pp. 167–169.

Spohr, D. (2017) "Fake news and ideological polarization: Filter bubbles and selective exposure on social media", *Business Information Review*, Vol. 34, No. 3, 150–160.

Statistics Finland (2020) *The use of information and communication technology*, [online], Statistics Finland, http://www.stat.fi/til/sutivi/2019/sutivi_2019_2019-11-07_tau_021_fi.html.

Tarafdar, M., Gupta, A., and Turel, O. (2015) "Editorial: Dark side of information technology use", *Information Systems Journal*, Vol. 25, No. 3, pp. 161–170.

Wasik, B (2011) "#Riot: self-organized, hyper-networked revolts – coming to a city near you", [online], *Wired Magazine,* https://www.wired.com/2011/12/ff_riots/.

White, R. (2006) "Swarming and the social dynamics of group violence", *Trends & Issues in Crime and Criminal Justice*, 326.

Fault Tolerance: An Approach Toward a Hybrid Fault Tolerance Framework Within a Distributed Computer Architecture

Christopher O'Flaherty, Louis Hyman, Malcolm Hodgson and Dewald Blaauw
Stellenbosch University, Stellenbosch, South Africa
Coflaherty16@gmail.com
Hymanlouis1@gmail.com
Malcolmhodgson7@gmail.com
Dnblaauw@sun.ac.za

DOI: 10.34190/EWS.20.016

Abstract: The increasing reliance on technical systems to achieve a desired level of organisational performance has significantly decreased the margin for system-related errors. The greater the number of interactions encountered by technical systems to provide the required functionality, the greater the fault exposure. These faults refer to the falling of a system into an undesirable state, hereafter referred to as failure. Faults may occur by many means, including malicious cyberattacks, bugs and malfunctions at component level. System failure often leads to substantial pecuniary losses and poses a high threat to system security owing to vulnerability (Castro and Liskov, 2002). To mitigate these issues, a hybrid fault framework is proposed. Through the implementation of a hybrid approach, where redundancy and consensus-based evaluation are implemented together as fault tolerance measures, system resilience will be greater when posed with failures. A network was created with three computers and security measures emplaced including Snort rules and an uncomplicated firewall. Thereafter, Pings, Port Scans and Denial of Service (DOS) attacks were performed in an attempt to induce a state of system failure with subsequent responses documented. One of the computers acted as a server, one as a user of the server, and one as an attacker. The Mean Time to Failure, Mean Time to Repair and Availability metric were used to monitor the effect on the network. It became evident that the computers faced strain and that if fault tolerance measures were emplaced, the computers would have responded to the attacks more efficiently. The metrics calculated indicated the need for fault tolerance in a system owing to the large discrepancies as the attacks occurred. To aid an infallible system, the implementation of a hybrid fault tolerance framework is necessary. Through the inclusion of Redundancy and consensus-based fault tolerance, which includes blockchain usage, the framework will result in resilient systems allowing various corporate ecosystems a safeguard against a multitude of system related issues.

Keywords: Cyberattacks, Fault tolerance, Redundancy, Resilience, Security, Byzantine

1. Introduction

Reliability is an imperative quality across all measures of system usage. However, it is difficult to ensure both system consistency as well as robustness. Computer systems are not always fault proof and are susceptible to malicious attacks as well as various other faults. Fault tolerance refers to the ability of a system to function as required when experiencing failure caused from a multitude of issues including, malicious attacks as well as software and hardware issues (Heimerdinger and Weinstock, 2019). Fault tolerance mechanisms, such as redundancy and consensus-based evaluations, are implemented to aid a system in situations wherein faults occur. Abundant resources are provided through fault tolerance for continuity as well as methods for identifying errors before critical failure is reached (Lee and Anderson, 1990). Ensuring fault tolerance within a system results in resilience and the ability to withstand various anomalies. Thereafter, if an issue does arise, a failsafe measure ensures a dependable system.

A controlled system with various nodes and well-defined security measures has been emplaced to ensure a scenario that simulates realistic events. By using simulated attacks, fault tolerance will be measured and tabulated through evaluation techniques (Kannan and Parker, 2006). Kali Linux and Ubuntu are the installed operating systems featuring a learned approach configuring necessary defence mechanisms with firewalls and other tools, including *Snort* and *Wireshark*. Simulated attacks will intrude the system, with the response being monitored and analysed to determine the value of a hybrid fault tolerance framework.

2. Research Question

This paper aims to determine and facilitate the need for and understanding of a Fault Tolerance Framework. The degree of assistance in system competence, as well as recovery times in the event of a cyberattack, will be explored.

3. Problem Statement

Systems are constantly developing and faults may occur in a system lifecycle. These faults should be accounted for during system inception and be detected, or deferred, through the use of implementing fault tolerance and redundancy. Implementing this may not be time or cost effective and thus this paper will examine the necessity of a hybrid fault tolerance framework and how it will optimize system resilience by means of redundancy and consensus-based fault tolerance.

4. Literature Review

4.1 Threats to information security

Information security threats are in abundance (Taylor, 2019). Having a secure system is imperative given the significance of information confidentiality and reliance on the system, as well as the information therein. Numerous malicious cyberattacks are performed daily, resulting in the common release of integral information and intellectual property. The loss of service through a malicious DOS attack may result in loss of customer trust (Hill, 2019). It is crucial to review the tolerance and performance of a system and, should a system is attacked, timeously ascertain what can be done to reduce the implications of this attack, as well as what can make the system less susceptible and therefore, more tolerant to such attacks.

The need for fault tolerance is emphasized by Geist and Trivedi (1990), owing to the fact that the durability of life critical systems, such as aircraft and spacecraft flight control, are affected. A fault in a system may result in an entire system malfunctioning.

To aid the fault tolerance framework, the Byzantine Model and Redundancy shall be examined.

4.2 The Byzantine Model

One of the most common forms of fault tolerance is the Byzantine Model. The Byzantine Model, which is a consensus-based method of fault tolerance, requires all nodes to be functioning and working together for a successful operation. A system may require a multitude of nodes to work systematically for system efficiency. Byzantine fault tolerance ensures system continuity, should one of the nodes malfunction. However, the system may not operate as efficiently following a malfunction (Castro and Liskov, 2002).

4.3 Redundancy

The implementation of redundancy permits another opportunity for a system to react to failures. Load balancing, a form of Redundancy granting more computational resources once demanded, ensures that traffic is effectively managed by a system and is therefore an infallible measure to mitigate potential cyberattacks (Hill, 2019).

4.4 Quantification of Fault Tolerance

When quantifying fault tolerance, it is auspicious to measure the effectiveness thereof in practice. The influence of Redundancy, as well as the effect of intelligence, reasoning and learning of the system should be investigated. Reliability is defined as the probability whereby a system will perform specified function(s), within specified conditions, without failure in the expected lifetime thereof (Kannan and Parker, 2006).

In order to measure the effects of fault tolerance on a system a variety of metrics have been instituted by Kannan and Parker (2006). To establish the reliability of a system, the Mean Time Before Failure (MTBF) model is suggested. This is defined as follows:

$$\frac{Number\ of\ Hours\ in\ Use}{Number\ of\ Failures\ Encountered}$$

Mean Time Taken to Repair (MTTR), as well as the metric for availability can be used to assist. The MTTR metric is defined as:

$$\frac{Number\ of\ Hours\ Spent\ Repairing}{Number\ of\ Repairs}$$

Availability can be defined as:

$$\frac{MTBF}{(MTTR + MTBF)} * 100$$

It is vital to consider whether a system has the ability to tackle failures and learn from results of previous failures, as this will determine the intelligence exhibited by the system (Kannan and Parker, 2006). When evaluating a system, the efficiency of the system in its entirety and the allocation of employable resources should be determined. The robustness of a system and the manner in which faults are identified and the subsequent recovery thereof, are also of importance. The ability of a system to learn and thus adapt to a magnitude of changes within a dynamic environment should be documented (Kannan and Parker, 2006).

It is assistive to classify the types of system faults by including known faults, which can be anticipated, unknown faults, which can be diagnosed by a system based on experience, and finally undiagnosable faults, that cannot be classified autonomously and thus require human intervention. The known and unknown faults may have redundancy implemented to ensure the system can recover after these faults. However, the undiagnosable faults will require human assistance (Kannan and Parker, 2006). Three states of a system are identified by Kannan and Parker, 2006, namely a normal state, fault state and recovery state.

In an attempt to quantify fault tolerance and the value of redundancy, the concept of Fuzzy Logic is to be employed. Fuzzy Logic is a form of many valued Logic whereby the truth values of variables may be any real number between 0 and 1, both inclusive (Hennessy and Patterson, 2012). Points of failure may differ and thus a decimal point will assist with accuracy. This is owing to the difficulties associated with distinguishing between what is acceptable and what is not.

5. Methodology

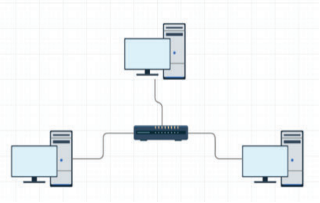
The investigation was divided into three main phases, namely building the system, the exploration of fault-inducing methods and data collection.

For this investigation a descriptive approach, accompanied by characteristics of explanatory research, were utilised. The study featured the implementation of a network of computers that was based on a Distributed Architecture. This network represented a simplified version of a technical system that would typically be found in practice and would thus enable the gathering of industry-relatable results. The computers were situated in a lab for approximately six months including set up and infiltration attempts. The lab was only accessible by the authors of this paper.

Through thorough research, a system set up was determined and created. A conceptual framework for system fault tolerance, created by Walter L. Heimerdinger and Charles B. Weinstock (2019), was utilised. This was used as a guide to develop a system that would be deemed academically suitable for collecting data relating to fault tolerance. This framework provided the reliability needed to draw conclusions from the data.

5.1 Phase 1: Building the System

5.1.1 The Network



Designed with Creately

The network was formed by connecting three personal computers to a router (Apple AirPort Extreme Base Station) using RJ45 Ethernet cables. The physical design of the network could be best described as that of a star topology, which allowed for high network speeds and easy troubleshooting when compared with other network designs (Anwar, 2019).

Each computer system was standardized to possess hardware specifications as follows; 8 GB memory, an Intel core i7-4770 CPU @3.40GHz x 8 processor and a 500 GB hard drive.

Both Computer 1 and Computer 2 employed version 19.01 of the Ubuntu operating system. Computer 3 featured version 2019.4 Kali GNU/Linux. This was selected as it offers a comprehensive range of tools for attacking.

The computers were each assigned a static IP address, with each having a unique purpose in the network. Computer 1 hosted a web streaming server, firewalls, and an intrusion detection system (*Snort*). Computer 2 acted as a user that would request from the server hosted by Computer 1 to watch the video it streamed. Computer 3 posed as the malicious entity in the network, from which all attacks were orchestrated.

5.1.2 The Firewall and Filtering Rules

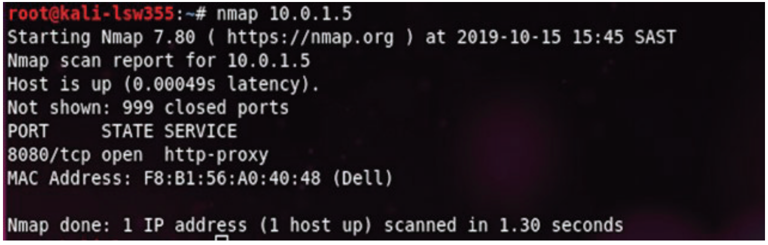
In an effort to ensure sufficient security measures were in place, an Uncomplicated Firewall (UFW) was installed alongside Mac address filtering.

Various stages were examined throughout the duration of the lifespan of this system, including the response of the system before and after the implementation of security measures. Before security measures were employed, infiltration methods were performed. Through the use of an intruder attempting to ping a network, or use a Port scan, the intruder was able to gain a variety of information about the network.
As depicted in *Figure 1*, the simulated attack entailed an intruder using the NMAP command to complete a Port scan. The intruder had access to information such as the open network (thus vulnerable Ports) as well as latency. The Mac address and the type of device in the network were also visible.

Snort rules, as well as the UFW, were installed on the network and the previous infiltration methods were then performed again. The UFW/Graphical UFW is an Ubuntu user interface for controlling the iptables (Help.ubuntu.com, 2019). The use of iptables through UFW also presented an opportunity for customised filtering and thus more control over access to the system. In order for Computer 2 to communicate with the server hosted on Computer 1, without exposing many vulnerabilities, Mac address filtering was implemented.

Iptables -I INPUT -p tcp –dPort 8080 -m mac –mac-source F8:B1:56: A0:4C:18 -j ACCEPT

The filtering rule above permitted Computer 2 to pass the firewall and send and receive packages from Port 8080 based on the computer's Mac address. This was necessary to facilitate the flow of information between the computers and gather further data on the network traffic.



```
root@kali-lsw355:~# nmap 10.0.1.5
Starting Nmap 7.80 ( https://nmap.org ) at 2019-10-15 15:45 SAST
Nmap scan report for 10.0.1.5
Host is up (0.00049s latency).
Not shown: 999 closed ports
PORT     STATE SERVICE
8080/tcp open  http-proxy
MAC Address: F8:B1:56:A0:40:48 (Dell)

Nmap done: 1 IP address (1 host up) scanned in 1.30 seconds
```

Figure 1: Nmap Scan on IP address 10.0.1.5

5.1.3 The Intrusion Detection System: SNORT

In order to intensify the security for the system and create a tangible way of observing system response when faced with faults, an intrusion detection system (IDS) was used.

Snort is a reputable intrusion detection system and allows for a custom set of rules to be created (Nayyar, 2019).

System Local *Snort* Rules:

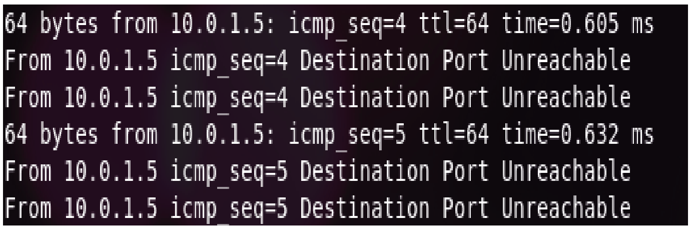
reject icmp $ EXTERNAL_NET any -> $HOME_NE any (msg: “Blocking ICMP Packets – “possible pin attempt”; sid:1000001; rev:1;)

reject tcp $ EXTERNAL_NET any -> $HOME_NET any (msg: “Blocking TCP Packets – “possible DDOS attempt”; sid:1000002; rev:1;)

Both pings, Nmap attempts, as well as DOS attacks are observed in network traffic as ICMP and TCP Packets. In the *Snort* rules above, the home net refers to the set of devices that *Snort* will attempt to protect. Moreover, the external net refers to any device that is not specified within the home network. In order to specify a home net using IP addresses, Computers 1 and 2 were both assigned static IP addresses for the duration of the investigation. The rules assessed all of the packets attempting to enter the home net to determine whether they were ICMP or TCP packets. It then accepts or rejects these packets accordingly. These packets are not always malicious in nature and if the computers were connected to the internet, the rules would be refined to account for this. All network traffic was controlled and thus it was known that ICMP or TCP would indicate a ping or DOS attempt respectively.

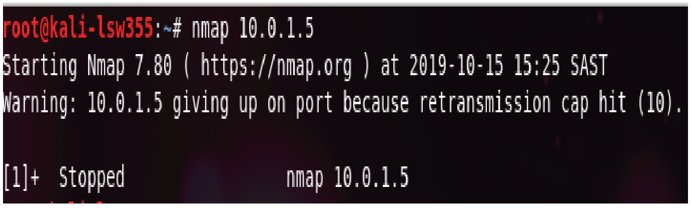
5.2 Phase 2: The Exploration of Fault Inducing Methods

Fault tolerance is centred on the notion that systems can keep operating, even if at sub-par level, when faced with failure (Rouse, 2019). Therefore, in order to assess how the system operates when experiencing failure, and how redundancy (combined with consensus-based evaluation) can have a positive impact on the systems fault tolerance, the system needed to be exposed to faults that would equate to failure.



```
64 bytes from 10.0.1.5: icmp_seq=4 ttl=64 time=0.605 ms
From 10.0.1.5 icmp_seq=4 Destination Port Unreachable
From 10.0.1.5 icmp_seq=4 Destination Port Unreachable
64 bytes from 10.0.1.5: icmp_seq=5 ttl=64 time=0.632 ms
From 10.0.1.5 icmp_seq=5 Destination Port Unreachable
From 10.0.1.5 icmp_seq=5 Destination Port Unreachable
```

Figure 2: Firewall & Snort Blocking Pings

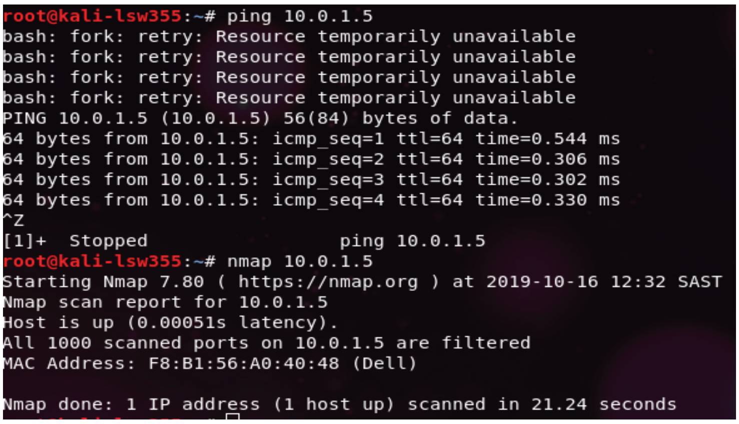


```
root@kali-lsw355:~# nmap 10.0.1.5
Starting Nmap 7.80 ( https://nmap.org ) at 2019-10-15 15:25 SAST
Warning: 10.0.1.5 giving up on port because retransmission cap hit (10).

[1]+  Stopped                 nmap 10.0.1.5
```

Figure 3: Firewall & Snort Block Port Scans

The information that was previously available, became invisible. This clearly illustrates fault tolerance and how implementing certain security methods ensures resilience and improves network redundancy against malicious intruders. *Figure 2* depicts the blocked Pings attempting to communicate with the network, whilst *Figure 3* depicts the rejected Port Scans.

As a measure to determine the need for fault tolerance, a DOS attack was performed as an intrusion method. The attack maliciously flooded the network and disrupted the regular traffic to gain control of the network. Through the use of a DOS attack, the infiltrator infected the network with malware and subsequently created a botnet. TCP packets flooded the network with an attempt to establish a three-way handshake. It is evident that through the use of this attack, as depicted in *Figures 4* and *5*, the network defences were infiltrated and weakened. The attack was initiated and the system was able to defer the packets. However, the attack was too powerful and later infiltrated the system, which thereafter lead to system failure. The Ping and Port Scan attacks were then performed and information was available. Results gathered for the DOS attack offer insight into the circumstances in which fault tolerance can be of value when attempting to mitigate the failure caused by such attacks. Pentmenu was used to perform this attack.

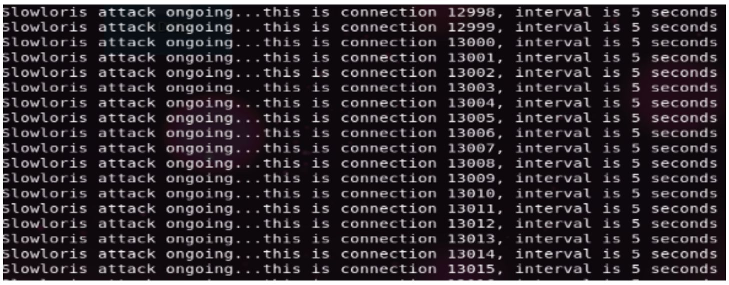


```
root@kali-lsw355:~# ping 10.0.1.5
bash: fork: retry: Resource temporarily unavailable
bash: fork: retry: Resource temporarily unavailable
bash: fork: retry: Resource temporarily unavailable
bash: fork: retry: Resource temporarily unavailable
PING 10.0.1.5 (10.0.1.5) 56(84) bytes of data.
64 bytes from 10.0.1.5: icmp_seq=1 ttl=64 time=0.544 ms
64 bytes from 10.0.1.5: icmp_seq=2 ttl=64 time=0.306 ms
64 bytes from 10.0.1.5: icmp_seq=3 ttl=64 time=0.302 ms
64 bytes from 10.0.1.5: icmp_seq=4 ttl=64 time=0.330 ms
^Z
[1]+  Stopped                 ping 10.0.1.5
root@kali-lsw355:~# nmap 10.0.1.5
Starting Nmap 7.80 ( https://nmap.org ) at 2019-10-16 12:32 SAST
Nmap scan report for 10.0.1.5
Host is up (0.00051s latency).
All 1000 scanned ports on 10.0.1.5 are filtered
MAC Address: F8:B1:56:A0:40:48 (Dell)

Nmap done: 1 IP address (1 host up) scanned in 21.24 seconds
```

Figure 4: Pings and Port Scans going through during attack

5.2.1 Pentmenu

Pentmenu is a Bash Script based on the principles of Pentbox (Penetration Testing, 2019). It was created to offer an intuitive process from which to implement Network Penetration Testing, including orchestrating network attacks. The Kali operating system installed on Computer 3 was equipped with the software required for this framework. Pentmenu has DOS modules including ICMP Echo Flood and ICMP Blacknurse (Zion, 2019). Both Modules use either hping or hping3 to launch traditional ICMP floods against specified targets. These attacks specifically strain CPU's (Mirkovic and Reiher, 2004) which has the potential to cause many forms of faults and lead to failure. Pentmenu was used for the DOS attack.



```
Slowloris attack ongoing...this is connection 12998, interval is 5 seconds
Slowloris attack ongoing...this is connection 12999, interval is 5 seconds
Slowloris attack ongoing...this is connection 13000, interval is 5 seconds
Slowloris attack ongoing...this is connection 13001, interval is 5 seconds
Slowloris attack ongoing...this is connection 13002, interval is 5 seconds
Slowloris attack ongoing...this is connection 13003, interval is 5 seconds
Slowloris attack ongoing...this is connection 13004, interval is 5 seconds
Slowloris attack ongoing...this is connection 13005, interval is 5 seconds
Slowloris attack ongoing...this is connection 13006, interval is 5 seconds
Slowloris attack ongoing...this is connection 13007, interval is 5 seconds
Slowloris attack ongoing...this is connection 13008, interval is 5 seconds
Slowloris attack ongoing...this is connection 13009, interval is 5 seconds
Slowloris attack ongoing...this is connection 13010, interval is 5 seconds
Slowloris attack ongoing...this is connection 13011, interval is 5 seconds
Slowloris attack ongoing...this is connection 13012, interval is 5 seconds
Slowloris attack ongoing...this is connection 13013, interval is 5 seconds
Slowloris attack ongoing...this is connection 13014, interval is 5 seconds
Slowloris attack ongoing...this is connection 13015, interval is 5 seconds
```

Figure 5: The Slowloris attack in progress

5.3 Phase 3: Data Collection

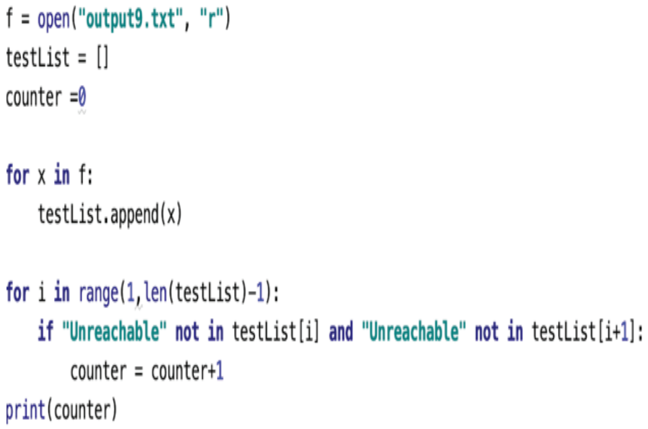
Faults at CPU level will cause various failures across the system. If rules defined for *Snort* were not being executed as expected at any point during the life of the system, a point of failure was reached.

Three DOS attacks were launched whereby the number of connections increased from 1000 to 2000, then to 4000, with an interval level of 10. Pings were orchestrated throughout the duration of the attack. In the event of a ping being recorded, a failure was recorded. In a situation whereby the *Snort* rules were active, the ping would be rejected with an alert stipulating "Port unreachable". In the situation that "Port unreachable" was consecutively reached in the attack, the system was considered stable. However, if the *Snort* rules failed, the ping would have been successful. The attack concluded and the system returned to its normal state.

A Python script *(Script 1)* was drafted and utilised to search the attack logs, then identify and quantify the failures. The script opened the text file and saved each line as an item in a list using a *for loop* The second *for loop* ran over this item in the list and if there were two consecutive items without "unreachable" being displayed, the failure counter was increased by one.

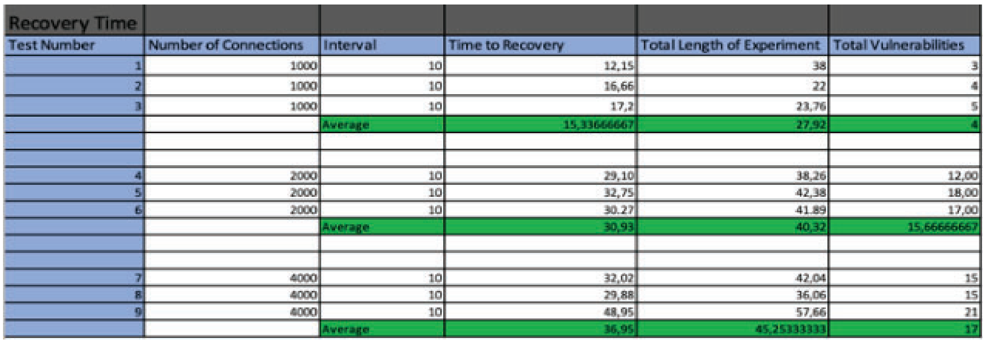
6. Analysis

The tables below detail the data obtained, which is subsequently analysed below.



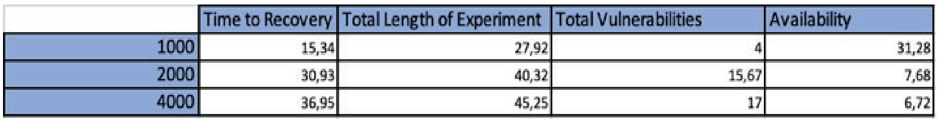
```
f = open("output9.txt", "r")
testList = []
counter =0

for x in f:
    testList.append(x)

for i in range(1,len(testList)-1):
    if "Unreachable" not in testList[i] and "Unreachable" not in testList[i+1]:
        counter = counter+1
print(counter)
```

Script 1: Python script to count vulnerabilities



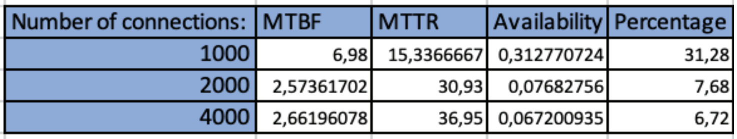
Recovery Time					
Test Number	Number of Connections	Interval	Time to Recovery	Total Length of Experiment	Total Vulnerabilities
1	1000	10	12,15	38	3
2	1000	10	16,66	22	4
3	1000	10	17,2	23,76	5
		Average	15,33666667	27,92	4
4	2000	10	29,10	38,26	12,00
5	2000	10	32,75	42,38	18,00
6	2000	10	30.27	41.89	17,00
		Average	30,93	40,32	15,66666667
7	4000	10	32,02	42,04	15
8	4000	10	29,88	36,06	15
9	4000	10	48,95	57,66	21
		Average	36,95	45,25333333	17

Table 1: Complete data collection



	Time to Recovery	Total Length of Experiment	Total Vulnerabilities	Availability
1000	15,34	27,92	4	31,28
2000	30,93	40,32	15,67	7,68
4000	36,95	45,25	17	6,72

Table 2: Averages of different severity



Number of connections:	MTBF	MTTR	Availability	Percentage
1000	6,98	15,3366667	0,312770724	31,28
2000	2,57361702	30,93	0,07682756	7,68
4000	2,66196078	36,95	0,067200935	6,72

Table 3: MTBF, MTTR calculations

MTBF, MTTR and the Availability can now be examined to quantify the results of the experiment. These metrics allow for estimation and for an indication of reliability in a quantitative manner (Hennessy and Patterson, 2012).

6.1 The Mean Time Before Failure Metric

The MTBF metric indicates the duration of the system's ability to overcome faults before failure wherein it is visible that faults are effective and produce constraints for the system.

Number of connections	Length	Failures Occurred	MTBF in Seconds
1000	27.92	4	6.98
2000	40.32	15.67	2.57
4000	45.25	17	2.66

This impact will differ between companies and thus the level of acceptable downtime will be determined by the companies themselves.

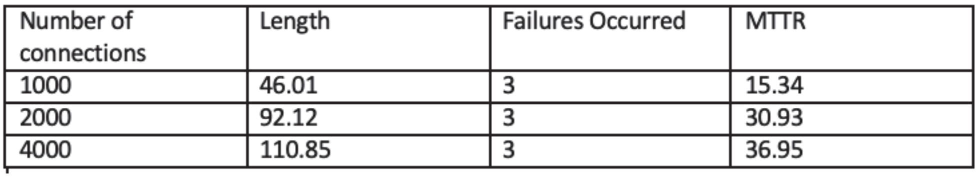
The results above indicate a 63.18% decrease in the MTBF from the first attempt average, to the second. Furthermore, there was an increase of 3.5% from the second attempt to the third, far less than the initial gap between attempt one and attempt two. Therefore, it is evident what the system is capable of and the consequences to be expected once the computer has near reached capacity.

As specified previously, the Pentmenu DOS attack placed high strain on the CPU. When the system experienced states of failure, the activity monitor indicated high CPU usage across all threads. Therefore, increased Mean Time to Failure and consequently decreased reliability can be attributed to the strain experienced by the CPU as a result of a DOS attack that occupies the computational resources needed to perform at a desired level.

6.2 The Mean Time to Repair Metric:

The MTTR metric is defined as:

$$\frac{Number\ of\ Hours\ Spent\ Repairing}{Total\ Failures\ (Repairs\ needed)}$$



Number of connections	Length	Failures Occurred	MTTR
1000	46.01	3	15.34
2000	92.12	3	30.93
4000	110.85	3	36.95

The table above indicates an 101.63% increase in recovery time from attempt one to attempt two, and a 19.46% increase from attempt two to attempt three. The processing power as well as various other resources are not increasing and thus it takes longer to deal with a heavier load and revert back to a normal state. The longer a system is in recovery for, the longer it is vulnerable and thus more likely to encounter problematic scenarios.

6.3 The Metric for Availability

Below are the calculations for the Availability metric which make use of the workings calculated from the MTTR and the MTBF metrics.

1000 connections:
*(6.98 / 15.34 + 6.98) * 100 = 31.27%*
2000 connections:
*(2.57 / 30.93 + 2.57) * 100 = 7.67%*
4000 connections:
*(2.66 / 36.95 + 2.66) * 100 = 6.71%*

From the above results we can see a 75.47% decrease in availability from attempt one to attempt two as well as a 12% decrease from attempt two to attempt three. This indicates that the availability of the system decreases as the severity of the attack increases. Thus, the availability of system processes decrease as the attack becomes more demanding in terms of resources. *Figure 6* created in *R* is available below to indicate this. This displays how the frequency tends to increase as the level of severity increases for the total length, time to recovery and total vulnerabilities. The frequency of availability drops exponentially as the level of severity increases, indicating the inverse relationship.

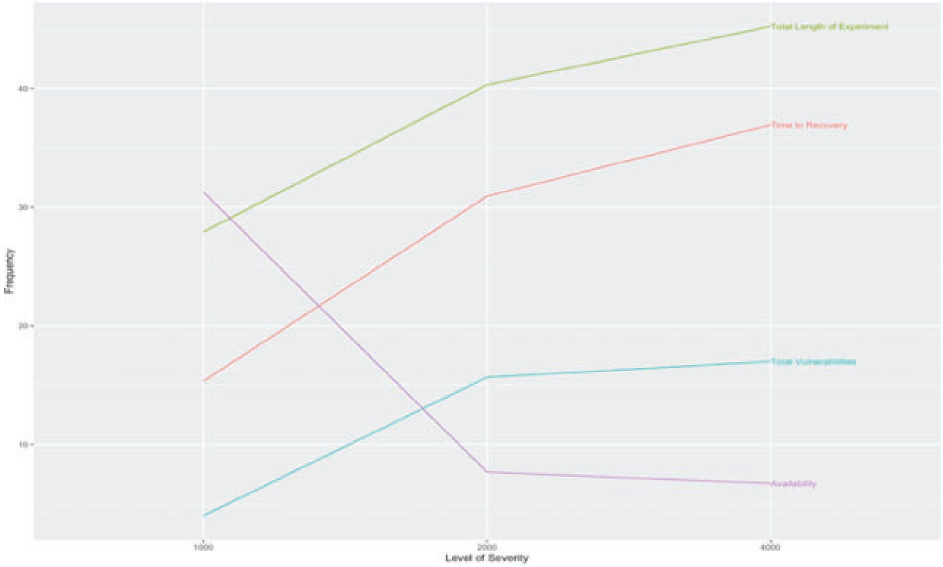


Figure 6: Severity vs. Availability analysis

7. Recommendations

The results gathered yielded sufficient evidence to support the notion that fault tolerance is critical for technological systems. Load Balancing and Byzantine Fault Tolerance are hereafter recommended.

Redundancy, as a solution to achieving fault tolerance, was introduced by John von Neumann in the 1950's and has proven to be successful (Han et al., 2005). However, the development of Cloud Technologies and advancements in hardware and software components, have allowed for new and improved methods of implementing redundancy with a multitude of faults (Cheraghlou, Khadem-Zadeh and Haghparast, 2015). Advancements in consensus-based technologies also indicate possibilities of implementing revised Byzantine Models (Block Chain will assist in this). The implementation of load balancing and Byzantine fault tolerance are key to a successful fault tolerance framework.

7.1 Load balancing

The attacks resulted in failure. However, in reality there were many factors that lead to the same faults and thus failure. Businesses that experience downtime face pecuniary losses, whereas entities such as hospitals might lose the ability to provide life critical support systems. 42% of IT security professionals from the Info Security conference in 2019 stated that there was loss of consumer trust owing to issues which arose from DDOS attacks. Moreover, 26% of the professionals listed data theft as the most damaging occurrence. (Hill, 2019)

Load balancing offers a viable answer to the complications that arise from overwhelming amounts of network traffic and lack of computational resources, regardless of the reason for failure. It also allows for effective distribution of incoming traffic across a group of other servers with high performance (NGINX, 2019). The implementation of such a feature allows redundancy along with scalability and ensured efficiency that equates to system reliability and availability.

Cloud load balancing overcomes some of the limitations posed by localised load balancing as it provides access to more server pools located in different geographical areas. Cloud-based solutions also allow for cost saving due to scalability. There are fewer limits with regards to how many resources a system can draw. This form of load balancing also mitigates risks associated with geographical areas such as power outages or other natural disasters. Load balancing accompanied by cloud-based technologies proves to be a viable method of redundancy to achieve fault tolerance when dealing with failures caused by network congestion and lack of computational resources.

7.2 Byzantine Fault Tolerance

A major problem that systems with a Distributed Architecture design have is achieving overall system reliability when faced with faults (Konstantopoulos, 2019). In a distributed architecture system, its functionality is spread across various components, each of which can be subjected to faults that could lead to a system failure

(Hennessy and Patterson, 2012). It is therefore important to have accountability amongst components to ensure correct and desired performance for the system as a whole. Achieving this consensus adds to the system's ability to be fault tolerant.

Byzantine fault tolerance ensures that at least two thirds of the actors in a system display the same data, therefore reaching a consensus that the system is, or is not, performing as desired (Konstantopoulos, 2019). This ensures that if one of the actors in the system is compromised it can be identified and proactive. Corrective measures can be initialised to return the system to the desired state. A functional way in which such a model could be implemented in a network would be to make use of Block Chain technology.

Block chain sees the implementation of a decentralised ledger, which is created and updated by the components a system is comprised of (Zainuddin, 2019). In order to achieve consensus and thus produce a validated block for the ledger, a consensus mechanism is used. There are many different options with regards to available consensus mechanisms, each of which have advantages and disadvantages. Practical Byzantine Fault Tolerant (PBFT) is an industry recommended consensus mechanism (Ismail, 2019). Through the use of this mechanism, all of the nodes in a system communicate and try to come to an agreement about the state of the system. This is based on what the majority reports back. It is highly evident that through the use of redundancy and the implementation of load balancing as well as Byzantine fault tolerance, a step toward a hybrid fault tolerance framework can be achieved.

8. Conclusion

System development has lead to a rise in fault occurrence. The margin for system related error is smaller than ever owing to system dependency. The necessity for a hybrid fault tolerance framework was analysed through the creation of a system with security measures emplaced and cyberattacks having been performed. The security measures, featuring an industry standard firewall and *Snort* rules, were infiltrated by denial of service attacks. The system experienced multiple counts of failure and was heavily influenced, as illustrated by the metrics used.

The robustness of systems will improve through the implementation of a hybrid fault tolerance framework with improved recovery times when faced with cyberattacks. It is evident that consensus-based methods and redundancy will aid the hybrid fault tolerance framework. Information systems need to be equipped with fault tolerance methods and a hybrid fault tolerance framework will ensure best practice and resilience. Further practical research into this topic can therefore include the implementation into systems, patches and other software improvements. The exploration of fault tolerance on a Social Engineering front as well as other emerging technologies, such as the Internet of Things, may also be explored. Limitations on resources, including the number of computers involved, as well as time limitations were faced. A larger architecture would have better simulated a real world environment. Further attacks could have been measured given more time. However, sufficient proof was gathered to prove that a hybrid fault tolerance framework, featuring redundancy and consensus-based methods, will aid information systems and cybersecurity.

Acknowledgments

For their guidance and assistance throughout the process of this paper we acknowledge Rudolph Le Roux, Malcolm Campbell, Kayleigh Campbell and Bruce Watson.

References

Anwar, S. (2019) What is Star Topology? | Advantages and Disadvantages (Online). Available: https://computernetworktopology.com/star-topology-advantages-disadvantages/ (2019, November 06).

Ball, R., Fink, G. and North, C. (2004) Home-Centric Visualization of Network Traffic for Security Administration. 55-64.

Bezdek, J. (1992) Fuzzy Models - What Are They, and Why?. IEEE Transactions on Fuzzy Logic, 1(1):1-6.

Castro, M. and Liskov, B. (2002) Practical Byzantine Fault Tolerance and Proactive Recovery. ACM Transactions on Computer Systems, 20(4):398–461.

Cheraghlou, M., Khadem-Zadeh, A. and Haghparast, M. (2015) A Survey of Fault Tolerance Architecture in Cloud Computing. Journal of Network and Computer Applications.

Geist, R. and Trivedi, K. (1990) Reliability estimation of fault-tolerant systems: tools and techniques - IEEE Journals & Magazine. (Online) Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=56852 (2019, June 14).

Han, J., Gao, J., Qi, Y. and A.B. Fortes, J. (2005) Toward Hardware-Redundant, Fault-Tolerant Logic for Nanoelectonics. IEEE Design & Test of Computers, 328-339.

Helmerdinger, W. and Weinstock, C. (2019) A Conceptual Framework for System Fault Tolerance. AD-A258467, 1-33.

Hennessy, J. and Patterson, D. (2012) Computer Architecture - A Quantitative Approach. 5th ed. Waltham, USA: Elsevier.

Hill, M. (2019) Loss of Customer Trust and Confidence Biggest Consequence of DDoS Attacks. (Online). Available: https://www.infosecurity-magazine.com/news/loss-trust-confidence-ddos/ (2019, November, 06).

Ismail, N. (2019) The gold standard for security in distributed systems like Blockchain. (Online). Available at: https://www.information-age.com/gold-standard-security-distributed-systems-like-blockchain-123469395/ (2019, November, 06).

Kannan, B. and Parker, L. (2006) Fault-Tolerance Based Metrics for Evaluating System Performance in Multi-Robot Teams. Performance Metrics for Intelligence Systems Workshop. IEEE International Workshop on Safety, and Rescue Robotics. (Electronic). Available: https://ws680.nist.gov/publication/get_pdf.cfm?pub_id=823600#page=56 (2019, June 16).

Konstantopoulos, G. (2019) Understanding Blockchain Fundamentals, Part 1: Byzantine Fault

Tolerance. (Electronic) Available: https://medium.com/loom-network/understanding-blockchain-fundamentals-part-1-byzantine-fault-tolerance-245f46fe8419 (2019, June 16).

Kotulic, A and Clark, J.G. (2004) Why There Aren't More Information Security Research Studies, Information & Management, 41, 597-607. Available: http://130.18.86.27/faculty/warkentin/SecurityPapers/Robert/KutolicClark2004_I&M_41_5_risk_management.pdf (2019, June 21)

Mirkovic, J. and Reiher, P. (2004) A Taxonomy of DDoS Attack and DDoS Defense Mechanisms. ACM SIGCOMM Computer Communications Review, 34(2):39-54.

Muelder, C., Ma, K. and Bartoletti, T. (2005) Interactive Visualization for Network and Port Scan Detection. RAID 2005: Recent Advances in Intrusion Detection, 265-283.

Nayyar, Dr Anand. (2019) The Best Open Source Network Intrusion Detection Tools. (Online) Available: https://opensourceforu.com/2017/04/best-open-source-network-intrusion-detection-tools/ (2019, October 30).

NGINX. (2019) What Is Load Balancing? How Load Balancers Work. (Online) Available: https://www.nginx.com/resources/glossary/load-balancing/ (2019, November 6).

Rouse, Margret. (2019) What Is Fault-Tolerance?. (Online) Available: https://searchdisasterrecovery.techtarget.com/definition/fault-tolerant (2019, November 6).

Steinberg, Joseph. (2019) Small Businesses Beware: Half of all Cyber-Attacks Target You. (Online) Available: https://www.inc.com/joseph-steinberg/small-businesses-beware-half-of-all-cyber-attacks-target-you.html (2019, November 6).

Taylor, Hugh. (2019) What Are Cyber Threats: How They Affect You and What to Do About Them - The Missing Report. (Online) Available: https://preyproject.com/blog/en/what-are-cyber-threats-how-they-affect-you-what-to-do-about-them/ (2019, November 5).

Zainuddin, Aziz. (2019) The Fundamentals of Blockchain: Byzantine Fault Tolerance Systems - eToroX. (Online) Available: https://www.etorox.com/news/opinions/the-fundamentals-of-blockchain-byzantine-fault-tolerance-systems/ (2019, November 6).

Zion, R. (2019) Pentmenu - A simple Bash Script for Recon and DOS Attacks. (Online) Available: https://www.hacking.land/2016/08/pentmenu-simple-bash-script-for-recon.html?m=1 (2019, November 6).

Mobile Forensics: Beyond Traditional Sources of Digital Evidence

Heloise Pieterse
Council for Scientific and Industrial Research, Pretoria, South Africa
hpieterse@csir.co.za

DOI: 10.34190/EWS.20.042

Abstract: Mobility is the future and people of the 21st century are continuously witnessing the fast-paced growth of mobile technology. The ever-increasing storage capacity of mobile devices allows for the capturing of the user activities in digital format. Traditionally, such digital data include contacts, text and instant messages, call history, electronic mail, web browsing history, documents and geographical data. These rich sources of digital data present on mobile devices become increasingly important when mobile devices are linked to civil or criminal digital investigations. However, these sheer quantities of traditional digital data available on mobile devices often cause other forms of noteworthy digital data to go unnoticed. This paper investigates and identifies other available sources of digital data present on mobile devices that can be of value to digital forensic investigations. The study focuses exclusively on the Android operating system and presents an extensive evaluation of Android's file system. Furthermore, the study aims to locate, extract and utilise non-traditional or contemporary sources of digital data, such as log files, usage statistics and event data, as potential digital evidence in civil or criminal digital investigations. The outcome of the study leads to the construction of the new Pre-Analysed Device Snapshot (PADS) model, which provides a summary of the current state of the mobile device at the time of acquiring the device.

Keywords: Digital Forensics, Mobile Forensics, Android, Smartphones, Digital Evidence, Data Analysis, Modelling

1. Introduction

Mobile devices, such as smartphones and tablet computers, have become an integral part of everyday living in the 21st century. These devices are constant companions and provide various advanced capabilities to allow end-users to perform a wide range of activities. Examples of such capabilities include cost-free communication mechanisms (e.g. Wi-Fi, Bluetooth and Near Field Communication), improved multimedia features (e.g. video recording, music playback and high-quality display) and the possibility of downloading additional applications with added functionality. These capabilities are a direct result of ever-improving hardware components and the evolution of mobile operating systems.

The reliance on and ubiquitous use of mobile devices by end-users, in conjunction with their ever-increasing storage capacity, allow for large quantities of digital data to reside in internal and external storage spaces. Such digital data includes contacts, text and instant messages, call history, geographical data, electronic mail, web browsing history and multimedia activities (Ayers et al., 2013). The evidential value offered by such traditional sources of digital data becomes increasingly important when mobile devices form part of criminal or illegal activities (Barmpatsalou, 2018). The traditional sources of digital data can help digital forensic professionals, the individual responsible to acquire, collect, preserve and analyse digital data retrieved from mobile devices (Hoelz & Maues, 2017), to form hypotheses and answer crucial questions during an investigation (Casey, 2011).

Although of great value, the sheer volume of the traditional sources of digital data often causes other contemporary forms of digital data to go unnoticed. Therefore, the purpose of this paper is to identify and discuss contemporary forms of digital data that can be potentially used as digital evidence during investigations. In order to simplify the use of contemporary digital data sources by digital forensic professionals, the paper also introduces the new Pre-Analysed Device Snapshot (PADS) model. The PADS model attempts to complement existing mobile forensic investigations by offering digital forensic professionals with a snapshot of the current state of the mobile device at the time of acquiring the device. The information produced by the PADS model will then offer guidance and direction to the digital forensic professionals during the investigation.

The efficiency of the PADS model is established by evaluating a collection of existing contemporary data present on a mobile device. The focus will be exclusively on the Android operating system since Google Android holds 74.17% of the mobile operating system market share as of December 2019 (StatsCounter, 2019). While the possibility exists to include other mobile operating systems such as iOS, the current dominance of Google Android justify the emphasis on this particular operating system. The focus of the study will be to locate, extract and utilise non-traditional or contemporary sources of digital data, such as log files, usage statistics and event data, present on Android smartphones. Such data will be provided as input to the PADS model, which in return will produce output that can be of use to an investigation as potential digital evidence.

The remainder of this paper is structured as follows. Section 2 briefly presents the forensic investigation of mobile devices and discusses the Android operating system, focussing on the current file system used by Android, as well as the most common partitions presents on Android smartphones. In Section 3, both traditional and contemporary sources of digital data are introduced and considered as potential digital evidence. Section 4 introduces the PADS model and illustrates the utilisation of contemporary digital data sources as forms of digital evidence. Section 5 concludes the paper.

2. Background

The expeditious development and growth of mobile technology continue to drive the evolution of mobile technology. These improvements coupled with their popularity among end-users ensure these devices collect and hold vast amounts of digital data. The following subsections further discuss the forensic investigation of mobile devices, with special attention given to the Android operating system as a platform often forming part of digital forensic investigations.

2.1 Mobile device forensics

Mobile devices are characterised as portable, lightweight and compact devices that provide personal computer-like functionality. Both the continuous improvements of mobile device technology coupled with the popularity of the devices among end-users ensure vast collections of digital data. Such data can become valuable sources of digital evidence, should the mobile device be linked to criminal or illegal activities. Therefore, mobile device forensic emerged to assist with the forensic investigation of mobile devices and is best described as "the science of recovering digital evidence from a mobile device under forensically sound conditions using accepted methods" (Ayers et al., 2013). The ability to capture and analyse the data present on mobile devices provides digital forensic professionals with powerful investigative capability. Attaining the available data requires of the digital forensic professional to follow the established digital forensic process, which consists of the following phases (Du, Le-Khac and Scanlon, 2017; Valjarevic & Venter, 2012):

- Planning Phase: involves the development of relevant procedures, identification of tools to be used, planning for use of appropriate human resources and the planning of all activities during other phases, which can include preparation of relevant equipment (software and hardware), infrastructure, human resources, raising awareness, training and documentation.
- Search and Seizure Phase: performed at the scene of an incident and requires the proper documentation of the entire incident scene.
- Acquisition Phase: acquiring and collecting (forensic imaging) the evidence without altering or damaging the original evidence.
- Preservation Phase: ensures the safety of collected evidence, as well as the transportation and storage of the evidence
- Analysis and Examination Phase: performs the analysis of the collected evidence and includes data filtering, validation, pattern matching and searching for particular keywords. Furthermore, this phase also conducts the identification of relationships between fragments of data, analysis of hidden data, determining the significance of the information obtained, reconstructing the event data, based on the extracted data and arriving at proper conclusions.
- Reporting Phase: involves the presentation of the results in the form of a formal report.

The presented digital forensic process provides the foundation for digital forensic professionals to investigate and obtain digital evidence from mobile devices. The Android operating system is a popular platform for mobile devices and often encountered by digital forensic professionals.

2.2 Android operating system

Operating systems form the foundation of advanced capabilities and improved functionality showcased by mobile devices today. They operate seamlessly and act as the intermediary layer between the end-user and the underlying hardware.

Google Android is an open-source mobile operating system created by the Open Handset Alliance (OHA) and officially announced in November 2007 (Lessard & Kessler, 2010). Developed for platforms such as mobile devices, the popularity of the operating system continued to grow steadily over the past decade. Part of the success of the Android operating system is the use and deployment of an appropriate file system.

YAFFS2, first released in 2004, was the primary file system for Android devices until Android version 2.2 (Froyo) (Barmpatsalou et al. 2013; Vidas et al. 2011; Zimmermann et al. 2012). However, the single-threaded design of YAFFS2 caused bottlenecks in devices released with dual-core or multi-core chipset systems (Tamma & Tindall 2015; Kim & Kim 2012). Therefore, with the release of Android version 2.3 (Gingerbread), Android switched from YAFFS2 to the extended file system (EXT) version 4 (Vidas et al. 2011; Zimmermann et al. 2012). The EXT4 file system runs smoothly on dual-core or multi-core mobile devices while providing stable performance (Kim & Kim 2012). Furthermore, the EXT4 file system also divides the disk space into logical blocks or partitions, which reduces overhead and improves throughput (Kim & Kim 2012). These key features of EXT4 make the file system ideal for newly released Android devices.

Partitioning of the file system allows the logical storage blocks, or partitions, to be accessed independently of each other and organises the stored data in a structured manner. Due to the various manufacturers of Android devices, partitioning can differ across different device models. The most common partitions of Android devices are the following (Tamma & Tindall 2015, Vidas et al. 2011):

- */boot*: holds all the information and files required by the boot process of the Android device.
- */recovery*: allows the Android device to boot into the recovery console and perform updates or maintenance operations.
- */data*: contains the bulk of user data (settings and customisations), as well as installed applications and their related data.
- */system*: holds the Android operating system, which includes libraries, system binaries and pre-installed system applications.
- */cache*: stores recovery logs, downloaded update packages and frequently accessed application data.
- */sdcard*: is found across all Android devices regardless of make or model and allows end-users to store additional files and data.

Although all partitions can hold digital data that may be of value during digital forensic investigations, digital forensic professionals usually focus on the */data*, */system* and */sdcard* partitions to obtain digital evidence.

3. Sources of digital data

The operation of mobile devices by end-users and the execution of installed applications ensure digital data is generated continuously in a deterministic and undisturbed manner (Mylonas et al, 2012). The digital data generally includes pre-generated data, data created due to the usage of the smartphone and data transferred to the device by the end-user. These various sources of digital data reside in different partitions and include both static and dynamic forms of data. Regardless of the location or nature of the digital data, the data can become valuable digital evidence should the smartphone be linked to civil or criminal investigations.

The available sources of digital data can be categorised as traditional or contemporary sources and each category is further discussed in the following subsections.

3.1 Traditional sources

Traditional sources of digital data refer to data that often form part of digital forensic investigations as digital evidence. These sources usually involve data residing on the Subscriber Identity Module (SIM) card, data created and transferred to the smartphone by the end-user and data generated by the installed applications.

Information collected on SIM cards include unique identifies (integrated circuit card identifier (ICCID) and the international mobile subscriber identity (IMSI) number), subscriber-related information and authentication keys (personal identification number (PIN) and personal unblocking key (PUK)) (Ayers et al., 2013; Curran et al. 2010; Ibrahim et al., 2016). Although all information present on SIM cards can be of value during a digital forensic investigation, the repository of subscriber information available on the SIM cards is of greater importance during the analysis and examination phases of an investigation. The available subscriber information commonly found on SIM cards include phonebook entries or contacts, text messages, call information (last numbers dialled), location area identity and service-related information (Ayers et al. 2013). However, the diminished storage capacity of SIM cards limits the amount of data that resides on the cards. Current improvements of smartphone technology allow the devices to hold larger volumes of data in internal or external storage locations.

Digital data found residing in internal or external storage locations are created as a direct result of the user's interaction with the smartphone. A small collection of the digital data present on a smartphone is user-created data, best described as data created externally to the smartphone that the end-user transfer to the device using either a wired (USB) or wireless (Bluetooth) connection. User-created data will most likely comprise of pictures, photographs, videos, audio clips and document files (i.e. presentations or spreadsheets) (Dlamini et al. 2016, Quick & Choo 2016). Other forms of user-created data also include personally identifiable information (phone number and e-mail address), accounts (Google or Apple, social media and bank) and personalised configurations (smartphone settings and themes). Although small in quantity, user-created digital data remains essential digital evidence since the data offers valuable information about the end-user.

Application-related data forms the most extensive collection of digital data present on smartphones and includes any data created or generated by smartphone applications. Smartphone applications, which consist of the user (third party) and system (pre-installed) applications, create both permanent and temporal data as the application executes (Pieterse 2019). The data is collected and stored in flat files or databases and only includes data specific to the application. These applications can act as witnesses (Pieterse & Olivier, 2014) and by analysing the collected data will provide digital forensic professionals with better context regarding the use of smartphones by end-users. Both the value and volume of application-related data emphasise the importance of such data to form part of criminal or civil investigations.

These traditional sources of digital data discussed above become valuable forms of digital evidence when the smartphones form part of criminal or civil investigations. Therefore, during analysis and examination, digital forensic professionals will pay special attention to these traditional sources of digital. However, the improvement of mobile technology and the current storage capacity of smartphones ensure various other sources of digital data is also present on smartphones.

3.2 Contemporary sources

Generally, digital forensic investigations rely on and make use of the traditional sources of digital found on mobile devices. However, various other sources of digital data exist on mobile devices that can be of value during digital forensic investigations. The purpose of this section is to explore other contemporary sources of digital data that can be found on smartphones.

3.2.1 Device logs and statistics

The first contemporary source of digital data involves collections of data relating to logging and capturing of device-specific activities and statistics. These sources of data give an overview of the activities that occurred on the smartphone and usually includes the following logged activities:

- Connections made to Wi-Fi access points or hotspots by the smartphone.
- Rebooting of the smartphone.
- Captured information regarding devices errors, irregular rebooting of the smartphone and crash logs.
- Event data such as system reboots and USB connections.

Furthermore, various statistics regarding network activities (mobile network connection, data usage per application and total data transferred) and battery usage (per application) are captured at regular intervals.

Digital forensic professionals cannot overlook the potential value of available device logs and statistics found on mobile devices. Digital forensic professionals can use such data to formulate a snapshot of the state of the smartphone, determine the location of the smartphone at the time of the acquisition, or combined the accompanying timestamps to construct a timeline, which a chronological ordering of captured events and activities. For example, digital forensic professionals can use the collected logs that captured information relating to connections made to Wi-Fi access points and hotspots to determine and pinpoint the location of the smartphone at the time an incident occurred. Additionally, the available event data, such as USB connections made by the smartphone, can direct the digital forensic professional to other devices such as a desktop computer or laptop that may also contain important digital evidence. Therefore, digital forensic professionals must review digital data, such as the available device logs and statistics, to avoid the potential loss of valuable digital evidence.

3.2.2 User information and usage statistics

The second contemporary source of digital data focus exclusively on the end-user and captures information pertaining to usage of the smartphone and activities performed by the end-user. The data captured relating to the end-user usually includes the following:

- Profile information of the smartphone end-user (name and photograph).
- User accounts such as “Admin” and “Guest” and restrictions on the accounts.
- The most recent tasks or activities performed by the end-user.
- The most recent applications used by the end-user.
- User login details and timestamp.

The obtainable user information and usage statistics as listed above offer a detailed overview of the end-user, as well as the end-user’s interaction with the smartphone. Digital forensic professionals can make use of the available user information to determine which user account was active at the time of the acquisition, as well as when the end-user logged into the account. Furthermore, chronological ordering of the recent tasks, activities and applications used by the end-user outlines to the digital forensic professional the final events that occurred before the acquisition of the smartphone. As a result, the digital forensic professional can conclude which applications were used by the end-user at the time an incident occurred. Therefore, the evaluation of the acquired digital data during the analysis and examination phases of the digital forensic investigation can be directed and prioritised based on the identified applications. Such guidance and direction during the investigation can significantly decrease the time required to conclude the analysis and the extraction of digital evidence.

3.2.3 Application information and usage

The third and final source of contemporary digital data relates specifically to additional application information such as metadata and the usage of the applications. Captured data relating to application information and usage can include the following:

- Metadata pertaining to each of the installed applications.
- Permissions and settings relating to each of the installed applications.
- Log files indicating when last the installed applications were used.
- Log files capturing the usage of the applications by the end-user.

As mentioned in Section 3.1, application-related data traditionally forms a substantial part of the digital evidence obtained from smartphones. However, additional data relating to the installed applications, as well as the usage of these applications, can offer digital forensic professionals with additional insights during a digital forensic investigation. For example, the available data showcasing the application usage can indicate to digital forensic professionals which were the most used applications used by the end-user. Such information offers guidance to the digital forensic professionals and provides direction to the digital forensic investigation during the analysis and examination phases.

4. The utilisation of contemporary digital data sources as digital evidence

Section 3 introduced and discussed both traditional and contemporary sources of digital data, highlighting the value of such data to digital forensic investigations. Currently, the utilisation of traditional digital data sources as forms of digital evidence is well understood and established. Various mobile forensic toolkits (e.g. Cellebrite and Mobile Phone Examiner Plus) exist to assist digital forensic professionals with the acquisition, preservation and analysis of digital data. Even though such mobile forensic toolkits support the acquisition of both traditional and contemporary sources of digital data from mobile devices, the analysis of the digital data is more focused on traditional sources. Overlooking the contemporary sources of digital data can cause digital forensic professionals to disregard valuable digital evidence. It becomes even more important for digital forensic professionals to consider and review other contemporary sources of digital data found on mobile devices. Therefore, the opportunity exists for the creation of a new model that focuses specifically on the processing of contemporary digital data sources with the aim of obtaining digital evidence.

The aim of this section is to introduce a new model, called Pre-Analysed Device Snapshot (PADS) that can evaluate and utilise contemporary digital data sources. The purpose of the PADS model is to provide an overview of the mobile device at the time acquisition is performed, allowing the digital forensic professional to attain a snapshot of the current state of the mobile device. The remainder of this section discusses the inner functionality of the PADS model, as well as the practical evaluation of the PADS model using contemporary digital data collected from an Android mobile device.

4.1 Pre-Analysed Device Snapshot (PADS) model

The high-level design of the PADS model follows the input-process-output (IPO) design pattern. As input, the PADS model can receive either structured (e.g. Extensible Markup Language (XML) or log files) or unstructured (e.g. plain text files) forms of digital data. Processing of the digital data accepted as input follows four distinct steps: (1) parsing of the input into logical components, (2) extraction of the relevant logical components, (3) formulation of the extracted logical components in collections of applicable information and (4) construction of purposeful output using the obtained information with the intent to be used as potential digital evidence. The provided output group the information according to three distinct categories: (1) User, (2) Device and (3) Apps. Figure 1 illustrates the process flow of the PADS model.

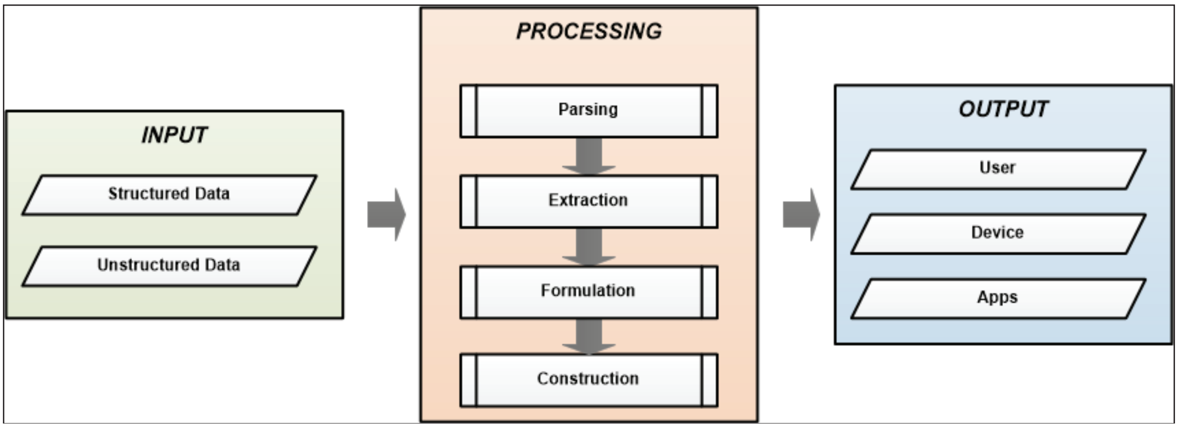


Figure 1: Illustration of the PADS Model

The PADS model expects to receive as input either structured or unstructured data sources. These data sources are contemporary digital data found on mobile devices and will include, for example, device log files and usage statistics. Once received, the PADS model proceeds to the processing phase. The parsing step is responsible to divide the received input into logical syntactic components that can easily be analysed. The extraction step will review all parsed components and extract the relevant components per pre-defined rules. These rules guide the processing of the PADS model to extract only information, such as timestamps of logged events, logged on user and last used applications, which can be of value as digital evidence. The formulation step is responsible to group the collected information into appropriate categories that best describe the state of the mobile device at the time acquisition is performed. Finally, the construction step utilises visual structures, such as timelines and graphs, to communicate the obtained information in an expressive manner. The categories forming the final output of the PADS model are the following:

- **User**: captures user-related information, such as the logged-on user and last activities conducted by the end-user. Results are illustrated using textual and visual (timeline) formats.
- **Device**: captures all events logged by the mobile device. Results are illustrated using a visual (timeline) format.
- **Apps**: captures and presents the usage of the applications by the end-user. Results are illustrated using a visual (graph) format.

The following section demonstrates the practical evaluation of the PADS model using inputs obtained from an Android mobile device.

4.2 Practical evaluation of the PADS model

The newly introduced PADS model offers needed assistance to digital forensic professionals by capturing the current state of the mobile device at the time acquisition is to be performed. However, confirming the effective use of the PADS model requires the practical evaluation of a mobile device. The test mobile device is a Samsung Galaxy S5 Mini running Android version 6.0.1 (Marshmallow). The practical evaluation involves the acquisition of the contemporary digital data sources from the test mobile device and the manual processing of the acquired data sources.

Using the test mobile device, contemporary digital data sources are acquired using traditional mobile forensic acquisition techniques. The acquired contemporary digital data sources are listed and described in Table 1.

Table 1: Collected contemporary digital data sources

Filename	Location	Description
1576540800000	/data/system/usagestats/0/daily	Plain text file capturing the usage statistics of the system applications for the past 24 hours. The file name is a timestamp that
1576108800000	/data/system/usagestats/0/weekly	Plain text file capturing the usage statistics of the system applications for the past week.
0.xml	/data/system/users	XML file capturing the user profile of the logged in user.
927_task.xml	/data/system/recent_tasks	XML file capturing information relating to the last activity performed by the logged in user.
926_task.xml	/data/system/recent_tasks	XML file capturing information relating to the second last activity performed by the logged in user.
925_task.xml	/data/system/recent_tasks	XML file capturing information relating to the third last activity performed by the logged in user.
SYSTEM_BOOT@1576575153041	/data/system/dropbox	Plain text file generated consistently at boot time, with the timestamp forming part of the file name and showing the mobile device was booted.

The collected data sources contain a combination of plain text and XML files, which will be used as input for the PADS model. Processing of the received input to attain the desired output categories is illustrated in Figure 2.

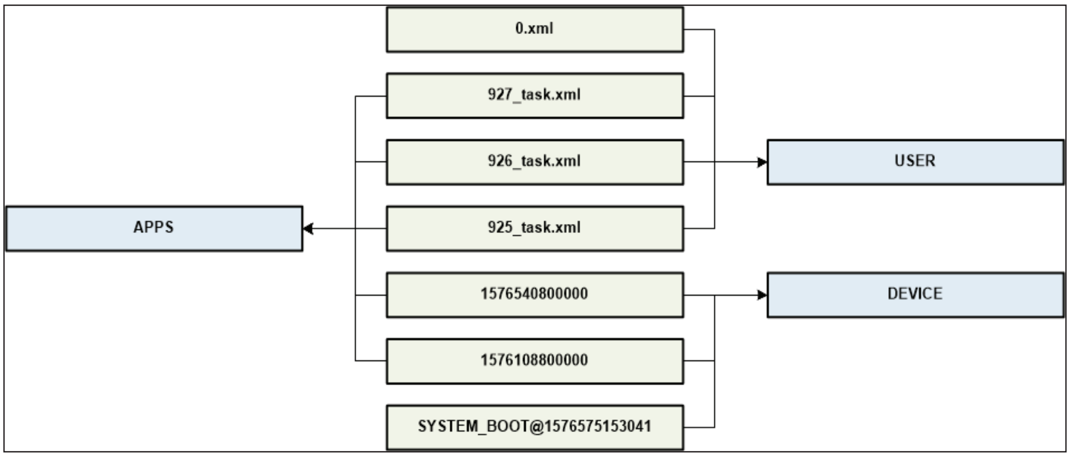


Figure 2: Selection of input to obtain the desired output

The final output produced by the PADS model is revealed in Figure 3 and presents the findings according to the categories defined in Section 4.1. All of the presented timestamps were captured in the Greenwich Mean Time (GMT) zone.

The information provided as output by the PADS model showcase the state of the test mobile device at the time the practical evaluation was performed. The following information can be deduced from the received output:

- The name of the user profile logged in during the evaluation is *Heloise*.
- The end-user only used the following three applications since booting the mobile device: *MyFiles*, *Camera* and *Gallery*.
- The end-user connected the mobile device to a computer during 9:40 and 9:42 and selected the Media Transfer Protocol (MTP) option.
- Application usage for the past 24 hours is the same as the past week, signifying the mobile device was not used prior to being switched on at 9:32:33.

Digital forensic professionals can now use the findings above to guide and direct the further analysis and examination of the digital data present on the mobile device.

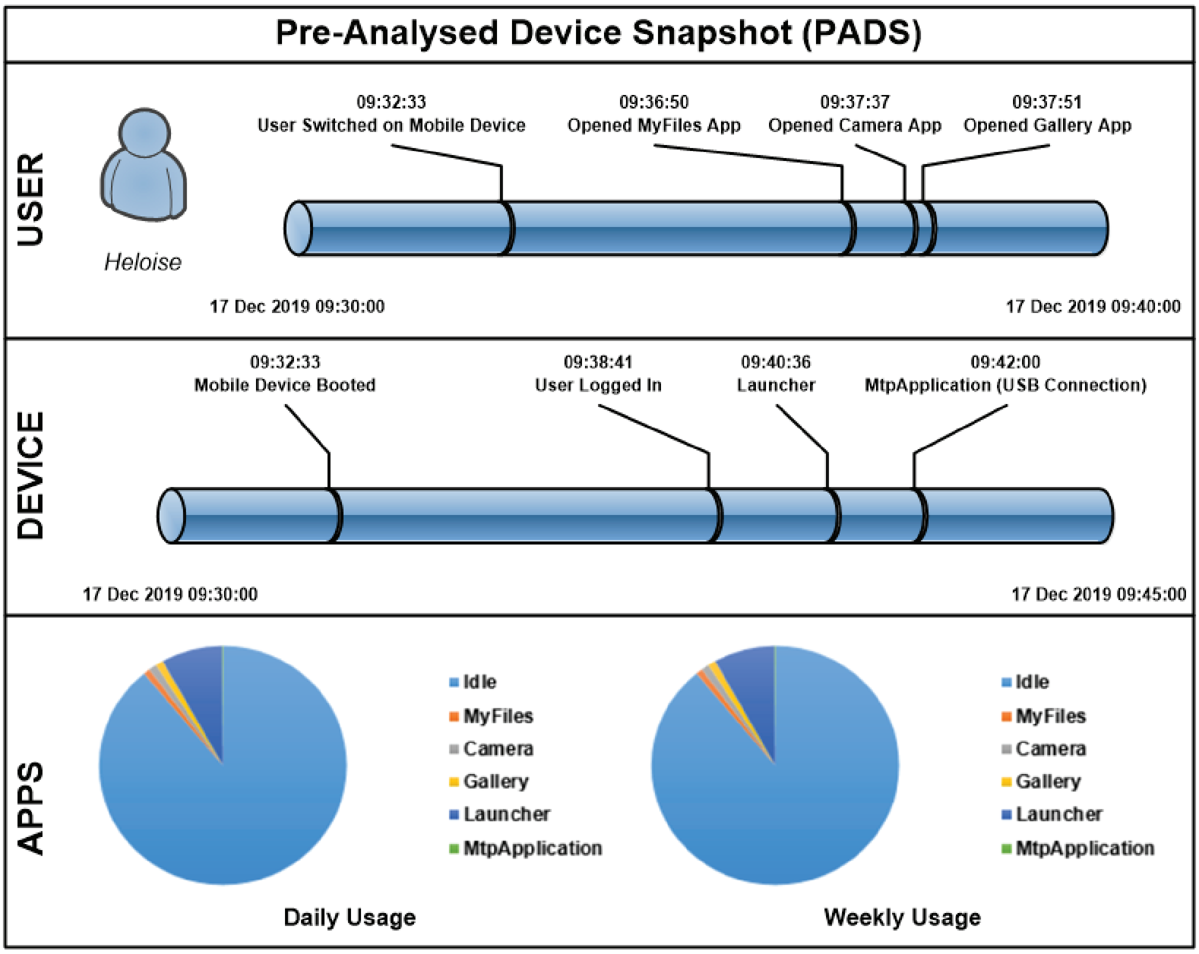


Figure 3: Visualisation of the produced output of the PADS model

5. Conclusion

As a discipline, mobile device forensics provides digital forensic professionals with the necessary concepts, processes and techniques to acquire and examine digital data obtained from devices. However, the current focus of mobile device forensics concentrates on traditional sources of digital data found on mobile devices. Although the utilisation of traditional digital data sources offers digital forensic professionals adequate results, an opportunity exists to enhance the obtained digital evidence by examining contemporary digital data sources also present on mobile devices. Therefore, this paper introduced the Pre-Analysed Device Snapshot or PADS model as a method to utilise contemporary digital data sources. The PADS model provide digital forensic professionals with an overview of the state of a mobile device at the time acquisition is performed. The output produced by the PADS model can either guide the digital forensic professional during the examination and analysis phase or be of value as digital evidence. The practical evaluation confirmed the worth of the PADS model as a method to extract additional digital evidence from mobile devices. Future work will focus on expanding the PADS model to evaluate contemporary digital data obtained from other mobile device operating systems such as iOS. Furthermore, the automation of the PADS model will also be explored to simplify the processing phase of the model.

References

Ayers, R., Brothers, S. and Jansen, W. (2013) “Guidelines on mobile device forensics (Draft)”, NIST Special Publication 800-101, Technical report, National Institute of Standards and Technology.

Barmpatsalou, K., Cruz, T., Monteiro, E. and Simoes, P. (2018) “Current and future trends in mobile device forensics: A survey”, *ACM Computing Surveys*, Vol. 51, No. 3, pp 1-31.

Barmpatsalou, K., Damopoulos, D., Kambourakis, G. and Katos, V. (2013) “A critical review of 7 years of mobile device forensics”, *Digital Investigation*, Vol. 10, No. 4, pp 323-349.

Casey, E. (2011) *Digital evidence and computer crime: Forensic science, computers, and the Internet*, 3rd ed, Academic Press, Cambridge, Massachusetts, USA.

Curran, K., Robinson, A., Peacocke, S. and Cassidy, S. (2010) "Mobile phone forensic analysis", *International Journal of Digital Crime and Forensics*, Vol. 2, No. 2, pp 15-27.

Dlamini, Z.I., Olivier, M.S. and Grobler, M.M. (2016) The smartphone evidence awareness framework for the users, in *Proceedings of the 11th International Conference on Cyber Warfare and Security, Boston, USA, 17-18 March*, pp 439-449.

Du, X., Le-Khac, N.-A. and Scanlon, M (2017) Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service, in *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS 2017)*, ACPI, pp 2876-2885.

Hoelz, B. and Maues, M. (2017) Anti-forensics threat modeling, in *Peterson G., Shenoi S. (eds) Advances in Digital Forensics XIII*, Vol. 511, Springer, Berlin, Heidelberg, pp 169–183.

Ibrahim, N., Al Naqbi, N., Iqbal, F. and AlFandi, O. (2016) SIM Card Forensics: Digital Evidence, in *Annual ADFSL Conference on Digital Forensics, Security and Law*, pp 219-234.

Kim, H.-J. and Kim, J.-S. (2012) Tuning the EXT4 file system performance for Android-based smartphones, in *Sambath S., Zhu E. (eds.) Frontiers in Computer Education*, Vol. 133, Springer, Berlin, Heidelberg, pp 745-752.

Lessard, J. and Kessler, G.C. (2010) "Android forensics: Simplifying cell phone examinations", *Small Scale Digital Device Forensics Journal*, Vol. 4, No. 1, pp 1-12.

Mylonas, A., Meletiadis, V., Tsoumas, B., Mitrou, L. and Gritzalis, D. (2012) Smartphone forensics: A proactive investigation scheme for evidence acquisition, in *Gritzalis D., Furnell S., Theoharidou M. (eds.) Information Security and Privacy Research, SEC 2012*, Vol. 376, Springer, Berlin, Heidelberg, pp. 249-260.

Pieterse, H. and Olivier M. (2014) Smartphones as Distributed Witnesses for Digital Forensics, in *Peterson G., Shenoi S. (eds.) Advances in Digital Forensics X*, Vol. 433, pp 237 - 251.

Pieterse, H. (2019) *Evaluation and Identification of Authentic Smartphone Data*. PhD Thesis. University of Pretoria.

Quick, D. and Choo, K.K.R. (2016) "Big forensics data reduction: Digital forensic images and electronic evidence", *Cluster Computing*, Vol. 19, No. 2, pp 723-740.

StatsCounter (2019) "Mobile Operating Market Share Worldwide", [online], https://gs.statcounter.com/os-market-share/mobile/worldwide/2019

Tamma, R. and Tindall, D. (2015) *Learning Android Forensics*, Packt Publishing Ltd, Birmingham, UK.

Valjarevic, A. and Venter, H.S. (2012) Harmonised digital forensic investigation process model, in *Information Security for South Africa (ISSA) Johannesburg, South Africa, 15-17 August*, IEEE, pp. 1-10.

Vidas, T., Zhang, C. and Christin, N. (2011) "Toward a general collection methodology for Android devices", *Digital Investigation*, Vol. 8, pp 14-24.

Cyber Security: Trust Based Architecture in the Management of an Organization‘s Security

Jouni Pöyhönen and Martti Lehto
University of Jyväskylä, Finland
jouni.a.poyhonen@jyu.fi
martti.j.lehto@jyu.fi

DOI: 10.34190/EWS.20.090

Abstract: The functioning of a modern society is based on the cooperation of several organisation, whose joint efficiency depends increasingly on trustable business processes. Trust is based on availability, reliability and integrity of ICT system data in the operating environment, whose cyber security risks are continuously augmented by threatening scenarios of the digital world. Information and Communications Technology (ICT) can be seen a critical asset of organization. To prevent loss of customers the trust and that way revenue and money, as well as to protect organisational reputation, this asset must be protected from cyberattacks. Trust in cyber environment is critical factor for organization in order to contentiously run the business processes that the customers can count on. This article emphasizes the system view, the trust and the trust-based architecture measures as an essential part of organization cyber security management. The article integrates several basic standards and three decision making viewpoints as an organisation cyber security architecture framework. The cyber trust factors for organization have been also defined by the researchers in this article. We have also examined how the measures can be considered part of the organization’s process structures while creating trust in its operation within a dynamic cyber environment. Thus, the article recommends and utilizes the PDCA (Plan, Do, Check, Act) method in developing cyber security management practices. In order to put the measures into practice, the leadership of an organisation regard trust-enhancing measures related to cyber security as a strategic goal, maintain efficient processes and communicate their implementation with a policy that supports the strategy.

Keywords: Organization, Cyber security, Process, Trust, Architecture, PDCA

1. Introduction

The functioning of a modern society is based on the cooperation of several critical infrastructures, whose joint efficiency depends increasingly on a cyber secure organization. Crucial in the cyber environment are the usability, reliability and integrity of system data in the operating environment, whose cyber security risks are continuously augmented by threatening scenarios of the digital world. A modern society depends entirely on a cyber environment that provides dynamic services.

In Finland’s first Cyber Security Strategy, the cyber environment (domain) is defined as an electronic information (data) processing environment that consists of one or more information technology infrastructures. According to the Strategy, cyber security refers to a desired end state in which the cyber environment is reliable and in which its functioning is ensured. Critical infrastructure, furthermore, refers to the structures and functions that are indispensable for the vital functions of society. They include physical facilities and structures as well as electronic functions and services. (Secretariat of the Security Committee, 2013)

The global threats within the cyber environment have remained at a high level over the past few years, as stated in the annual international business world surveys by the World Economic Forum. They are seen to be among the major global threats based on the probability and impact of their realization. (World Economic Forum, 2019)

The role of organization ICT-systems in the cyber world as well as the factors that affect their cyber security, it is of primary importance to be aware of the most central features of the systems. Organizations can use different standards, frameworks and best practices when addressing cyber security. These governance documents are used typically to implement different kind of controls. In addition to that most of these documents also describe very specific capabilities, and an organization can use them to develop in securing their cyber domain. By enhancing capabilities, consisting of people, processes and technology, are meant to achieve outcomes or effects, that are applicable to the operational domain. (Jacobs, P. C., von Solms, S. H. and Grobler, M. M., 2016)

The ICT and industrial automation systems are part of the common cyber world, in which the primary risks are related to the loss of money, sensitive information and reputation as well as to business hindrance. Security solutions are hereby the key elements in risk management. The vulnerabilities behind the risks can be analysed as insufficient technology in relation to attack technology, insufficient staff competence or inappropriate working methods, deficiencies in the management of organizations, and lacks in operating processes or their technologies. The most common motives of attackers are related to the aim of causing destructive effects on processes, making inquiries about process vulnerabilities, and anarchism or egoism. These attacks can even be carried out by state-level actors, but perhaps most commonly by organized activists, hackers or individuals acting independently. (Lehto M., 2015)

The basic research question of this paper is "How can we gain cyber trust at the organizational level?"

2. An organization's cyber structure

Martin C. Libicki has created a structure for the cyber world, whose idea is based on the Open Systems Interconnection Reference Model (OSI). The OSI model groups communication protocols into seven layers. Each layer serves the layer above it and is served by the layer below it. The Libicki cyber world model has the following four layers: physical, syntactic, semantic and pragmatic. (Libicki, 2007)

Organization´s supply chains are complex systems of systems characterized by a conglomeration of interconnected networks and dependencies. The general networks and working processes involved in the operation of an organization can be illustrated with a logistics framework that comprises a supplier network, a production process, a client network, and information and material flows that connect them. According to EU commission "the ICT sector is vital for all segments of society". Information and communication technology (ICT) systems are part of critical organization's infrastructure and thus constitute a significant part of the operations that support an organization's core processes. Corporate-level ICT systems are related to administration and to the management of information and material flows in the network. The production level includes industrial automation systems (industrial control systems, ICS). (Edwards N, et al. 2016, EU Commission, 2009)

Cyber security professor in University of Jyväskylä, Martti Lehto, has updated the Libicki`s four layers cyber world model by adding the fifth layer in order to consider organisation networking needs. The structure is described in figure 1.

In the case of five-layers model structure, the physical layer contains the physical elements of the communications network, such as network devices, switches and routers as well as wired and wireless connections. The syntactic layer is formed of various system control and management programs and features which facilitate interaction between the devices connected to the network, such as network protocols, error correction, handshaking, etc. The semantic layer contains the information and datasets in the user's computer terminals as well as different user-administered functions, such as printer control. The service layer is the heart of the entire network. It contains such as administrative services, ICT-services, security services, IT based manufacturing services, supply and logistics services. The cognitive layer portrays the user's information-awareness environment: a world in which information is being interpreted and where one's contextual understanding of information is created.

Protecting the ICT-systems against threats implies measures taken based on risk assessment, and they ensure the availability of primarily digital information in the operating processes being examined. The measures are highly significant for the overall availability of the systems that support the organization´s business processes. Availability plays a key role in achieving business results and promoting the reliability of activities. Further central goals include the reliability and content integrity of information within the processes and used by the processes. Overall trust should be built from these starting points, based on the target organization's realistic idea of its own capabilities to reliably manage the challenges involved in operations within the cyber world. The following section addresses the significance of trust in the cyber environment for the operations of an organization. Moreover, trust-enhancing measures applicable to an organization will be mapped.

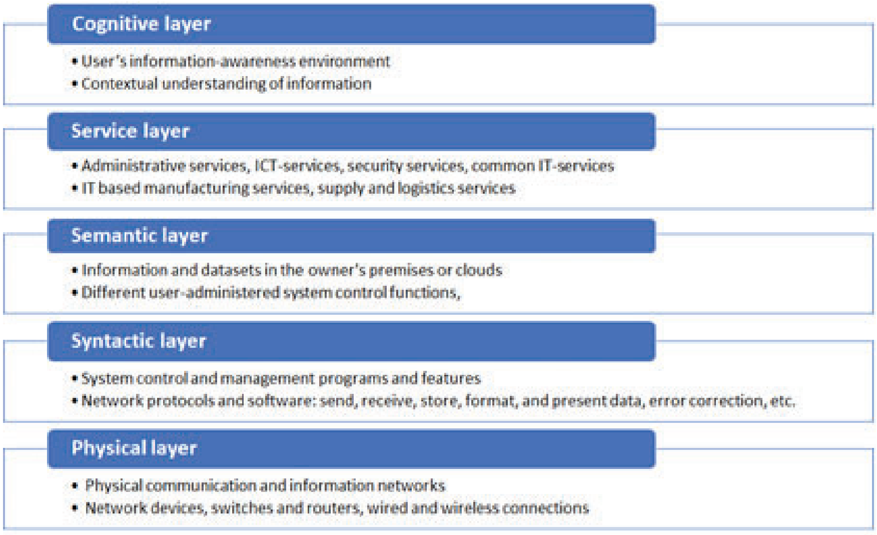

Cognitive layer
• User's information-awareness environment
• Contextual understanding of information
Service layer
• Administrative services, ICT-services, security services, common IT-services
• IT based manufacturing services, supply and logistics services
Semantic layer
• Information and datasets in the owner's premises or clouds
• Different user-administered system control functions,
Syntactic layer
• System control and management programs and features
• Network protocols and software: send, receive, store, format, and present data, error correction, etc.
Physical layer
• Physical communication and information networks
• Network devices, switches and routers, wired and wireless connections

Figure 1: The fife layer structure for the cyber world (modified from Lehto & Neittaanmäki, 2018)

3. Trust in an organization's cyber security

3.1 The significance of trust for cyber security

Trust in the operation of organizations and its continuous maintenance with effective measures are central factors affecting cyber security. Security is based on trust. Without trust there is no security, and vice versa. It is also good to be conscious of the fact that perfect safety is in general hardly achievable, and this also applies to the cyber world, which is a dynamic environment difficult to anticipate. Therefore, it is particularly important to understand the great significance of trust in the cyber world and its security. The role of measures enhancing trust is emphasized. When we build operations in the cyber world on a foundation that is as sustainable as possible, we can utilize the diverse opportunities it offers. (Limnell J., Majewski K. and Salminen M., 2014)

Finland's national cyber security strategy highlights the need to increase general cyber trust throughout the entire society. The strategy emphasizes that all actors, from individuals to enterprises and public administration, are responsible for their own preparedness in order to achieve cyber trust. (Secretariat of the Security Committee, 2013)

The ISO 9000 Standard states that an organization achieves success by acquiring and maintaining the trust of clients and other relevant interest groups. Understanding their present and future needs contributes to the organization's continuous success. The standard includes the central concepts of quality management and the principles for building trust. It can be applied by organizations that pursue ongoing success in their operation by utilizing a quality management system of their own. The quality of an organization's products and services is determined by how its clients experience that their needs and expectations are met. Clients also look for guarantees on the organization's ability to systematically produce products and services that correspond to their requirements. The ISO 9000 Standard comprises seven quality management principles, which constitute a commonly accepted basis for applying the standard series. The standard also specifies the benefits to an organization that has adopted the principles in its operation. The seven basic quality management principles are related to customer focus, leadership, the engagement of staff, a process approach, continuous improvement, evidence-based decision-making, and relationship management. (Finnish Standards Association SFS, 2016)

3.2 Cyber trust and process management

Establishing measures that increase cyber world security and trust in an organization is primarily the responsibility of corporate leadership. Integrating the necessary measures with the idea of ensured business activities increases their significance and benefits through better processes for the entire organization, interest groups and society. If security is not considered, risk analysis reveals potential damages as well as their costs and social consequences. The leadership's views and requirements brought out in the analysis play a central role in developing security planning for the operating process. The costs and other resources allocated to the activities are simultaneously specified. (Stouffer K., Falco J. and Scarfone K., 2011)

Process management theory has developed along with industrial production. The development of industrial mass production led to the use of variation theory while developing production process control: to perform control measures, uniform product quality was monitored with statistical methods. Statistical process analysis led to the observation that variation occurs everywhere in nature and in the processes and systems created by humans. After analysing distributions that involved variation, variation was classified into two types according to its causes: variation due to common causes (or the system itself) and variation due to special causes (i.e. named and assignable causes). Systemic variation has random causes and it is therefore often normally distributed, according to Gaussian distribution. Variation resulting from special causes does not follow any regularities. The common causes of variation are thus constantly present in the process. An individual cause produces only little deviation, but several causes together generate considerable variation. The causes of special variation, on the other hand, are not constantly present in the process. They come from outside of the process and usually generate more variation in the process than the common causes. In uncontrolled processes, deviation as a result of both types occurs simultaneously. (Lillrank P., 1998)

In principle, Lillrank's theory on the causes of process variation can also be generalized to the processes of an organization. The measures taken by organization leadership can be targeted at reducing variations resulting from both aforementioned types of causes. Proper planning and control of process performance reduce variation generated by random causes. At a general level, it is always recommended to aim at reducing this variation. If organization leadership, in particular, concentrates too much on process changes resulting from random causes, it can lead to overreactions in process control due to the measures chosen. At its worst, this can lead to loss of control in managing the overall process. The actions of organization leadership should indeed be targeted primarily at proactively preventing variation generated by special causes.

Almost without exception, serious cyber security disturbances occurring in the operating process cause harms. They do not represent normal process variation but are deviations resulting from special causes. They are not in the normal range of variation. Taking these special causes into account in planning and proactively implementing managerial activities reduces related risks and improves the overall reliability of the organization's operations.

3.3 Measures increasing cyber trust

The following measures related to cyber security management in an organization encompass of leadership, process management and seven principles quality management.

In order to comprehensively build organization cyber security, organization leadership must define and guide actions at the strategic, operational and technical-tactical levels. The strategic level provides answers to 'why' and 'what' questions. The operational and tactical levels answer the 'how' question. The approach guided by questions ensures that the right things are done and that they are done in line with the set goal. The technical-tactical level must implement the goal-oriented activities defined at the strategic level, not create it. The organization's organizational capability in implementing the cyber security measures required by the technical-tactical level ultimately determines how the organization manages potential disturbance situations. (Limnell J., Majewski K. and Salminen M., 2014)

Building organization cyber security management begins from the level of vision and strategy work. The visions created by organization leadership to enhance cyber trust are translated into strategic goals, operational-level actions, guidelines and a policy. The practical measures derived from the strategy are realized at the technical-tactical level. Organizational capability factors enable the success of the measures.

The strategic choices supporting the creation of visions are primarily related to organization social responsibility, organization reputation, and ensuring business and its economic efficiency. The leadership is expected to make concrete strategic choices as well as support and guide the execution of the chosen measures throughout the organization. It is also important that the leadership ensures sufficient resource allocation to the measures. The chosen measures should be comprehensively communicated to the organization's interest groups. (Stouffer K., Falco J. and Scarfone K., 2011, Finnish Standards Association SFS, 2016)

The measures at the operational level promote the strategic goals. Comprehensive measures that increase security and trust call for holistic cyber security management. It must be based on risk assessment and analyses of the measures based on the assessment. It is also important that the organization declares and communicates the policy with which the leadership commits to the measures required to develop cyber security management. The declaration of a policy that ensures cyber security and the development of related procedures must be integrated with the organization's general policies. The highest organizational level is responsible for creating a policy that defines acceptable risk levels and the measures used in the reduction of risks (Stouffer K., Falco J. and Scarfone K., 2011). The concrete measures at the operational level must be targeted at ensuring data security solutions and at creating business continuity and recovery plans (Finnish Standards Association SFS, 2012). The maintenance of situational awareness regarding the cyber environment of the organization's processes, furthermore, makes it possible to monitor the effects of the operational measures and, when needed, to react efficiently to events that constitute a threat within the organization's operating environment. The aim must be to continuously monitor the availability of processes and to support decision-making in disturbance situations that require analyses and decisions. (Faber, 2015)

The tactical organization level encompasses the systems and processes that comply with the he tactical organization level encompasses the systems and processes that comply with the structure of the cyber world. Consistent and predictable results are achieved more efficiently when operations are handled and managed as interrelated processes that function as a coherent system (Finnish Standards Association SFS, 2016). Cyber security threats set special requirements for these processes in addition to other operational requirements. At a general level, the performance of processes is determined according to their client-based demands. In the cyber environment, the target can be achieved by defining the processes to be protected, choosing process control mechanisms successfully, and by using expedient technological solutions and services to protect the processes (Stouffer K., Falco J. and Scarfone K., 2011). Successful operation also calls for the adoption of security-oriented values to guide the activities of staff (Lillrank P., 1998). The aforementioned solutions suitable for the cyber environment constitute an entity that can be called a technical-tactical level.

The continuous improvement of activities related to cyber security as well as the development of staff competence enhance the organization's capability to proactively prevent disturbances and tolerate potential changes in process operation caused by them. Taking the staff into account at all organizational levels, as well as focusing on competence and the possibilities it opens to fully influence in the organization, develops the overall operations of the organization (Finnish Standards Association SFS, 2016).

Staff competence is a crucial factor that determines the level of the entire organization's activities. The capacity of human resources can be increased by developing employees' knowledge and skills related to cyber security and thus developing the organization's capabilities. Challenges related to capabilities grow when an enterprise's cyber environment becomes more complex along with globalization and technological development. Investment in staff competence can transform the enterprise's capability into core competence, which can be used to pursue competitive advantage through trust. This will provide unique added value to both the organization and its customers. Valuable capabilities are helpful in an organization's threat and risk management and can consequently facilitate, in particular, the utilization of profitable opportunities. (Hitt, M. A., Ireland, D. R. and Hoskisson, R. E., 1997)

The Finnish Cyber-Trust research program was conducted 2015-2017 and University of Jyväskylä was one of the research partners in this program. We have determined the cyber trust for organization and described the relationship between cyber trust and comprehensive actions at organizations strategic, operational and technical-tactical levels at one of the research programs reports. The information from the report has been transferred to this scientific article in the way we have now shown. The cyber security and continuous development activities and staff competence support the measures taken at the strategic, operational and technical-tactical level. The summarized measures taken to increase an organization cyber trust is illustrated in Figure 2.

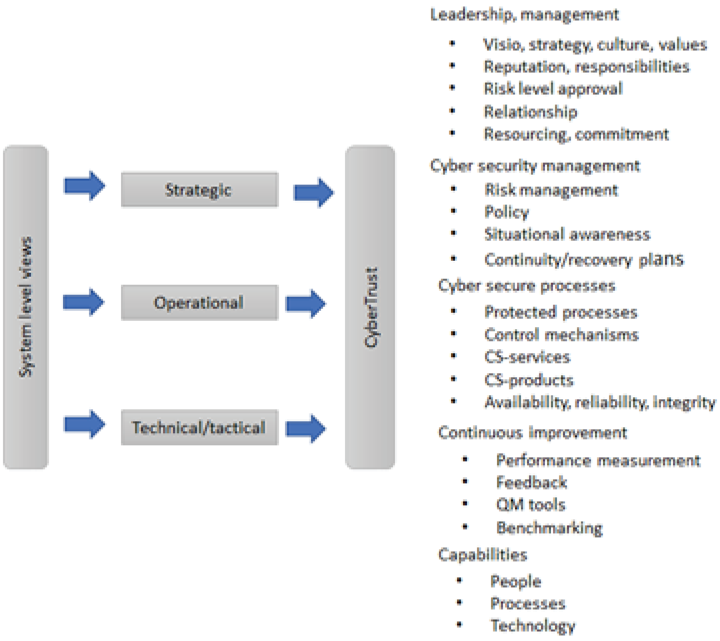


Figure 2: Measures increasing an organization cyber trust

4. Implementing the measures that enhance cyber trust

4.1 An integrated management system and its components

The latest standard of the ISO 9000 series (ISO/IEC 9004: 2018) emphasizes the importance of trust in an organization's ability to achieve continuous success and to recognize at all levels of management the factors influencing its operations in a constantly changing operating environment ((ISO/IEC 9004: 2018)

When management in an organization is performed systematically, we talk about the organization’s management system. A management system can comprise various control systems that comply with different standards, such as a quality management system, an information security management system and an environmental management system. In order to put into practice principles that comply with different standards, an organization may describe the required measures in its integrated management system (IMS). The IMS is a description of the procedures everyone should apply in the organization. With the help of guidelines and operations models jointly defined by the leadership and staff, the aim is to purposefully maintain a high level of activities and to develop the activities with an eye on set goals as well as the needs of clients and interest groups. The integrated management system compiles process descriptions, guidelines, recordings, indicators, tasks and feedback into a functional whole, which guides and supports the organization’s mission and vision as well as the actions taken to realize and assess them.

4.2 Decision-making levels and system view

Organizations operate in very complex, interrelated cyber environments, in which the new and long used information technical system entities (e.g. system of systems) are utilized. Organizations are depended on these systems and their apparatus in order to accomplish their missions. The management must recognize that clear, rational and risk-based decisions are necessary from the point of view of business continuity. The risk management at best combines the best collective risk assessments of the organization’s individuals and different groups related to the strategic planning, and also the operative and daily business management. The understanding and dealing of risks are an organization’s strategic capabilities and key tasks when organizing the operations. This requires for example the continuous recognition and understanding of the security risks on the different levels of the management. The security risks may be targeted not only at the organization’s own operation but also at individuals, other organizations and the whole society. (Joint Task Force Transformation Initiative, 2011)

Joint Task Force Transformation Initiative (2011) recommends implementing the organization’s cyber risk management as a comprehensive operation, in which the risks are dealt with from the strategic to tactical level. That way, the risk-based decision-making is integrated into all parts of an organization. In Joint Task Force Transformation Initiative's research, the follow-up operations of the risks are emphasized in every decision-making level. For example, in the tactical level, the follow-up operations may include constant threat evaluations about how the changes in an area can affect the strategic and operational levels. The operational level's follow-up operations, in turn, may contain for example the analysis of the new or present technologies in order to recognize the risks to the business continuity. The follow-up operations of the strategic level can often concentrate on the organization's information system entities, the standardization of the operation and for example on the continuous monitoring of the security operation. (Joint Task Force Transformation Initiative, 2011)

From the necessity of the organization's cyber security operations can be drawn as a necessity of comprehensive awareness in system level. The organization's and decision-maker's awareness can be seen as a system level awareness arrangement. Thus, an appropriate awareness supports the cyber risk management and more extensively the evaluation of the organization's whole cyber capability. We have integrated organization´s three decision-making levels to five-layers cyber structure in order to have comprehensive system view from organization cyber security environment. It is system thinking approach to the organization cyber security subsect and subject and the principle is described in figure 3.

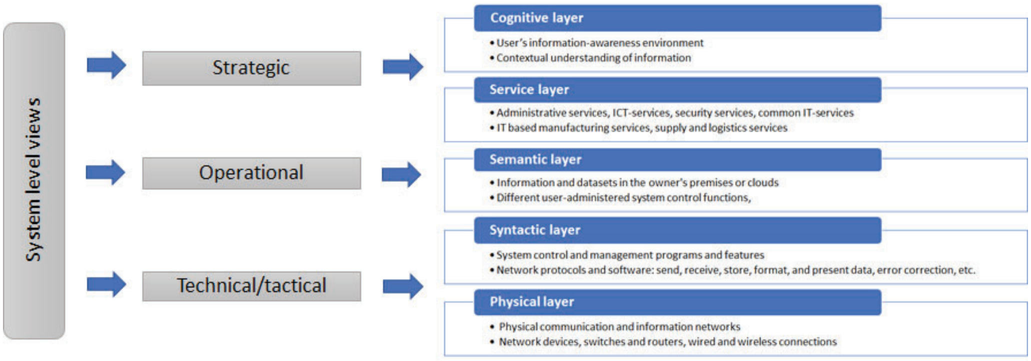

System level views
Strategic
Operational
Technical/tactical
Cognitive layer
• User's information-awareness environment
• Contextual understanding of information
Service layer
• Administrative services, ICT-services, security services, common IT-services
• IT based manufacturing services, supply and logistics services
Semantic layer
• Information and datasets in the owner's premises or clouds
• Different user-administered system control functions,
Syntactic layer
• System control and management programs and features
• Network protocols and software: send, receive, store, format, and present data, error correction, etc.
Physical layer
• Physical communication and information networks
• Network devices, switches and routers, wired and wireless connections

Figure 3: System level view to organization cyber security

4.3 Trust-enhancing measures based on trusted architecture

By adding three decision-making levels to five-layers cyber structure in order to have comprehensive system view from organization cyber security environment. The following standards were utilized to support the idea to create trust based cyber security architecture framework based on comprehensive system view from organization cyber world.

NIST 800-39 publication places information security into the broader organizational context of achieving mission/business success. The objective is to: (NIST 800-39, 2011)

- Ensure; senior leaders/executives recognize cyber security risks and managing such risks;
- Ensure; risk management process are being conducted across the three tiers of organization, mission/business processes, and information systems;
- Foster; cyber security risks in mission/business processes are designed within comprehensive enterprise architecture, and system development life cycle processes; and
- Help; people in system implementation or operation understand how cyber security risks in systems affect the mission/business success (organization-wide risk).

The ISO 9000 standard family of quality management systems helps organizations ensure meets of stakeholders needs related to products or services. The main goal is the customers satisfaction. The fundamentals of quality management systems, including the seven quality management principles (customer focus, leadership, the engagement of staff, a process approach, continuous improvement, evidence-based decision-making, and relationship management) are the basic principles of the standard family. (Finnish Standards Association SFS., 2016)

The ISO 27000 standard family provides recommendations for information security management system (integrated elements of an organization to establish policies and objectives and processes to achieve those objectives), risk treatments and controls. (ISO/IEC 27000: 2018)

Integration of the measures increasing an organization cyber trust (Fig 2.) and the system thinking approach to organization fife layer cyber structure (Fig. 4) makes possible to have trust based cyber security architecture framework. The content of the aspects of it is derived from the organization-wide risk management standard, NIST 800-39 (NIST, 2011) and perspectives from ISO 9000 standard family (seven quality management principles) and ISO27000 standard family.

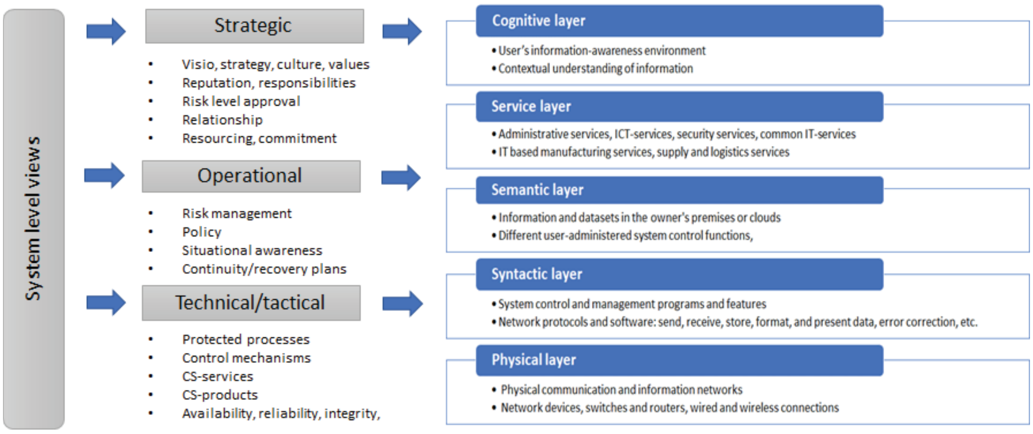

System level views
Strategic
Visio, strategy, culture, values
Reputation, responsibilities
Risk level approval
Relationship
Resourcing, commitment
Operational
Risk management
Policy
Situational awareness
Continuity/recovery plans
Technical/tactical
Protected processes
Control mechanisms
CS-services
CS-products
Availability, reliability, integrity,
Cognitive layer
User's information-awareness environment
Contextual understanding of information
Service layer
Administrative services, ICT-services, security services, common IT-services
IT based manufacturing services, supply and logistics services
Semantic layer
Information and datasets in the owner's premises or clouds
Different user-administered system control functions,
Syntactic layer
System control and management programs and features
Network protocols and software: send, receive, store, format, and present data, error correction, etc.
Physical layer
Physical communication and information networks
Network devices, switches and routers, wired and wireless connections

Figure 4: Trust based cyber security architecture framework.

Business Dictionary.com defines a capability in general as the "measure of the ability of an entity (department, organization, person, system) to achieve its objectives, especially in relation to its overall mission", and in quality perspective as the "total range of inherent variations in a stable process". (BusinessDictionary.com 2020). Dickenson and Mavris defines a capability as "the ability to achieve a desired effect under specified standards and conditions through combinations of ways and means to perform a set of tasks" and also "the ability to execute a specified course of action". (Dickerson C. and Mavris D. N., 2010). Thus, the capability of organization can be seen also as an ability to learn from its experiences and use relevant information to improve cyber security processes. Organization capabilities support actions of the architecture framework.

The process approach promoted by ISO 9001 systematically identifies processes that are part of organization quality system. Related to the quality management system, the PDCA cycle is a dynamic cycle that could be implemented in each process throughout the organization. It combines planning, implementing, controlling and continual improvement. That way organization would achieve continual improvement once it will implement the PDCA cycle. (9001 quality)

4.4 The PDCA method as a tool for developing activities

An organization's cyber security architecture framework could demonstrate how cyber thrust measures can be addressed, committed and implemented in all decision-making levels, strategic, operational and technical/tactical. In that means continuous improvement of processes can based on activities, whereby risk analysis has been considered. In order to create the measures, the organization must have a systematic approach to developing its operations.

The ISO9000 Standard recommends the PDCA (Plan, Do, Check, Act) method for a systematic development of an organization's activities. The method is based on a cycle of four development phases. The first phase (Plan) set the objectives of the system and processes to deliver results ("What to do" and "how to do it"). In the second phase realization (Do), implement and control what was planned. The third phase (Check), monitor and measure processes and results against policies, objectives and requirements and report results. At the last (Act), take actions to improve the performance of processesphase. After the cycle has been implemented once, one will return to the first phase and start a new cycle with improvement actions based on a new situation analysis. Development can thus proceed as an endless process, in which a new level of activities is achieved after each cycle. The method is based on the idea of continuous improvement, with risk-based thinking at each stage of activities. (ISO/IEC 9001: 2015)

The measures during one round of the cycle usually require a lot of planning, so sufficient time should be reserved for it. It is important to select the measures in relation to the resources needed for their implementation. The maturity level of the organization's development activities affects the evaluation of the implemented measures. When developing cyber security, at an initial stage the aim can be to recognize the need for cyber security management and to define cyber security risks for business. Hereby, the PDCA cycle may comprise the administrative actions most necessary according to risk assessment, such as a coherent information security policy in production, practical guidelines for maintaining information security in production, and potential preliminary system-specific cyber security checks. The targets for development must later be chosen according to risk prioritization.

5. Conclusion

Ensuring the availability and reliability of processes and data integrity is vital for the efficient functioning of trust based cyber security. Therefore, the measures taken in organizations in order to manage and control its processes and implement continuous improvement actions are an essential component of the trust of business processes.

The major cyber environment risks within the processes of an organization require that trust is enhanced and maintained at all levels of business activity. For the basic research question "How can we gain cyber trust at the organizational level?", this paper includes system view and standards-based way to form cyber trust architecture framework for organization use. Firstly, the cyber trust activities for organization have been defined in this article. After that the secure architecture. Comprehensive measures to increase cyber trust and trusted architecture framework together with the development of organization capabilities related to cyber activity, also improve an organization's continuity and competitive edge. The integrated management system (IMS) of organization supports the way for assess the actions to be taken.

Serious cyber security disturbances occurring in the organization operating process cause in many cases harms. They are resulting from special causes in the process. So, they do not represent normal process variation. Implementing the cyber trust architecture framework, these special causes would be taken into account proactively and measures reduce process related risks and improves the overall reliability of the organization's operations.

References

Faber S., 2015. Flow Analytics for Cyber Situational Awareness. SEI Blog, 2015. https://insights.sei.cmu.edu/sei_blog/2015/12/flow-analytics-for-cyber-situational-awareness.html

Dickerson C. and Mavris D. N., 2010. Architecture and Principles of Systems Engineering, CRC Press.

Edwards N., Kao G., Hamlet J., Bailon J. and Liptak S., 2016. Supply Chain Decision Analytics: Application and Case Study for Critical Infrastructure Security. Proceedings of The11th International Conference on Cyber Warfare and Security ICCWS 2016 s. 99-106.

EU Commission, 2009. Critical information infrastructure protection (2009), COM (2009) 149 final, Commission of the European Communities, Brussels, 30.3.2009

Finnish Standards Association SFS., 2012. SFS-käsikirja 327. Informaatioteknologia. Turvallisuus. Tietoturvallisuuden hallintajärjestelmät. SFS ry, Helsinki, 2012.

Finnish Standards Association SFS., 2016. Johdanto laadunhallinnan ISO 9000 -standardeihin. Available on 27 January 2020: slideplayer.fi/slide/11133323/

Hitt, M. A., Ireland, D. R. and Hoskisson, R. E., 1997. Strategic Management, 2th edison, St Paul, West Publishing Company, 438 s.

Jacobs, P. C., von Solms, S. H. and Grobler, M. M., 2016. Towards a framework for the development of business cybersecurity capabilities. International Conference on Business and Cyber Security (ICBCS), London, UK. The Business and Management Review, Volume 7 Number 4, 51–61.

Joint Task Force Transformation Initiative, 2011. NIST Special Publication 800-39: Managing Information Security Risk - Organization, Mission, and Information System View, Gaithersburg: National Institute of Standards and Technology.

ISO/IEC 27000: 2018. International Organization for Standardization. Information technology. Security techniques. Information security management systems. Overview and vocabulary. Retrieved on 27 January 2020 from https://www.iso.org/standard/73906.html

ISO/IEC 9001: 2015, International Organization for Standardization. THE PROCESS APPROACH IN ISO 9001:2015. Retrieved on 27 January 2020 from https://www.iso.org/files/live/sites/isoorg/files/archive/pdf/en/iso9001-2015-process-appr.pdf

ISO/IEC 9004: 2018. International Organization for Standardization. Quality management. Quality of an organization. Guidance to achieve sustained success Retrieved on 27 January 2020 from https://www.iso.org/standard/70397.htmlhttps://www.iso.org/standard/70397.html

Lehto M., 2015. Cyber Security: Analytics, Technology and Automation. Springer.

Lehto, M. & Neittaanmäki, P., 2018. The modern strategies in the cyber warfare. Cyber Security: Cyber power and technology. Berlin: Springer.

Libicki, M. C., 2007. Conquest in Cyberspace – National Security and Information Warfare, Cambridge University Press, New York 2007.

Lillrank P., 1998. Laatuajattelu. Laadun filosofia, tekniikka ja johtaminen tietoyhteiskunnassa. Otavan Kirjapaino Oy, Keuruu, 1998.

Limnell J., Majewski K., and Salminen M., 2014. Kyberturvallisuus, Docendo Oy, Jyväskylä, 2014.

Secretariat of the Security Committee. Finland’s Cyber Security Strategy. [online document], 2013. http://www.defmin.fi/files/2378/Finland_s_Cyber_Security_Strategy.pdf

Stouffer K., Falco J., Scarfone K., 2011. NIST Special Publication 800-82. Guide to Industrial Control Systems (ICS) Security. Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce. [online document]. http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf

World Economic Forum. The Global Risks Report 2019. 14th Edition. Retrieved on 27 January 2020 from http://reports.weforum.org/global-risks-2015/part-1-global-risks-2015/technological-risks-back-to-the-future/#frame/20ad6

9001 quality. 2020. The Plan Do Check Act (PDCA) cycle. Retrieved on 27 January 2020 from http://9001quality.com/plan-do-check-act-pcda-iso-9001/

Cybersecurity Awareness in an Industrial Control Systems Company

Stefan Prins, Annlizé Marnewick and Suné von Solms
University of Johannesburg, South Africa
svonsolms@uj.ac.za

DOI: 10.34190/EWS.20.010

Abstract: This paper investigates the cybersecurity awareness levels of employees at an industrial control systems organization and measures their knowledge on the potential impact of cyber-related attacks on their systems through a case study. Attacks on industrial control systems as well as the information technology infrastructure which it relies on, are becoming a growing problem for governments and organizations. Cybersecurity policies of organizations are critical to ensure that industrial control systems environments are adequately protected. It is equally important for the organizations to ensure that their employees are aware of the cybersecurity policies and why they must be implemented. In many cases, however, organizations are faced with employees who are not aware of the potential cyber-related security threats posed to their industrial control systems, nor the impact these attacks might have. Results show that although employees understand the severity of cyber vulnerabilities their awareness is low.

Keywords: Awareness, Cybersecurity, Industrial control systems, Vulnerabilities

1. Introduction

Individuals as well as organizations rely on information technology (IT) to assist them in their daily activities using equipment such as computers. Connected systems carry the risk of various cyber-related attacks, ranging from theft of data, personal information exposure, financial loss, sabotage to name a few. In the Industrial Control Systems (ICS) environment, the control systems are becoming more integrated with IT technology, which means that ICS systems are also more susceptible to cyber threats (Byres & Lowe, 2018). Organizations with highly connected ICS are faced with risks as attacks on these systems can have severe effects on the operations of the company. The data shows that cyber-related attacks on ICS systems are on the rise (Byres & Lowe, 2018). In 2017, the WannaCry worm infected Windows based computers and caused worldwide disruption (Symantec Security Response, 2017). The Stuxnet worm was designed to attack specific ICS systems, such as supervisory control and data acquisition (SCADA) as well as programmable logic controller (PLC) systems, and is considered as a weapon of cyberwarfare by some specialists (Singer, 2015). In 2015, the Industrial Control Systems Cyber Emergency Response Team (ICS-CERT) discovered the BlackEnergy malware on computers of a Ukrainian power utility after the company had unscheduled power outages impacting many of their customers. The findings indicated that the intruder used remote access to infiltrate the network (ICS-CERT, 2018). The risk and vulnerability for ICS networks as well as the challenge to address these were discussed by the authors (Pretorius & van Niekerk, 2016). The head of global research and analysis at Kaspersky in the Middle East, Turkey and Africa stated that one of the biggest trends in Africa in 2018 was increased cyber-attacks on ICS and critical infrastructure systems (Smith, 2019).

Organizations must reassess their risks and consider strategies such as protecting the systems with multiple layers of security (Thomson, & von Solms, 2006). The practice of cybersecurity is there to protect the IT systems of an organization by preventing data breaches, forming part of business operations (Thomson & von Solms, 2006). Businesses in the ICS environment can no longer consider cybersecurity practice as a low priority and must change their position on how to deal with it in better ways to protect a critical system (Knowles et al, 2015). Governments from developed countries such as the United States (US) and United Kingdom (UK), developed cybersecurity policies to make their citizens more knowledgeable about cybersecurity. In South Africa, however, many companies lack cybersecurity policies and awareness relating to Industrial Control Systems security, amongst others. According to a study by Nexia International, 39% of South African organizations do not provide any training in cybersecurity, and only 30% train their employees occasionally (Nexia International, 2017). The South African government developed a cybersecurity policy, but the policy is unclear how the nation will be made aware of cybersecurity (Kortjan & von Solms, 2013).

The threats which exist in the domain of cyberspace is an increasing problem for organizations in the ICS environment and thus creates a situation where organizations must implement strategies to counteract the threats like educating the employees through awareness programs. This research aims to determine the level of cybersecurity awareness in the organization in the ICS environment and to determine the impact of the potential risk for the organization. This paper is structured as follows: Section 1 provided the introduction where section

2 provides an overview on the related work. section 3 describes the research methodology followed in this research. Sections 4 and 5 provides the Data Collection and Results sections, respectively. Section 6 includes a discussion and section 7 concludes this paper.

2. Related Work

Industrial Control Systems is a term used to refer to the overall use of control systems (Stouffer et al, 2015). It is common to see terms such as ICS or SCADA interpreted as the same thing, but in reality, the terms refer to different types of control systems (Krotofil & Gollmann, 2013), where ICS is viewed as covering all types of systems. ICS used to be a complex system designed for dependability, durability, and safe to use by operators (Krotofil & Gollmann, 2013). These type of design characteristics allowed the control system to function separately from any other systems with the minimum risk. Through the evolution of technology in the ICS environment, the IT infrastructure started to become more integrated with the control systems and replacing or enhancing some of the other physical systems (Stouffer et al, 2015). According to (Krotofil & Gollmann, 2013), the evolution of technology allows for better integration with other systems and integration with cyber systems. The evolution of technology improves the efficiency of the systems, but it also causes more risk which is difficult to foresee (Krotofil & Gollmann, 2013).

ICS systems require companies with the right set of knowledge and skills to develop a system according to the requirements of the client and who can support the ICS system through its product lifecycle. Cyber attackers have learned of the value of targeting ICS as it can lead to severe damage to daily business operations, including operational shutdowns, equipment damage, financial loss, intellectual property loss, as well as health and safety risks (TrendMicro, 2016). In order to secure ICS infrastructure, personnel, policy makers and engineering experts must be involved to determine and analyze ICS infrastructure risks. The cybersecurity maturity level of an ICS organization is dependent on how well it understands its ICS and network environments. In the context of more connected systems and the evolution of ICS to be fully connected, developers, users and support personnel must therefore have the knowledge and understanding of the threats, vulnerabilities and risks that come with the technology and must grasp the idea of how to implement cybersecurity.

Commenting on the increase in cyber-related attacks on ICS in Africa, the head of global research and analysis at Kaspersky in the Middle East, Turkey and Africa stated that “(i)n many countries – especially in Africa – there is a lack of awareness of this threat to critical infrastructure. Many of these facilities are not aware that they are exposed to cyber risk and can be controlled if hacked” (Smith, 2019). According to Kortjan & von Solms (2013), governments and organizations must use policies to ensure that the desired goals, such as employee awareness, are met. Management of an organization should ensure the workers follow the company policies. Company policies must set out what the cybersecurity procedures are and must then be used to control the worker's behavior towards cybersecurity. A cybersecurity policy of an organization must enable the workers to be more aware when using company information or assets and help to protect the company from potential damages (Vroom & von Solms, 2002). Companies must run their awareness programs on a regular basis to ensure the culture remains intact (von Solms & von Solms, 2004). Awareness programs help the workers understand the technical insights and how procedural processes can be used to prevent cybersecurity attacks in the ICS environment. Awareness programs in an organization are there to continuously drive the importance of cybersecurity and can't be considered as a quick fix. In addition, ICS cybersecurity awareness programs must also be adapted from IT because of the different characteristics (Stouffer et al, 2015) and must form part of the policies of an organization or government (Byres & Cusimano, 2012). To increase awareness, programs to inform workers about the threats, vulnerabilities and risks in the ICS environment should continuously be integrated with organizational communication drives.

Similar to awareness programs, an organization must develop a plan to improve a workers’ skills in cybersecurity through training programs. The training programs must build and enhance the skills of the workers. In addition, the ICS and IT communities in an organization must form a good working relationship. Management in organizations must ensure the two different groups can work together and ensure there is knowledge transfer between the groups. Sharing knowledge between the two groups will allow better solutions, better use of workers, and better use of monetary resources (National Cyber Security Centre, 2015). Piggin (2013) identified 14 security themes that leads to specific challenges when ICS and SCADA systems are integrated into organisations business environments (Piggin, 2013). Two of the themes relates to the awareness as displayed in Table 1.

Table 1: Cyber security themes

Theme	Information technology	Industrial control system	Challenge
IT	Good	Generally a poor understanding of cyber security	Control engineers generally lack cyber security education or training
ICS	A poor understanding of control systems operating environments	Recently events have started to raise more awareness but lack of implementation.	Knowledge often limits required understanding to implement security. Fragmented team leads to potential security weaknesses.

Although security priorities within an ICS landscape is different from traditional IT landscapes, these two are integrated and cannot be addressed separate from each other in ICS implementations. From Table 1 it can be seen that a low level of cybersecurity awareness in the ICS environment can lead to cyber-related challenges. The requirements of a higher level of cybersecurity awareness by control systems engineers is therefore highlighted in (Piggin, 2013). Cybersecurity awareness in an organization is critical, as human error due to a lack of cybersecurity knowledge and awareness is one of the leading cause of cyber-incidents (Kaspersky, 2019). A framework for improving awareness and skills of employees' in organizations by (Torten, Reaiche & Boyle, 2018) suggested three parallel approaches:

1. Increase ongoing awareness by implementing awareness programs.
2. Establish training frameworks and focus on countermeasure training.
3. Continuously evaluate program effectiveness.

To determine the potential risks faced by an ICS organization, this paper investigates the cybersecurity awareness of employees of an ICS organization in South Africa. The methodology followed in this investigation is detailed in the next section.

3. Methodology

The aim of this research is to determine the level of cybersecurity awareness and the level of impact thereof for the organization. A case study was used as it is ideal to gain an in-depth understanding of a problem (Zikmund et al, 2010). The organization researched in the case study is a small South African company which specializes in the ICS environment by providing solutions, products, and services for the industrial market. The employees of the organization are typically involved in designing, implementation, commissioning, and maintenance of control systems for industrial customers. The employees' level of cybersecurity awareness and their perceived level of impact regarding risks were the unit of analysis in this research. The case study data collection made use of a questionnaire and document analysis to get the necessary information so that the researchers can determine the current condition within the organization. The study groups in the organization included the service department, project department, and internal sales department. The occupational respondents included the engineering manager, systems engineers, service engineers, and internal sales. As a small company, the number of professionals in these positions totals 10. The questionnaire was electronically distributed to all the identified respondents (total of 10) as a PDF form.

During the data collection process, the purpose was to collect the following data: (i) the participant's knowledge of cybersecurity awareness and (ii) the participant's perceived level of impact regarding risks. To assess these levels, a Likert scale was used as the measuring instrument to determine each respondent's knowledge of cybersecurity awareness and perceived level of impact. Table 1 illustrates the two Likert scales used for each statement testing knowledge of cybersecurity awareness and perceived level of impact.

Table 1: Likert scales for questionnaire

	1	2	3	4	5
Knowledge of cybersecurity awareness	Not aware	Somewhat aware	Moderately aware	Extremely aware	
Perceived level of impact	No impact	Minor impact	Moderate impact	Significant impact	Severe impact

Multiple data sources were used to triangulate the results of this study. The following data were utilized:

1. Benchmark data was collected which categorized vulnerabilities into different groups to present the different vulnerabilities of the ICS environment from the literature (Stamp, Young & Mae Depoy, 2003).
2. Document analysis related to the organization in the ICS environment is measured against the benchmark data collected in step 1. Organizational documents were divided according to the categories identified in the benchmark data.
3. Data collection though a questionnaire, collecting data on the level of awareness and the level of perceived impact from the participants. The structure of the questionnaire was derived from the benchmark data collected in step 1.

The common vulnerabilities as identified from the Guide to Industrial Control Systems (ICS) Security by the National Institute of Standards and Technology (NIST) was categorized into these different groups and was used as the benchmark data, shown in Table 2 (Stouffer et al, 2015). This list of vulnerabilities was the basis used in the questionnaire design as well as the document analysis checklist.

It can be seen from Table 2 that 41 vulnerabilities falling within 5 categories were identified. These vulnerability statements were utilized as the baseline data for the data collection discussed in Section 4. In order to obtain information from the organisation's documentation, every identified vulnerably listed in the table above was compared it to the organisation's documents. When a document was found where the vulnerability was addressed, the document was recorded.

Table 2: Vulnerability statements per ICS dimension

Data	
Data 1	Data is not correctly classified according to the sensitivity categories of industrial control systems.
Data 2	Sensitivity categories are not determined according to the level of impact.
Security administration	
Security admin 2	No adequate security policies for ICS such as config management, access control, and authentication.
Security admin 2	There are no official awareness and training programs about cybersecurity in industrial control systems.
Security admin 3	There is a lack of or no implementation guides on industrial control systems equipment.
Security admin 4	Organizational mechanisms are absent for security policy enforcement.
Security admin 5	The effectiveness of the security controls is not adequately reviewed.
Security admin 6	Inadequate security plans for ICSs, contingency plan, incident detection plan and response plan, not used.
Networks	
Networks 1	Data flow control is not utilised between systems such as the ICS and business networks.
Networks 2	Firewalls are not used between networks or not configured correctly.
Networks 3	There exist no adequate firewall and router logs for the industrial control system.
Networks 4	Standard, well-documented communication protocols are used in plain text.
Networks 5	Authentication of users, data, or devices of an industrial control system is substandard or non-existent.
Networks 6	Unsecured industrial control system protocols with no integrity checking for communications are utilised.
Networks 7	Inadequate authentication and data protection between wireless connections and access points.
Architecture	
Architecture 1	The addition of security controls into the architectural designs of ICS are inadequate.
Architecture 2	Insecure modifications are allowed to the architecture of industrial control systems.
Architecture 3	Non-control traffic allowed on the control network of industrial control systems.
Architecture 4	Control network services such as a domain controller are not controlled within the control network of ICS.
Architecture 5	Redundant components used for critical functions are absent from industrial control systems.
Platforms	
Platforms 1	Modifications of hardware, software, and firmware of industrial control system are not properly managed.
Platforms 2	There exist delays in OS and vendor software patches for the industrial control systems.
Platforms 3	Security patches are not maintained for industrial control systems.
Platforms 4	Security changes are not tested adequately before implementation.
Platforms 5	Poor remote access controls to access the industrial control systems are utilised
Platforms 6	Default or poor configurations on the industrial control platforms are implemented.
Platforms 7	Critical configurations of industrial control systems are not stored or backed up.
Platforms 8	There exist no data encryption on portable devices for the added protection.
Platforms 9	Password implementation policy is not implemented according to the organizational policies.
Platforms 10	Strong access controls are not applied across industrial control systems.
Platforms 11	Improper sharing methods are used to replicate data to other systems.
Platforms 12	Anti-malware software is not installed or updated on platforms.
Platforms 13	Anti-malware is loaded on platforms without adequate testing.

Platforms	
Platforms 14	Software on industrial control platforms are vulnerable to denial of service attacks.
Platforms 15	Intrusion detection or prevention software on industrial control platforms are not installed.
Platforms 16	Logs from industrial control systems are not collected or examined.
Platforms 17	Unauthorised access to the physical equipment is allowed.
Platforms 18	Physical ports on industrial control platforms are not secured.
Platforms 19	Data on industrial control platforms are not adequately validated.
Platforms 20	Security capabilities installed on industrial control platforms are not activated.
Platforms 21	There exists inadequate user authentication, access privileges, and access control in the software.

4. Data Collection

The organization in the ICS environment are measured against the benchmark data collected in step 1. The document analysis as well as the questionnaire design are discussed below.

4.1 Document Analysis

As part of the case study, documents were reviewed to determine how the organizational policies, guidelines and procedures addresses the ICS vulnerabilities. The documents investigated were the documents used by the organization to reduce the vulnerabilities and increase the security of the ICS environment. The organizational documents, which includes policy documents, guidelines and specification documents, are mainly used by the employees in ensuring the ICS infrastructure is protected and explain the processes to protect the products from cyber threats. The documents surveyed were the following:

1. Functional design specifications
2. Security countermeasures
3. Security guide
4. Security Requirements

Byres and Cusimano (2012) recommends a risk assessment to identify vulnerabilities in the ICS environment. Following this guideline, the document review was used to determine whether the organization is addressing these vulnerabilities. The organization documents were divided according to the categories identified in the benchmark data in Table 2. Every vulnerability identified in the benchmark data were compared it to the organization's documents.

4.2 Questionnaire

The respondents identified in the case study were asked to indicate their role in the organization. The aim of the question was to determine the demographics of the roles in the organization. The collected data were summarized in Table 3 to indicate the roles in the organization.

Table 3: Respondents' role in ICS organization

Respondent role	**Number of respondents**
Engineering Manager	1
Service Engineer	5
Systems Engineer	2
Project Manager	1
Internal Sales	1

The benchmark data of ICS vulnerabilities from Table 2 was used to develop the list of statements for measuring the participants. The validity of the questionnaire was obtained by building research questions, case study proposition and principles on the benchmark data shown in Table 2. Reliability was achieved using the stated data collection procedures and the case study database. The questionnaire was divided into two main sections, detailed below.

Cybersecurity awareness: The questionnaire's first objective was to determine the respondent's level of awareness. The questionnaire was designed to measure awareness of the vulnerabilities using a 4-point Likert scale (not aware, somewhat aware, moderately aware and extremely aware). The respondents indicated their awareness of each vulnerability statement identified in the benchmark data. To measure the cybersecurity awareness, the question was stated as follows "How aware do you rate yourself for each statement".

Cybersecurity impact: The second objective of the questionnaire was to determine the level of impact of each vulnerability statement. The questionnaire was designed to measure the impact of vulnerabilities using a 5-point Likert scale (no impact, minor-impact, moderate impact, significant impact and severe impact). The respondents indicated the level of impact for each vulnerability statement. To measure the impact, the question was as follows "What level of impact do you rate each vulnerability statement".

5. Results

The results obtained from the document analysis and the questionnaires are discussed in the sections below. The calculated weighted average for each vulnerability element in the benchmark data was calculated. A weighted average was also determined per category.

5.1 Document Analysis

The document analysis shows that there are five areas which are not covered in the ICS organization's policies and governance documents. The vulnerability descriptions which are not addressed in the document are shown in Table 5.

Table 5: Unaddressed elements in document analysis

Element	Description	Document analysis
Data 1	Data is not correctly classified according to the sensitivity categories of ICS.	No
Data 2	Sensitivity categories are not determined according to the level of impact.	No
Security admin 4	Organizational mechanisms are absent for security policy enforcement.	No
Architecture 4	Control network services such as a domain controller are not controlled within the control network of ICS.	No
Platforms 2	There exist delays in OS and vendor software patches for the ICS.	No

Of the 41 vulnerabilities analyzed, only five were not documented in the company documents. The organization therefore understand the vulnerabilities. It can be seen that the vulnerabilities not addressed in the documents relates to all the vulnerability categories determined in the benchmark data, which includes Data, Security Administration, Architecture and Platforms.

5.2 Cybersecurity Awareness

The data collected in the first section of the questionnaire was analyzed to determine the weighted totals for each vulnerability statement. The weight mean was calculated in order to rank the responses (Robson & McCartan, 2016) . The vulnerability statements were ranked based on a weighted mean to determine the lowest to highest level of awareness based on weighted averages determined. If the weighted average was found to be lower or equal to 2.7, it was considered as low awareness. The 14 elements rated for low awareness are displayed in Table 6. The weighted average column of each vulnerability statement is displayed in a greyscale to indicate the level of awareness amongst respondents.

Table 6: Level of awareness (lowest rankings)

Element	Description	Awareness Level
Network 7	There exist inadequate authentication and data protection between wireless connections and access points.	2.3
Network 4	Standard, well-documented communication protocols are used in plain text.	2.4
Platforms 19	Data on industrial control platforms are not adequately validated.	2.5
Security admin 4	Organizational mechanisms are absent for security policy enforcement.	2.6
Security admin 6	Inadequate security plans for industrial control systems such as a contingency plan, incident detection plan and response plan, are not used.	2.6
Network 3	There exist no adequate firewall and router logs for the industrial control system.	2.6
Architecture 4	Control network services such as a domain controller are not controlled within the control network of industrial control systems.	2.6
Platforms 6	Default or poor configs on the industrial control platforms are implemented.	2.6
Security admin 2	No official awareness and training programs about cybersecurity in ICS.	2.7
Security admin 3	There is a lack of or no implementation guides on ICS equipment.	2.7
Security admin 5	The effectiveness of the security controls is not adequately reviewed.	2.7
Architecture 5	Redundant components used for critical functions are absent from ICS.	2.7
Platforms 4	Security changes are not tested adequately before implementation.	2.7

Element	Description	Awareness Level
Platforms 20	Security capabilities installed on industrial control platforms are not activated.	2.7

It can be seen from the results in Table 6 that the lowest awareness items are distributed across Network, Platforms, Security administration and Architecture. If the weighted average was found to be higher or equal to 3.1, it was considered as high awareness. Twelve of the 41 vulnerabilities were rated with awareness above 3.1 out of a scale of 4 and are shown in Table 7.

Table 7: Level of awareness (highest ranking)

Element	Description	Awareness Level
Network 1	Data flow control is not utilised between systems such as the ICS and business networks.	3.5
Network 2	Firewalls are not used between networks or not configured correctly.	3.4
Platforms 1	Modifications of hardware, software & firmware of ICS not properly managed.	3.4
Platforms 17	Unauthorised access to the physical equipment is allowed.	3.4
Platforms 2	There exist delays in OS and vendor software patches for the ICS.	3.3
Platforms 14	Software on IC platforms are vulnerable to denial of service attacks.	3.3
Platforms 3	Security patches are not maintained for industrial control systems.	3.2
Platforms 15	Intrusion detection or prevention software on IC platforms are not installed.	3.2
Network 5	Authentication of users, data, or devices of ICS is substandard or non-existent.	3.1
Architecture 2	Insecure modifications are allowed to the architecture of ICS.	3.1
Platforms 10	Strong access controls are not applied across industrial control systems.	3.1
Platforms 13	Anti-malware is loaded on platforms without adequate testing.	3.1

The other vulnerabilities were all rated between 2.8 and 3. In total 29 vulnerabilities rated below 3.1, which translates to 71% of the vulnerabilities that have a low awareness. Six of the 8 vulnerability items with low awareness in Table 6 are mentioned in the company's documentation. Therefor the vulnerabilities are known in the organisation. However, the results show that the staff's awareness levels are low. As shown in Table 5, no documentation is available on Security administration 4 & Architecture 4.

5.3 Cybersecurity Impact

The data collected in the questionnaire related to the respondents' views on the possible impact for each vulnerability statement were analyzed to determine the weight mean per vulnerability. The vulnerability statements were ranked based on weighted mean to determine the highest level of impact based on the results of the respondents. If the weighted average was found to be higher or equal to 4.2, it was considered as high impact. The 15 vulnerability elements with the highest rated impact is displayed in Table 8. The weighted average column of each vulnerability statement is displayed in a greyscale to indicate the level of impact rated amongst respondents.

Table 8: Level of impact

Element	Description	Impact Level
Platforms 17	Unauthorized access to the physical equipment is allowed.	4.7
Security admin 1	There does not exist adequate security policies for industrial control systems such as configuration management, access control, and authentication.	4.6
Platforms 7	Critical configurations of industrial control systems are not stored or backed up.	4.6
Platforms 12	Anti-malware software is not installed or updated on platforms.	4.6
Security admin 3	There is a lack of or no implementation guides on ICS equipment.	4.5
Architecture 2	Insecure modifications are allowed to the architecture of ICS.	4.5
Platforms 15	Intrusion detection or prevention software on IC platforms are not installed.	4.5
Network 6	Unsecured ICS protocols with no integrity checking for communications are used.	4.4
Platforms 5	Poor remote access controls to access the industrial control systems are utilised	4.3
Platforms 10	Strong access controls are not applied across industrial control systems.	4.3
Security admin 2	There are no official awareness and training programs about cybersecurity in ICS.	4.2
Network 5	Authentication of users, data, or devices of an ICS is substandard or non-existent.	4.2
Architecture 5	Redundant components used for critical functions are absent from ICS.	4.2
Platforms 1	Modifications of hardware, software, and firmware of ICS not properly managed.	4.2
Platforms 21	Inadequate user authentication, access privileges & access control in software.	4.2

It can be seen from the results in Table 8 that the vulnerability elements rated as having the highest impact is all covered in the policy documents. Elements from the Platform and Security Architecture ICS dimensions are most prevalent in the top 10 rankings. The respondents rated 21 of the vulnerabilities to have an impact higher than 4 out of a 5 point scale, and 15 of the vulnerabilities had an impact between 3.8 and 3.3.

6. Discussion

Figure 1 shows the normalized weighted averages for all 5 ICS dimensions. When comparing the overall in terms of the 5 ICS dimensions, it can be seen that security administration and architecture dimensions had the lowest weighted averages for awareness amongst respondents. These two ICS dimensions also shows high ratings of impact, where Security Administration obtained the highest rating for impact of the ICS dimensions. Security administration also has one vulnerability element (Security Administration 4) which was not covered in the organization’s documentation.

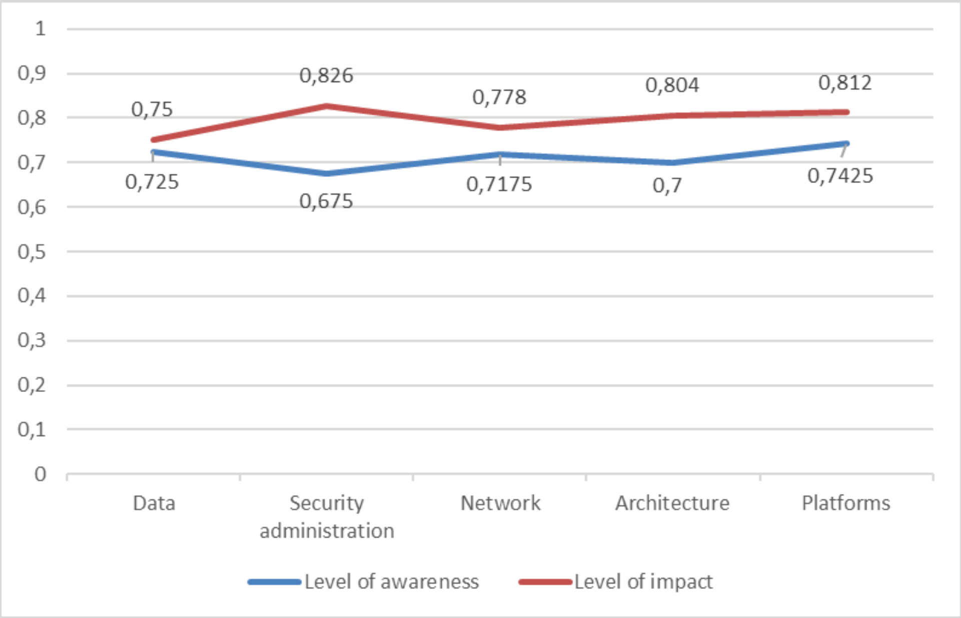


Figure 1: Awareness and Impact for the 5 ICS dimensions

The security administration involves the organization's policies and procedures which are used to address the security vulnerabilities in the ICS environment. The potential cause for a low awareness can be attributed to little or no training in the organization's cybersecurity policies and procedures. When looking at the results of the respondent’s background information, it can be seen that the majority of the respondent’s role in the organization is engineering driven in the ICS environment. The security administration in the organization is an administrative activity and could be a weakness for the engineers. The dimension with the highest level of impact was the security administration. This means that the respondents consider the security administration to have the biggest impact on the organization's operational function and that it is recognized as a problem by the respondents. Comparing the low level of awareness to the large impact of security administration as indicated by the respondents, it is evident that the security vulnerabilities pose a risk and that awareness for this dimension should be a top priority for the organization.

The data dimension was considered by the respondents to have the lowest impact on the organization. This means that respondents may not realize how important it is to protect the data infrastructure in the ICS environment, which can pose a danger to the operations of the organization.

7. Conclusion

This paper investigated the cybersecurity awareness levels of employees at an ICS organization in South Africa and measured their knowledge on the potential impact of cyber-related attacks on their systems through a case study. In this case study, benchmark data was collected from the literature to categorize vulnerabilities 5 dimensions of the ICS environment. This data was used to create a questionnaire to collect data on the level of awareness and the level of impact from employees of an ICS organization. The benchmark data was also used determine the documentation that the organization has relating to the benchmark ICS vulnerabilities.

Industrial control systems are becoming more integrated with supporting IT technologies, which means that ICS systems are also becoming more susceptible to cyber-related attacks, including theft of data, personal information exposure, financial loss and sabotage. These attacks can have severe effects on the operations of the organizations. Cybersecurity policies of organizations as well as cybersecurity training relating to these policies are critical factors to ensure that the ICS environments in organizations are adequately protected. There exist cases, such as the case study presented in this research, where organizations have employees who understand the severity of cyber vulnerabilities, however, their awareness is low. In addition, companies might not have adequate cybersecurity training programs.

As case study was conducted on a small ICS organisation in South Africa, the results are not generalisable, but does provide an indication of the shortcomings many small ICS organizations may face. A shift is required in organizations to move from a governance checklist to a cybersecurity culture. This shift should be facilitated though communication programs, followed up by training programs and a continuous monitoring of these to establish a cybersecurity culture. In addition, it can be recommended that the management of small companies start to take responsibility in enforcing the cybersecurity policies and guidelines and keep track thereof. The employees will start to improve their cybersecurity by becoming more knowledgeable about the cybersecurity policies and procedures of the organisation.

References

Byres, E. and Cusimano, J. (2012) “7 Steps to ICS and SCADA Security”. Tofino Security. Available at: https://www.tofinosecurity.com/sites/default/files/WP-7-Steps-to-ICS-Security.docx.

Byres, E. and Lowe, J. (2004) “The myths and facts behind cyber security risks for industrial control systems”. VDE Kongress, pp 213-218.

ICS-CERT. (2018) “Cyber-Attack Against Ukrainian Critical Infrastructure”. Available at: https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01 (Accessed: 16 April 2019).

Kaspersky. (2019) “Kaspersky Industrial Cybersecurity Training Program Training with Kaspersky Lab ICS CERT”.

Knowles, W., Prince, D., Hutchison, D., Disso, J.F.P. and Jones, K. (2015) “A survey of cyber security management in industrial control systems”. International Journal of Critical Infrastructure Protection, 9, pp 52-80.

Kortjan, N. and von Solms, R. (2013) “Cyber Security Education in Developing Countries: A South African Perspective”. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 289-297.

Krotofil, M. and Gollmann, D. (2013) “Industrial control systems security: What is happening?”. 11th International Conference on Industrial Informatics (INDIN), IEEE, Bochum, Germany, pp 670-675.

National Cyber Security Centre Security for Industrial Control Systems. (2015) “Improve Awareness and Skills - A Good Practice Guide”. Available at: https://www.ncsc.gov.uk/content/files/protected_files/guidance_files/SICS - Improve Awareness and Skills Final v1.0.pdf.

Nexia International. (2017) Global Cybersecurity Report. URL Available at: https://www.cohnreznick.com/-/media/resources/global_cybersecurity_report_2017.pdf (Accessed: 15 April 2018).

Piggin, R.S.H. (2013) “Development of industrial cyber security standards: IEC 62443 for SCADA and Industrial Control System security”. in IET Conference on Control and Automation 2013: Uniting Problems and Solutions, pp 1-6.

Pretorius, B. and van Niekerk, B. (2016) “Cyber-security for ICS/SCADA: a South African perspective”. International Journal of Cyber Warfare and Terrorism, 6 (3), pp 1-16.

Robson, C. and McCartan, K. (2016) “Real World Research”. Wiley.

Singer, P.W. (2015) “Stuxnet and its hidden lessons on the ethics of cyberweapons Case”. Western Reserve Journal of International Law, pp 79.

Smith, C. (2019) “Major spike in SA cyber attacks, over 10 000 attempts a day - security company”. Fin 24.

Stamp, E., Young, Y., and Mae Depoy, J. (2003) “Common vulnerabilities in critical infrastructure control systems”. Sandia National Laboratories.

Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M. and Hahn, A. (2015) “Guide to Industrial Control Systems (ICS) Security”. National Institute of Standards and Technology (NIST).

Symantec Security Response. (2017) “What you need to know about the WannaCry Ransomware”. Available at: https://www.symantec.com/blogs/threat-intelligence/wannacry-ransomware-attack.

Thomson, K.-L., von Solms, R. and Louw, L. (2006) “Cultivating an organizational information security culture”. Computer Fraud & Security, 10, pp 7-11.

TrendMicro. (2016) “Securing ICS environments in a connected world”. Available at: https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/securing-ics-environments-in-a-connected-world

Von Solms, R. and von Solms, B. (2004) “From policies to culture”. Computers & Security, 23 (4), pp 275-279.

Vroom, C. and von Solms, R. (2002) “A Practical Approach to Information Security Awareness in the Organization”. in Ghonaimy, M.A., El-Hadidi, M.T. and Aslan, H.K. eds. Security in the Information Society: Visions and Perspectives, Springer US, Boston, MA, pp 19-37.

Zikmund, W.G., Babin, B.J., Carr, J.C. and Griffin, M. (2010) “Business Research Methods”. South-Western, Cengage Learning.

Ransomware Attacks Against Critical Infrastructure

Aunshul Rege and Rachel Bleiman
Temple University, USA
rege@temple.edu
rachel.bleiman@temple.edu

DOI: 10.34190/EWS.20.115

Abstract: Critical infrastructures are those entities (energy, communications, finance, etc.) critical to the everyday functioning of a country, that their incapacitation or destruction would have devastating consequences for the country's security and daily operations. This paper analyzes a critical infrastructure ransomware dataset (created by the authors), which has 512 incidents from 2013 to present day gathered from publicly accessible media sources. This paper provides an exploratory overview of current trends in frequency of ransomware attacks against critical infrastructure from 2013 to date; frequency and duration of attacks; which ransom amounts are more often demanded by cybercriminals; and which ransomware strains are more popular.

Keywords: ransomware; critical infrastructure; cybercrime; extortion

1. Introduction

Critical infrastructures are those entities that are critical to a society's everyday functioning and security. The United States Department of Homeland Security has identified 16 critical infrastructures, namely: chemical; commercial facilities; communications; critical manufacturing; dams; defense industrial base; emergency services; energy; financial services; food and agriculture; government facilities; healthcare and public health; information technology; nuclear reactors and waste; transportation systems; and water and wastewater (CISA 2019).

Ransomware attacks targeting critical infrastructures have surged in the past few years. Ransomware is a malicious software that allows cybercriminals to lock and encrypt a target's computer or data, and then demand a ransom payment to restore access (Johansen 2019). Targets are expected to pay this amount within a set amount of time or risk losing access to their data (documents, photos, financial data) (Johansen 2019). Cybercriminals can target individuals, but are more likely to target those who can, and are more likely to, pay larger ransoms (Johanson 2019). These targets could include: organizations having smaller security teams (and thus cannot effectively respond); organizations that can and will pay quickly (hospitals, banks, and government agencies); organizations that hold sensitive data (law firms and research labs with intellectual property); and businesses in Western countries (as more of their operations are conducted online and are more likely to be crippled by a ransomware attack) (Johanson 2019).

This paper offers an exploratory analysis of ransomware attacks against critical infrastructure using a dataset of 512 incidents. This paper will first provide a review of the existing literature and information about ransomware, including its taxonomy, recent incidents, group dynamics, responses, and impacts. Next, the paper details the methods the authors used to generate the ransomware dataset and any limitations of it. The next section of the paper provides the analysis of several frequency tables created from the dataset. Finally, the paper concludes by acknowledging the limitations of this dataset and highlighting the need for more accessible data on ransomware incidents against critical infrastructures. The authors acknowledge that this dataset serves as a starting point, which can be used to explore current trends and identify future areas of research.

2. Literature review

This section offers a review of the most recent information on ransomware, focusing on ransomware taxonomy, recent critical infrastructure ransomware cases, ransomware groups' organizational dynamics and operations, responses to ransomware attacks, and impacts to critical infrastructures.

2.1 Ransomware taxonomy

Research has identified seven common types of ransomware (Johanson 2019). *Crypto malware* encrypts files, folders, and hard drives. A popular crypto malware is 'WannaCry', which targeted thousands of computer systems globally and demanded ransom payments in Bitcoins (Johanson 2019). *Locker-ransomware* infects operating systems to completely lock victims out, which prevent access to files and applications (Johanson 2019).

Scareware is fake software that pretends to operate as an antivirus or cleaning tool; it identifies 'issues' in the target's computer that requires payments to resolve the problem (Johanson 2019). Scareware can either lock computers or flood screens with numerous alerts and pop-up messages. *Doxware*, also known as leakware or extortionware, threatens to publish a victim's information (obtained through a data breach and then sold on black markets); this information could be sensitive files or photos (Johanson 2019). *RaaS*, or ransomware-as-a-service (detailed further in section 2.3), is a type of malware where cybercriminals develop and share their ransomware, which is used by others to target victims; cybercriminals then get a cut of any ransom payment collected through this ransomware (Johanson 2019). *Mac ransomware* targets mac operating systems and encrypts victims' files (Johanson 2019). *Mobile ransomware* is delivered through a malicious application that leaves a message on the device stating that it has been locked due to illicit activity (Johanson 2019).

2.2 Recent critical infrastructure ransomware incidents

In 2019 alone, several ransomware attacks have targeted critical infrastructure. The LockerGoga ransomware family impacted two chemical manufacturers in the United States and the Norsk Hydro in Norway just last March (Dragos 2019). Not only did this ransomware result in costly operational disruptions, but also made restoration difficult. In July 2019, IT Systems at South Africa's City Power suffered a ransomware attack, which prevented them from purchasing prepaid electricity via online systems and also from loading prepaid power to their meter boxes (Dragos 2019). Also consider the October 2019 ransomware attack experienced by the German company Pilz, which is one of the world's largest producers of automation tools. The BitPaymer ransomware impacted "all servers and PC workstations", email service, and ability to check orders (Cimpanu 2019, p. 1.). Petroleos Mexicano, a Mexican state oil and gas company, experienced a ransomware attack in November 2019 that "crashed its servers, encrypted files and halted some administrative work" (Kass 2019, p.1.).

In 2020, a United States natural gas compression facility suffered from a ransomware attack that led to a two-day operational shutdown (Palli 2020). Attackers accessed the facility's information technology (IT) network via a spear phishing attack, and then pivoted to the operational technology (OT) network (Palli 2020). Attackers then used ransomware in both the IT and OT networks to encrypt data throughout the facility (Palli 2020). A new form of ransomware 'Ekans' discovered in 2020 goes beyond the typical ransomware traits encrypting files and disabling data backups. Before the encryption process, the ransomware "kills 64 processes listed by process name in a hard-coded list" (Goodin 2020, p.1.). Experts have identified Ekans as a 'primitive' ransomware attack, because it simply kills processes (unlike malware that can cause more serious damage to critical infrastructure) and has no mechanism to spread, it is nonetheless an indication of increasing intentionality that can now impact OT (Goodin 2020).

2.3 Criminal modus operandi and organizational dynamics

Cybercriminals behind ransomware attacks are difficult to identify as they operate under the veil of anonymity: bitcoin and anonymous messaging platforms (Popper 2020). Jurisdictional matters also complicate matters as cybercriminals operate in countries that are outside the reach of the targeted country's legal system, do not face the threat of extradition, and in some cases operate under the protection of the country where the cybercriminals are based (Popper 2020).

Despite these challenges, some insights into the criminal modus operandi of ransomware groups can be obtained. Cybersecurity experts suggest that ransomware has developed into a lucrative business with cybercriminals competing for the most lucrative victims (Popper 2020). Cybercriminals can spend months quietly conducting reconnaissance on potential targets' computer systems to identify all critical assets and files that would need to be encrypted to ensure a higher likelihood of receiving ransomware payments (Popper 2020). Many sophisticated ransomware groups immediately offer the decryption keys to their victims upon receiving payments to ensure that their group retains its 'good business reputation' (Popper 2020; Vijayan 2020).

In addition to 'brand management', ransomware groups engage in 'return on investment' decisions like legitimate businesses, and make payment demands accordingly. One of these investments include infrastructures; ransomware operators leverage other networks for distributing their product, which has an average estimated operational cost of USD 1,044 (Hurley 2019). Ransomware operators must also invest in research and development to ensure that their product can do maximum damage to reduce the chances of a quick recovery; this requires continuous improvements to their codes and encryption (Hurley 2019). Ransomware code is polymorphic, which ensure the code to change its signature with each infection thereby making it harder to detect (Hurley 2019). Encryption updates are also essential as this hinders security researchers from developing decryption keys for victims (Hurley 2019). Finally, ransomware operators must maintain a system for tracking their victims and payments to monitor the victims that have been targeted, which have paid (and thus who should get decryption keys), and which have not cooperated (to follow up and ensure they do not receive decryption codes) (Hurley 2019). Collectively the costs of infrastructure, research and development, and victim trafficking determine the amount of ransom payments, and gives the ransomware group the "flexibility to negotiate or walk away from a transaction, or to create pricing tiers for different types of victims" (Hurley 2019, p. 4).

Cybercriminals also specialize in 'ransomware-as-a-service' or RaaS, which is available for rent to affiliates who can afford the service subscription fee. Alternatively, RaaS developers may collect all ransom payments from affiliates, take out their portion as commission, before passing the remainder back to the affiliates (Peters 2019). RaaS are deployed using simple, easy-to-use interfaces and requires no coding (Peters 2019). Furthermore, RaaS developers offer online customer support to "help subscribers get their ransomware campaigns up and running" (Peters 2019, p. 3). One such service, GrandCrab, generated USD 2 billion in 2019 and went off the market in June, "when its creators claimed they were retiring" (ProofPoint 2020). This RaaS service surpassed its competitors because its developers worked diligently to provide five major revisions with bug fixes to ensure their product would bypass antivirus and other defense measures (KrebsOnSecurity 2019). Furthermore, GrandCrab developers offered their users support, such as providing step-by-step instructions and customization options (Palmer 2018).

These ransomware and RaaS groups can have sophisticated organizational dynamics. Consider again the GrandCrab group. While the exact details of group membership are difficult to ascertain, research suggests that the RaaS group followed an 'affiliate marketing business model'. In this model, the enterprising cybercriminals (affiliates) use the product (ransomware) to do the "heavy lifting" of spreading the ransomware and finding victims, which allowed the developers to revise their product, add new features, and improve its encryption technology (Malwarebytes 2019). Any ransom payment paid is then split between the affiliate and the GrandCrab RaaS developers "60/40 or as high as 70/30 for top affiliates" (Malwarebytes 2019, p.2).

Unlike data breaches, where cybercriminals must find buyers and engage in price negotiation, ransomware and RaaS attacks offer "built-in buyers: businesses who are locked out of their systems, who are often not in a position to negotiate on price" (Peters 2019, p. 3). Approximately 80% of organizations in the United States got their data back after initial payment (Proofpoint 2020). However, not all ransomware operators follow up on their word and offer decryption tools for their targets upon receiving payment; 18% of organizations in the United States never got access to their data, and 2% had to pay additional ransom demands than the initial ransom demand and still did not get data access (Proofpoint 2020).

2.4 Responding to ransomware attacks

Three possible responses to ransomware attacks are found in the literature. The first involves *targets giving into their ransomware attackers' demands*. Targeted organizations may be more willing to pay as this might be the cheapest option for them to ensure the availability of their systems or services (Stokel-Walker 2019). Some organizations may even have insurance to assist with the payments. Consider Lake City, which paid its attackers USD 10,000 of the USD 530,000 ransom, and the insurance company paid the balance (Stokel-Walker 2019). Security researchers note, however, that paying does not always guarantee unlocked access to the encrypted data. In fact, victims might get targeted again with further demands or be subjected to future attacks; "criminals distribute 'suckers lists' of those who have proven susceptible to extortion" (Stokel-Walker 2019, p. 2).

Organizations may wish to *negotiate* with their attackers, as ransomware is a rare type of attack where victims have an opportunity to interface with the ransomware group (Marshall 2019). Experts advice for organizations is "first, you size up your opponent, then you exchange messages, and ultimately, you try to reach some kind of agreement" (Marshall 2019, p. 2). Negotiations might be beneficial to organizations, especially if they do not have any real-time backup systems in place (Marshall 2019). Negotiating for extra time and a lower ransom amount can also prove beneficial as cybercriminals would "much rather get something than nothing" (Murphy 2019, p. 3). Alternatively, negotiations are a great way to assess the attackers' knowledge and 'sincerity'. Organizations can ask for a "digital equivalent of a 'proof-of-life'", where the attackers provide decryption for part of the locked files taken hostage (Murphy 2019, p. 3). If the attackers refuse, it may indicate that they either do not have the decryption key and would not have unlocked the files even after receiving payment (Murphy 2019).

Security experts therefore recommend that *organizations have their data correctly and frequently backed up*. Organizations can select backups made prior to the ransomware infection to resume functionality and operations (Bauer 2019). Security experts recommend following the "3-2-1 rule, which advocates having 3 different copies of data on 2 different media with one of them being off-site" (Abrams 2020, p. 5). Unfortunately, backups may not always work. During the initial ransomware infection, attackers can spread through the entire system and obtain administrator credentials, which can give them access to backup software (Abrams 2020). They may delete backups to ensure that they cannot be used to restore encrypted files (Abrams 2020). Alternatively, cybercriminals may not just encrypt sensitive data; they may also exfiltrate or copy the data. If an organization refuses to pay because it has backups, attackers may then demand payment to ensure they do not leak the sensitive data they exfiltrated.

2.5 Impacts to critical infrastructures

The United Sates Federal Bureau of Investigation noted that between October 2013 and November 2019, ransomware cybercriminals received USD 144 million in bitcoins from their victims (Kass 2020). EMSISOFT (2020) provided estimated ransomware demand costs across different countries, with the United States at the top (USD 1,373,248,80), followed by Italy (USD 641,995,200), Germany (USD 592,542,720), and Spain (490,089,600). When downtime costs were added to the ransomware demand costs, the United States again ranked highest (USD 9,299,648,800), followed by Italy (USD 4,347,595,200), Germany (USD 4,012,702,720), and Spain (USD 3,318,889,600) (EMSISOFT 2020).

Of course, the impact of ransomware attacks is not limited to financial loss (ransomware payment or downtime costs). Other consequences will depend on the type of infrastructure being targeted and the good and services it offers. In 2019, the United States experienced an unprecedented amount of ransomware attacks that targeted "113 state and municipal governments and agencies, 764 healthcare providers, and 89 universities, colleges, and school districts" (EMSISOFT 2019). In addition to the financial loss discussed earlier, these ransomware attacks caused several disruptions and inconvenience to both the organizations and their customers. Some hurdles experienced by government agencies included "police [being] locked out of background check systems and unable to access details about criminal histories or active warrants; Surveillance systems [going] offline; badge scanners and building access systems [not working]; and [inability to issue or renew] driver's licenses" (EMSISOFT 2019). For healthcare providers, "emergency patients had to be redirected to other hospitals; medical records were inaccessible and, in some cases, permanently lost; and surgical procedures were canceled, tests were postponed, and admissions halted" (EMSISOFT 2019). Education institutions suffered as they "could not access data about students' medications or allergies; lost students' grades; and [even] closed [down]" (EMSISOFT 2019).

Ransomware attacks against critical infrastructure have grown tremendously in the past few years and are run by organizations that are very sophisticated in operation and organization. Furthermore, the impacts of ransomware attacks are indeed devastating and include financial loss, disruption to critical services and information, threat to public health and safety, and overall inconvenience. It is difficult to get a handle on the exact costs of these attacks and identify the details of these groups as not all incidents are reported, and groups operate through covert and anonymous mechanisms. Despite the growing relevance of ransomware attacks, there is a severe lack of publicly accessible, aggregated data on ransomware incidents against critical infrastructure.

3. Methodology

This section details the methods the authors used to generate and maintain a dataset on ransomware attacks against critical infrastructure.

3.1 Dataset generation and maintenance

In order to study the impact of ransomware attacks on critical infrastructure, the authors generated a ransomware dataset based on incidents that were publicly disclosed to the media. To develop this dataset, the authors utilized the service Google Alerts using the keyword *ransomware.* This service sent them daily email notifications about newly posted articles whose titles contained the specified keyword. The alerts produced a list of links to articles on ransomware incidents as well as on general information about new ransomware strains, preventative measures, and recent trends; the authors sorted through these lists to find the appropriate articles. In order to include older ransomware incidents, the authors used Google search engine to discover prior attacks; they also followed hyperlinks found in articles about recent incidents that linked to reports on previous attacks. News sources included local news websites, organizations' website postings, and technology-oriented news websites, among others. For each ransomware incident, the authors recorded data about several variables (Table 1) which they maintained in a Microsoft Excel document that was generated over a five-month period. Because the data on these incidents was gathered from media sources, information on some variables was often undisclosed and missing data was reported as such in the dataset. This process generated a dataset of 512 critical infrastructure ransomware attacks.

Table 1: Dataset variables and definitions

Variable Name	***Variable Definition***
ID Number	The unique identifier of each incident, assigned sequentially with incidents sorted from oldest to newest
Date Began	The date that the ransomware attack was first identified by the organization
Year	The year that the ransomware attack was first identified by the organization
General Date	If the *Date Began* is unknown: the first day of the month or year that is known (e.g. "In March of 2016" → 03/01/2016; "In 2016" → 01/01/2016)
Organization Name	The name of the organization that was attacked with ransomware
Location	The US state (or country, if outside US) of the organization that was attacked.
CIS Targeted	The critical infrastructure sector(s) to which the organization belonged, categorized using the United States Department of Homeland Security's 16 identified critical infrastructures, with the additional sector of Education Facilities
HowTargeted2:	The specific strain of ransomware used in the attack
Duration	If ransom was paid: the time it took for the organization to pay the ransom and recover If ransom was not paid: the time it took for the organization to recover
Duration Rank:	The nearest (without exceeding) duration rank: 1 day or less, 1 week or less, 1 month or less, more than 1 month
Ransom Amount	The ransom demand in the original currency of the demand (bitcoin, USD, Euro, etc.)
Local Currency	The monetary amount demanded by the ransom, converted to USD
Ransom Amount Rank	The nearest (without exceeding) ransom amount rank in USD: $1,000 or less, $50,000 or less, $100,000 or less, $1 million or less, $5 million or less, more than $5 million
Paid Status	Whether or not the organization paid the ransom to the attackers (Yes/No)
Pay Method	If *Paid Status* is Yes, the method the organization used to pay the ransom to the attackers (Bitcoin, Cash, Western Union, etc.)
Amount Paid	If *Paid Status* is Yes, the ransom amount that the organization paid to the attackers, in the Pay Method unit
Source:	The URL of the article from which the incident was identified
Related incidents	The list of incidents by *ID Number* thought to be part of the same attack
Comments	Any significant details not covered by other variables (e.g. declaration of state of emergency, total cost to rebuild, etc.)

3.2 Limitations

As with any dataset and methodology, the ones used in this paper also have their own limitations. First, the dataset was assembled along the critical infrastructure taxonomy identified by the United States Department of Homeland Security. Other countries, and even other organizations in the United States, may use a different taxonomy, which may impact the analysis of incident occurrence, frequency, duration, and comparisons. Second, the dataset only documents those incidents that are publicly disclosed (primarily in the media). Depending on the type of resource, the information provided in the source may lack depth on type of ransomware, technical details, duration of attacks/recoveries, and whether the amount was paid (and how much). Additionally, not all incidents are reported to the media. Victims may avoid disclosing attacks to the media to avoid reputational losses or to hide vulnerabilities that could serve as an invitation for future attacks. These issues also limit the analysis offered in this paper. Third, despite the authors search for both older and newer ransomware incidents, the use of Google Alerts as the main tool to collect data directed the focus to more recent cases. While there was easily accessible information on well-publicized older cases, it was more difficult to find less-publicized cases from previous years than the current year. Finally, the use of other search engines or keywords for alerts may have resulted in a slightly different set of cases.

However, given that this is an exploratory paper that examines critical infrastructure ransomware incidents over time, across infrastructures, attack duration, and payments demanded, the dataset serves as a starting point to understand the overall trends of this phenomenon, and to inform future areas of research (see section 5).

4. Analysis

Based on the dataset of ransomware attacks, we examined five of the variables in terms of frequency of attacks. These variables included the following: CI targeted, year of attack, attack/recovery duration, ransom amount, and ransomware strain.

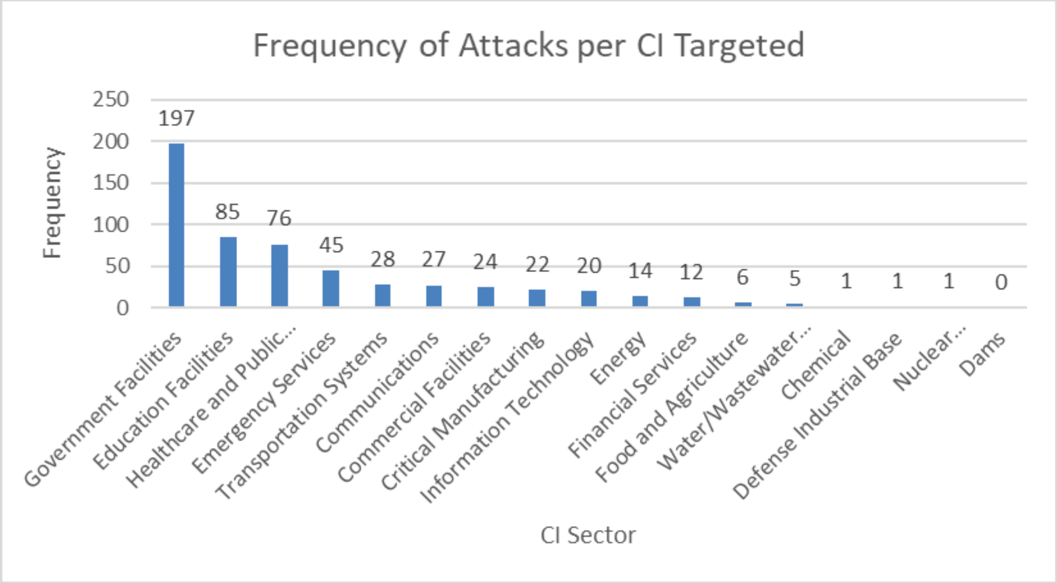


Figure 1: Frequency of Ransomware attacks per CI

Figure 1 describes the frequencies of attacks per critical infrastructure targeted. According to this data, government facilities (embassies, courthouses, general-use office buildings) were the most frequently targeted CI with at least 197 reported attacks gathered from media sources. This may imply a vulnerability in government computer systems that allow attackers to easily target them or it implies that government facilities consistently lack sufficient data back-ups to be used in cases of emergencies. Education facilities, healthcare/public health, and emergency CI subsectors were also frequent ransomware victims with 85, 76, and 45 reported instances, respectively. In addition to possible system vulnerabilities, this data may imply that attackers target CI organizations that can and will pay quickly, due to the need and urgency to restore their operations, such as schools, hospitals, or police and fire departments. While CI sectors such as chemical or defense industrial bases were still targeted, they were either targeted much less frequently (because they were more secure) or they were unwilling to publicly disclose their cyberattack incidents to the media.

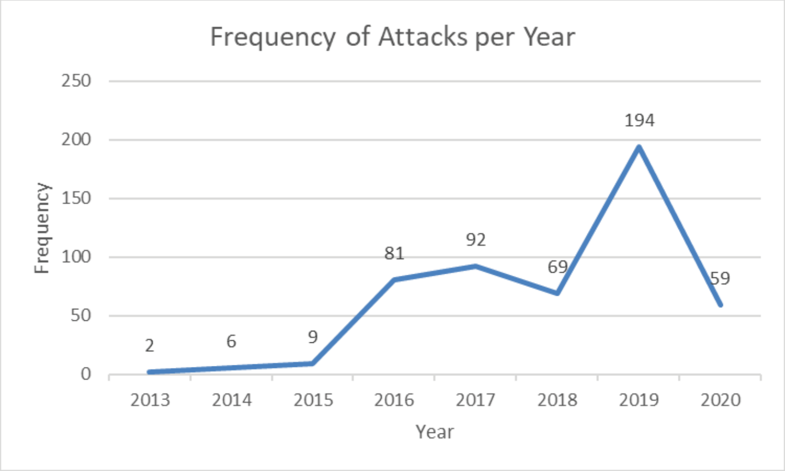


Figure 2: Frequency of Ransomware attacks per year

Figure 2 depicts the frequency of ransomware attacks per year. With data ranging from 2013 to March 2020, this data confirms that there has been a rise in attacks over recent years. Particularly in 2019, the frequency of attacks more than doubled from previous years. This implies that ransomware attacks are becoming a more common attack method and are being used more frequently. This may be due to the growing online operations of corporations and more sensitive data being stored online, increasing the vulnerability to cyberattacks. As noted above, not all incidents are not reported to the media, which may alter the exact frequency numbers.

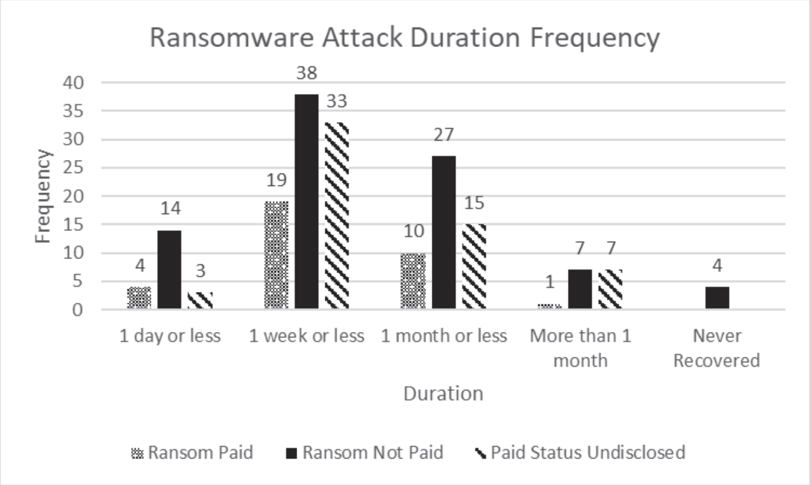


Figure 3: Attack duration frequency, grouped by payment status *(paid, not paid, paid status undisclosed)*

Note: Attack duration covers the time from attack initiation to target recovery (irrespective of ransomware payment)

Figure 3 depicts the ransomware attack duration frequency, which includes the amount of time between the start of the attack and when the organization recovered from the attack. Incidents are grouped by payment status (paid, not paid, paid status undisclosed). The ‘paid’ group recovered via paying the ransom; the ‘not paid’ group recovered via backups or other means. The recovery duration of the ‘paid’ group is tantamount with the duration of time the organization took to pay the ransom, as recovery after payment was instantaneous. Among every group, the most frequent recovery duration was in the range of one week or less, which suggests two possibilities. The first possibility is that targeted organizations who paid the ransom recovered their data immediately after being targeted (they received decryption codes). The second possibility is that targeted organizations who did not pay the ransom recovered quickly as they had sufficient backup data to rebuild and continue operations. This finding suggests that security experts’ recommendations for maintaining backups is an effective strategy for managing ransomware attacks and not giving into the demands of extortionists. There were at least 4 reported cases of organizations, none of which paid the ransoms, that were unable to recover from ransomware attacks to the point that they either permanently lost significant data or had to close permanently. For example, Wood Ranch Medical, a healthcare practice, was unable to recover from an attack in September 2019, resulting in its shut down in two months.

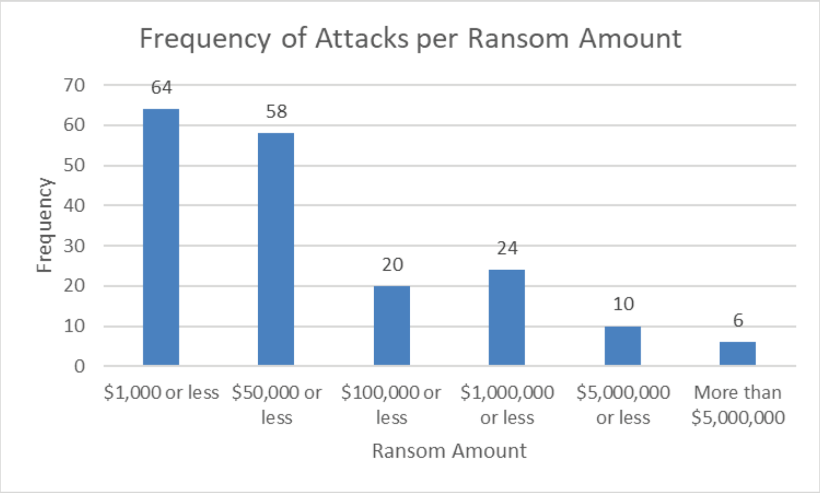


Figure 4: Frequency of ransomware attacks per ransom amount

Figure 4 describes the frequency of attacks per ransomware amounts. Most frequently, attacks had ransoms of either USD 1,000 or less, or USD 50,000 or less. Ransom amounts negatively correlated with frequencies of attacks; attacks with greater ransom demands happened less frequently than attacks with smaller ransom demands. This may imply that when attackers demand ransoms in this range, the demands are small enough that the victims consider paying them. Fewer numbers of attacks demanded greater ransom amounts, but at least 6 attacks demanded more than USD 5 million, one of which was an attack on the city of Sarasota that demanded up to USD 33 million in bitcoin. While the payment status was undisclosed for 185 of the incidents, 67 cases have confirmed ransom payments made. Of these 33 incidents, 13 ransoms demands were USD 1,000 or less, 24 were USD 50,000 or less, 1 was USD 100,000 or less, 9 were USD 1 million or less, 2 were USD 5 million or less, and 18 ransom demand amounts were undisclosed. This shows that the smaller ransom amounts were indeed the ones that were paid most frequently.

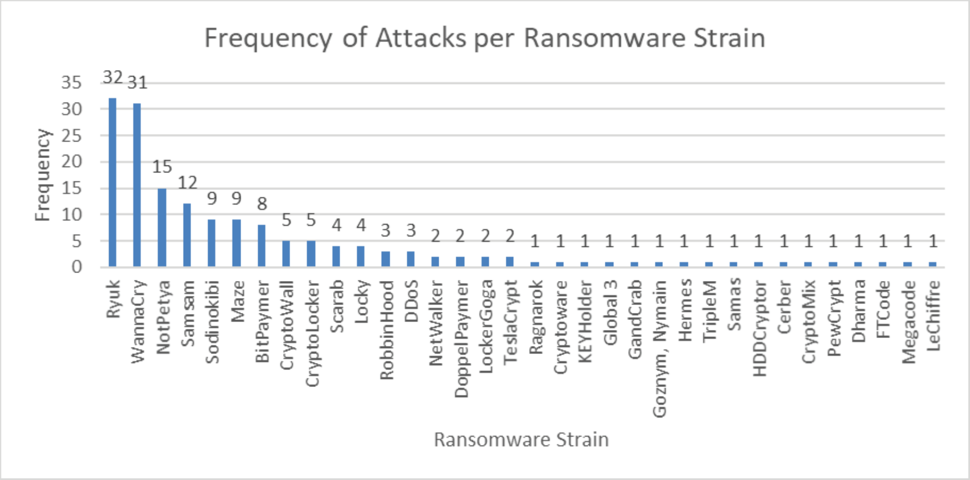


Figure 5: Frequency of ransomware attacks per ransom strain

Figure 5 depicts the frequencies of attacks per ransomware strain. The most frequently used strain of ransomware was 'Ryuk' with at least 32 reported instances. Other commonly utilized strains were 'WannaCry' and 'NotPetya' with 31 and 15 reported instances, respectively. The high frequencies of attacks using these strains imply coordinated attacks. One particular strain may be used in a series of attacks, possibly from the same group of attackers. This may also imply that particular strains of ransomware are more successful in an attack and are harder for security experts to detect or fight against. There were also several strains that were only reported to be used once. This may imply that there were one-time instances of attacks or that certain strains of ransomware were less successful than others.

Overall, the results of the analyses support the claims that ransomware attacks are on the rise; attackers target less secure CI that can and will pay ransom demands quickly; targeted organizations are more likely to pay smaller ransoms; ransoms are most frequently paid within one week of the demand; and ransomware groups may repeatedly use the same strain of malware to target their victims.

5. Conclusion

This paper provides an exploratory study of the state of critical infrastructure ransomware cyberattacks globally. The analyses provided here are based on publicly disclosed incidents covered in the media, and we are aware that they limit the generalizability of our findings. The authors recognize and appreciate that critical infrastructures may not wish to publicly disclose ransomware incidents as these may harm the infrastructures' reputation, publicize their vulnerabilities and possibly subject them to further ransomware (or other) attacks, and result in liability issues. While the authors are using the dataset as a starting point to explore trends in ransomware attacks, this dataset can help identify directions for future research. For instance, researchers can interview security experts about the characteristics of the most commonly used ransomware strains and determine what makes them so successful. Researchers can also work with the infrastructure sector to develop mechanisms for disclosing and sharing incidents, which could improve the quality of the dataset analysed in this paper. The authors are not aware of any critical infrastructure ransomware datasets that are currently freely available for research. The authors assembled this dataset based on publicly disclosed incidents that received media coverage (as noted in section 3.1). The authors would like to note that this dataset has been made available as a free resource, and has already been utilized by *industry* (for generating trends reports to inform board of directors and budget allocations), *researchers* (for using unique approaches such as econometrics and machine learning), *educators* (for their classrooms), and *undergraduate and graduate students* (for their research projects and dissertations). These download requests clearly demonstrate the need for such resources to advance research and education in this space.

Indeed, the authors were able to analyze the dataset they created to offer an important starting point; it provides an exploratory overview of current trends in frequency of ransomware attacks against critical infrastructure from 2013 to date; frequency and duration of attacks; which ransom amounts are more often demanded by cybercriminals; and which ransomware strains are more popular. Of course, more research is needed in this space, which requires having relevant and good quality datasets.

Acknowledgements

This material is based upon work supported by the National Science Foundation CAREER Award, Grant No. CNS-1453040.

The critical infrastructure ransomware incident dataset is available for download request at sites.temple.edu/care/downloads.

References

Abrams, L. (2020). "Ransomware Attackers Use Your Cloud Backups Against You". Retrieved March 8, 2020. Online at https://www.bleepingcomputer.com/news/security/ransomware-attackers-use-your-cloud-backups-against-you/

Bauer, R. (2019). "Ransomware: How to Prevent Being Attacked and Recover After an Attack". Retrieved January 25, 2020. Online at https://www.backblaze.com/blog/complete-guide-ransomware/

CISA. (Cybersecurity and Infrastructure Security Agency) (2019). "Critical Infrastructure Sectors". Retrieved February 1, 2020. Online at https://www.cisa.gov/critical-infrastructure-sectors

Cimpanu, C. (2019). "Major German manufacturer still down a week after getting hit by ransomware". Retrieved January 25, 2020. Online at https://www.zdnet.com/article/major-german-manufacturer-still-down-a-week-after-getting-hit-by-ransomware/

Dragos. (2019). "2019 Year in Review: The ICS Threat Landscape and Threat Activity Groups". Retrieved February 1, 2020. Online at https://dragos.com/wp-content/uploads/The-ICS-Threat-Landscape.pdf

EMSISOFT. (2019). "The State of Ransomware in the US: Report and Statistics 2019". Retrieved February 29, 2019. Online at https://blog.emsisoft.com/en/34822/the-state-of-ransomware-in-the-us-report-and-statistics-2019/

EMSISOFT. (2020). "Report: The cost of ransomware in 2020. A country-by-country analysis". Retrieved February 29, 2020. Online at https://blog.emsisoft.com/en/35583/report-the-cost-of-ransomware-in-2020-a-country-by-country-analysis/

Goodin, D. (2020). "New ransomware doesn't just encrypt data. It also meddles with critical infrastructure". Retrieved February 29, 2020. Online at https://arstechnica.com/information-technology/2020/02/new-ransomware-intentionally-meddles-with-critical-infrastructure/

Hurley, A. (2019). "Ransomware ROI from the criminal perspective". Retrieved January 25, 2020. Online at blog.barracuda.com/2019/08/22/ransomware-roi-from-the-criminal-perspective/

Johansen, A. (2019). “What is ransomware and how to help prevent ransomware atacks”. Retrieved January 25, 2020. Online at https://us.norton.com/internetsecurity-malware-ransomware-5-dos-and-donts.html

Kass, D.H. (2019). “Pemex Ransomware Attack: Mexico Oil, Gas Recovery Update”. Retrieved January 25, 2020. Online at https://www.msspalert.com/cybersecurity-breaches-and-attacks/ransomware/pemex-recovery-update/

Kass, D. H. (2020). “Ransomware Payments Top $140 million, FBI Reports”. Retrieved March 5, 2020. Online at https://www.msspalert.com/cybersecurity-breaches-and-attacks/ransomware/payments-top-140-million-fbi-reports/

KrebsOnSecurity.com (2010). “Who’s Behind the GrandCrab Ransomware?”. Retrieved January 25, 2020. Online at https://krebsonsecurity.com/2019/07/whos-behind-the-gandcrab-ransomware/

Malwarebytes (2019). “GrandCrab”. Retrieved January 25, 2020. Online at https://www.malwarebytes.com/gandcrab/

Marshall, P. (2019). “When ransomware strikes… negotiate?”. Retrieved January 25, 2020. Online at https://gcn.com/articles/2019/05/13/ransomware-negotiation.aspx

Palli, C. (2020). “Ransomware Attack Hit US Natural Gas Facility”. Retrieved February 29, 2020. Online at https://www.bankinfosecurity.com/ransomware-attack-hit-us-natural-gas-facility-a-13740

Palmer, D. (2018). “Ransomware gets easier for would-be crooks as developers offer malware-as-a-service”. Retrieved January 25, 2020. Online at https://www.zdnet.com/article/ransomware-gets-easier-for-would-be-crooks-easier-as-developers-offer-malware-as-a-service/

Peters, M. (2019). “What Is Ransomware-as-a-Service? Understanding RaaS”. Retrieved January 25, 2020. Online at https://securityboulevard.com/2019/02/what-is-ransomware-as-a-service-understanding-raas/

Popper, N. (2020). “Ransomware Attacks Grow, Crippling Cities and Businesses”. Retrieved February 29, 2020. Online at https://www.nytimes.com/2020/02/09/technology/ransomware-attacks.html

Proofpoint. (2020). “2020 State of the Phish Report”. Retrieved February 29, 2020. Online at https://www.proofpoint.com/sites/default/files/gtd-pfpt-us-tr-state-of-the-phish-2020.pdf

Stokel-Walker, C. (2019). “Ransomware attacks are on the rise and the criminals are winning”. Retrieved January 25, 2020. Online at https://www.newscientist.com/article/2208897-ransomware-attacks-are-on-the-rise-and-the-criminals-are-winning/

Vijayan, J. (2020). “Average Ransomware Payments More Than Doubled in Q4 2019”. Retrieved February 29, 2020. Online at https://www.darkreading.com/risk/average-ransomware-payments-more-than-doubled-in-q4-2019/d/d-id/1336893

Cyber-Informed: Bridging Cybersecurity and Other Disciplines

Char Sample[1,2], Sin Ming Loo[2,1], Connie Justice[3], Eleanor Taylor[1] and Clay Hampton[3]'
[1] Idaho National Laboratory, Idaho Falls, USA
[2]Boise State University, Idaho, USA
[3]Indiana University at Purdue, Indianapolis, USA
Charmaine.Sample@inl.gov, Eleanor.Taylor@inl.gov
smloo@boisestate.edu
cjustice@iupui.edu, cthampton@iupui.edu
DOI: 10.34190/EWS.20.092

Abstract: A recent study by Cybersecurity Ventures (Morgan 2018), predicts that 3.5 million cybersecurity jobs around the world will be unfilled by 2021. In the United States, the demand for professionals with cybersecurity expertise is outpacing all other occupations (NIST 2018). These reports, along with many others, underpin the need for increasing workforce development initiatives founded in cybersecurity principles. The workforce shortage is across all cybersecurity domains, yet problems continue to persist, as the lines between combatants and non-combatants are blurred. Combating this persistent threat, which is a 24/7 operation, requires a more aggressive and inclusive approach. Higher education institutions are positioned to fully support cybersecurity workforce development; cybersecurity needs people with different perspectives, approaches, ways of thinking, and methods to solve current and emerging cyber challenges. This need is especially pressing when assessing the digital landscape — a tireless and ever-expanding connectivity supported by societal needs, and economic development yet compromised by the common criminal to nation-state sponsored felonious activity. Educators need to consider augmenting their approaches to educating students to include cybersecurity content. In this technology forward world, one that is expanding more rapidly than society and policy can react, increases the imperative for fundamental cyber defence skills. Accordingly, all students, no matter the major, should, minimally, understand the implications of good versus bad cyber hygiene. STEM graduates will require awareness of cyber issues that impact the security of programs, systems, codes or algorithms that they design. Operationally focused cyber-security graduates require a curriculum for careers dedicated to protecting and defending cyber systems in domain specific environments. In a world of Internet of Things (IoT), the ability for individual disciplines to understand the impact of cyber events in environments outside of traditional cybersecurity networks is critically important. This will provide the next generation defenders with domain specific cybersecurity knowledge that is applicable to specific operating environments.

Keywords: cybersecurity curriculum, workforce development, cybersecurity curriculum design, cybersecurity degrees

1. Introduction

The warnings for the cyber-workforce shortfall have been sounded for years, yet the gap continues to grow (ISC2 2019, NIST 2018) Cybersecurity Ventures (2019) recently predicting a 3.5 million worker shortfall. The growth of the Internet of Things (IoT) (Morgan 2018) promises to exacerbate this problem. In general, the future digital landscape promises increasingly more complex cyber-attacks from criminals to nation-state actors making basic cybersecurity hygiene the responsibility of all users not just the cybersecurity indoctrinated.

In addition to the cyber workforce development gap, a divide between technology savvy citizens and their counterparts is also noteworthy. This coincides with the introduction of technology courses in the public-school system (Blazer 2008), along with the popularity of newer technologies, the ubiquitous use of devices, and proliferation of social media and mobile apps. A recent Pew research survey (2018) found that more than nine-in-ten teenagers, 13 to 17, reported owing or having access to a smartphone and almost half reported being on-line on a near constant basis. Technology adoption rates have soared across the US (Ritchie & Roser 2019) and serve to illustrate the need for greater cyber awareness and proficiency (aka cyber hygiene). As more of the population becomes cyber-aware, cyber hygiene knowledge could potentially address the cybersecurity skills gap while introducing different disciplines to cybersecurity.

The past 10 years Internet enabled home electronics sales increased making possible the ability for many home users to become unwitting accomplices in cyberattacks. In 2017 there were approximately 8.4 billion connected internet of things (IoT) with well over 20 billion predicted in 2020 (Gartner 2017). IoT devices such as thermostats, smart TVs, and appliances, offer greater convenience by leveraging an Internet connection, but at the cost of granting access to uninvited guests. Most users remain unaware of basic security settings on their Wireless Access Point (WAP), or the firewall settings provided by their Internet Service Provider (ISP) and how to modify these settings. For example, the re-purposing of IoT devices into bots for botnets creates an inadvertent increase in installation base and scale through IoT device default set-up increasing vulnerabilities, and opportunities for malicious attacks.

Widespread IoT deployment presently, and for the foreseeable future, supports the argument that cyber proficiency is as important as reading. In cybersecurity terms, our education system fails to keep up in the age of IoT. While adoption rates of apps and the latest incarnation of technology remains healthy, few have the deep understanding of how these products actually work, and the security ramifications of using these products in their native installed state (Lorenz 2020). Consider the outrage over Facebook's selling of personal data (Isaak & Hanna 2018), Amazon's application of AI/ML to use personal data to sell more products (Levy 2018), the Marai botnet attack using open ports and default passwords (Antonakakis 2017) and most recently the Zoom hacking (Lorenz 2020). These examples underscore the need for a curriculum of cybersecurity for non-professionals training.

The need to introduce cybersecurity into other disciplines is well documented (LeClair 2013; Henry 2017), the problem is in the details of "how" to do so. Efforts to integrate various disciplines and cybersecurity have had mixed results, suggesting that while the general idea has merit, the implementation may have faults. Boise State University (BSU) and Indiana University at Purdue (IUPUI) are two institutions in a growing number that have recognized the need for cybersecurity to extend beyond the traditional academic silo and make teachings available to wider audiences. What follows is a review of the experiences of two universities in extending cybersecurity beyond the traditional academic disciplines to develop a more interdisciplinary and inclusive approach which include, but are not limited to, computer science and engineering.

2. Background

Fundamentally, any digital system can be "hacked" by a persistent adversary and is often only bound by the imagination of the attacker and creativity of the defender (Carey and Jin 2019). As these events continue to gain complexity and sophistication, a cyber informed approach across several disciplines may increase both the skill and size of the future workforce. Much like the move to remote work resulted in many institutions adding Microsoft Office courses and certifications, the interdisciplinary nature of cyber security suggests a need to educate the masses in a similar fashion and approach, from introductory and basic usage through advanced and expert levels.

Introducing cybersecurity education into existing curriculum offers the opportunity to increase the number of cyber-aware workers in the workforce. In doing so, the workload decreases on cybersecurity professionals, and an additional potential outcome is attracting non-traditional majors to cybersecurity, especially as cybersecurity definitions continue evolving and expanding from STEM domains into other domains. Although challenges remain at universities for students, as course catalogues can be difficult to navigate, prerequisites can create barriers to entry, especially across colleges within the same system, and inabilities to identify or track extracurricular learning opportunities (Cybersecurity Curricula 2017).

The skills gap remains and there is currently no viable way to fill the demand without addressing future university curriculum and workforce development training processes. Meeting this challenge requires strategic education efforts to create a foundation for a broader talent pipeline in support of the needs facing industry, government and academia. The objective is not only to create awareness and interest in cybersecurity careers but also to develop and deliver cyber hygiene courses across universities and community colleges and increasing workforce development training while providing the community at large with tools and resources, to significantly increase education and training aligned with cyber objectives and career paths.

The current approaches to cybersecurity are not sustainable, scalable, or anticipatory (Borg 2018). The rapid adoption of digital technology, IoT, the promise of 5G, and integration into a vast and often unrecognized communications web will eclipse existing methods to identify and mitigate cyber risks. These factors drive the urgent need for expanding the cybersecurity talent pool, for addressing current needs, future innovations and protecting critical infrastructure.

Attaining digital safety in this increasingly connected world, requires problem solvers with different perspectives, approaches, ways of thinking, and methods (Nisbett 2010; LeClair 2013). The solution cannot be a purely computer-based curriculum approach. Historically, cybersecurity workforce preparation came from computer science and information technology majors, yet many industry leaders (i.e. Cliff Stoll, Dan Geer and Radia Perlman), do not have degrees in Computer Science or Cybersecurity. In fact, a gap exists between materials covered and field requirements. Computer Science is algorithm focused and IT courses are designed to train students to run and/or manage IT systems. Cybersecurity requires broader knowledge on protecting information, technology, processes and people.

There are several cybersecurity professionals, without computer science and information technology degrees, that have found their way to the discipline through informal methods and have learned the required skills and received the formal certifications as they progressed in their careers (Tribe of Hackers1, Tribe of Hackers2). Cybersecurity is multidisciplinary and benefits from a workforce of various majors across a university campus, expanding cybersecurity into holism and creating proactive problem solvers. By pushing the boundaries of a traditional STEM-based approach, with the goal of leveraging the learning environment and endorsing critical thought across all disciplines, students can learn valuable insights from each other. Borg (2018, p.2) discussed the perspective issue providing evidence that an over reliance on STEM courses can be detrimental leaving students "less equipped to do cyber security well".

Creating this new course curricula requires a true mix of disciplines where cybersecurity concepts can be shaped and shared in environments beyond traditional computer networks, where other data, beyond cyber event data, shapes the narrative. In this environment, students will learn comprehensively about cybersecurity with an operational as well as an informational focus covering information assurance, security operations, forensics, and disaster recovery — while providing the foundational knowledge needed to become a cybersecurity leader. Preparing the next generation of cybersecurity professionals with the necessary skills, requires continual investment in all areas of cybersecurity research, education, training, communication and outreach. This effort complements the work being supported by other education institutions; thus, combining our forces to help meet the ever-increasing cybersecurity workforce needs.

Cybersecurity is more than a technology solution, the discipline reflects the interactions of humans, machines, and networks. Cyber-safety prevails when human factors and other disciplines are included in the solution. The lateral thinking needed for cybersecurity is similar to epidemiology, requiring many different areas of expertise to understand the agent, host and environment. Using this metaphor, (Zhang et al., 2014; Zhang et al. 2015), cybersecurity depends on herd immunity - the resistance to the spread of a contagious disease within a population that occurs if a sufficiently high percentage of people are immune to the disease, generally through vaccination, or in this case, education.

Much like the work performed at Cambridge University on inoculating people against fake news (Roozenbeck & Van der Linden 2019), cybersecurity relies on proper training to prevent security breaches. Education and training are proven methods that help inoculate populations from cyber threats delivered through innovative and inclusive, broad programs, two of which are outlined in this paper. At Boise State University, three identified levels reflect general delineations in cyber security including users, operators and full-time cyber security professionals. Another effort to address the diverse needs of the cybersecurity workforce from the Purdue School of Engineering and Technology at Indianapolis, IUPUI, offers courses, a certificate, minors, and concentrations within a degree program to tackle broader demands.

3. Cybersecurity Course Offerings

Education as commonly defined, is the pursuit of "knowledge, skills, beliefs, and habits." [1] Education is a life-long pursuit, and develops skills in critical and strategic thinking, problem solving, and research. Training is singularly focused on a single objective. Upon completion, a specific task or skill is mastered. The learner need not understand the theory behind the skill as long as the student demonstrates task mastery. Education can include training as reinforcement after studying the critical thinking and other components of life-long learning, but the main focus of education is to prepare the learner to adapt to the ever-changing cybersecurity landscape.

The higher educational system provides mechanisms for education and training. Community colleges (CC) or two-year colleges serve multiple purposes, preparing learners for entry into four-year colleges or universities while also offering applied or certificate-based programs that are more task based or skill specific.

Four-year colleges and universities responsibilities are to educate learners, offering bachelor, master and doctoral degrees in support of lifelong learning, research, and critical thinking. The primary goal of higher education is to create a well-rounded learner, proficient in the chosen major, adaptable to new technologies while avoiding the technology specific traps that can come at the expense of rapid technology changes and adoption. Learners who understand the underlying foundation and theory and can think abstractly, and in research terms, will be able to lead in future new solutions.

3.1 Boise State University (BSU)

Boise State offers several certificate programs along with the degree programs, attempting to address the diverse needs of the future cybersecurity workforce (Boise State Catalog, 2020). Boise State observes the needs of three student types in the cybersecurity education; users, operators and creators. These levels are presented in no particular importance and encompass necessary components for student preparation, production and fortification in cybersecurity and the IoT.

3.1.1 Level 1: Cybersecurity for All

A set of courses providing a cyber security knowledge for everyone. The objective is to raise the cybersecurity awareness across the campus. The Cybersecurity for All certification consists of four-courses offering awareness training covering cyber-physical systems (CPS), Internet of Things (IoT), home office security and incident response recovery. These four courses provide a basic understanding of cybersecurity hygiene that is necessary for any Internet user.

3.1.2 Level 2: Cyber Operations

The objective of this operationally focused certificate program is to prepare students by providing industry certification opportunities for those interested in cybersecurity as a career. Students are trained to think critically and develop strong problem-solving skills that can be applied to understand and solve data-driven cybersecurity problems. This certification requires the student to demonstrate proficiency in four courses that cover information assurance with critical thinking, cybersecurity foundations, cyber operations and cyber forensics with recovery. In addition to certification training, this course offering covers gaps not addressed in many industry certifications thereby developing students capable of solving new problems that appear in the cyber domain as well as being able to translate those skills into other domains. This certification consists of twelve credit hours and includes: information assurance, applied critical thinking, offensive security, defensive security, recovery and forensics while requiring one industry security certification. The program includes a multidisciplinary group exercise designed to apply the lessons learned in the courses to a real-world setting requiring both technical and interpersonal skills (McChrystal 2015, Lencioni 2002, Lencioni 2016).

3.1.3 Level 3: Cyber-Informed Engineering

A set of courses that provide cyber-informed engineering skill sets with the objective is to educate the future engineers and scientists to incorporate cybersecurity (as a requirement) into the design process for more resilient systems. In order to protect against cyber threats, the course developer(s) require awareness of the system being designed (software, hardware, firmware, control systems etc.) in addition to understanding different ways in which an adversary can compromise the system through denial, deception, disruption, destruction and deterrence. The courses teach countering compromise through robustness, resilience and other methods to design security controls into a system that meet performance criteria and enhance protection, hence the broad interdisciplinary knowledge requirement. This program was introduced in August 2018 with an introduction course in cybersecurity. Other cyber contents are incorporated into existing STEM courses.

3.2 Purdue School of Engineering and Technology Indiana University-Purdue University Indianapolis (IUPUI)

The Purdue School of Engineering and Technology, IUPUI, offers courses, a certificate, and minors, and concentrations within a degree program that attempt to address the diverse needs of the cybersecurity workforce (IUPUI Catalog 2020).

3.2.1 Level 1: Information Security Specialty

A minor that provides a cybersecurity background for everyone. The objective is to raise awareness of cybersecurity across different disciplines throughout campus. The minor consists of four courses that offer base knowledge in computer architecture and data communications. Additionally, the minor consists of a cybersecurity awareness foundational course that teaches the novice about threats and vulnerabilities on the Internet; how to protect home systems, small business systems, IoT devices, and what to do to recover from an incident. The end user is the most vulnerable asset to the cybersecurity ecosystem and providing an education to all users, regardless of discipline, will help to protect our cyber domain. The courses are available to the entire campus community for educational purposes with aid in the overall education of the larger community.

3.2.2 Level 2:

The Information security concentration in Purdue School of Engineering and Technology, IUPUI, Computer Information Technology (CIT) program focuses on preparing students with a solid foundation in information assurance and network security. These students gain skills in practicing information assurance data management security by understanding networking and concepts on protecting systems connected to the internet. Below are a few courses available for the concentration:

- The Programming for NetSec class prepares students though "hands on" experience with scripting and programming within the network security realm. These skills are frequently used in future courses or professional career when task automation or script writing can be used to reduce repetitive tasks or job functions that exists as a part of their job duties.
- The Networking Operating System Administration course is designed to engage students to develop a deeper understanding of operating systems and how networked servers' function, as well as introducing the concepts of virtualized environments compared to minimal installs on physical hardware.
- The Advanced Network Security course introduces additional disciplines by analyzing the mentality of attackers and reviewing the Attacker Methodology/Cyber Kill Chain. These viewpoints provide keen insights as to why security is a crucial step in networked systems.
- Digital Forensics course demonstrates the needed skills for a digital forensics career and requisite proficiencies to present evidence that pass legal proceedings standards.
- The IT Risk Assessment course provides an overview of a typical IT risk assessment and audit while teaching students how to give recommendations and reports for IT systems designed to help ensure a safer and more secure environment.

3.2.3 Level 3:

IUPUI's new cybersecurity degree builds off the Information Security concentration and offers the students more involvement in the offensive and defensive aspects of cybersecurity, applying their knowledge to learn more in-depth attacking and defending skills and processes. The Offensive Cybersecurity course enables students to practice and demonstrate concepts in an actual, connected, systems environment. The coursework includes a competition with a complete report on findings including what additional measures might be implemented to help mitigate future incidents, exposures or vulnerabilities.

In addition to the aforementioned courses, development plans exist for a security architecture course with different areas of concentration, including: Industrial Controls Cybersecurity, Cybersecurity Risk Assessment, Defensive Security, and Digital Forensics. These areas of focus would enable a deeper concentration to obtain in-depth knowledge for each related career path, allowing students to tailor their interests to a specific job role or research area post-graduation.

Students must be able to integrate cybersecurity learning with other disciplines. Experimental integrated learning was implemented in two undergraduate STEM courses through cross-curricular instructional units. The courses, Big Data Analytics and IT Risk Assessment, have content connections and were taught in parallel. The goals of this approach were to investigate the effectiveness of cross-curricular integrated instructional units for courses that are content connected and study the effectiveness of these projects based on inquiry levels. These units were integrated to enhance students' capabilities in connecting and transferring knowledge from one setting to another, improve students' understanding of the subjects in the courses and interpreting the larger impacts of information technology. We hypothesize that the hands-on projects will enhance students' interdisciplinary problem-solving skills by connecting knowledge, as well as research experiences and capabilities.

4. Observations on course execution

Higher education has the "build-it they will come" mentality in some cases and in other cases a hyper-response to industry at the expense of vision. Although both problems are very different the end result is oftentimes the same, a failure to graduate intellectually adaptable students capable of solving current and future cybersecurity problems. Students who have gone through a degree program have neither the thinking skills nor job skills. Student who complete training certifications have specific skills that do not always adapt to novel problems.

One way to solve this problem is collaboration between security industry practitioners and universities. The universities need to work with industry to develop the program and curriculum. Academia offers abstract concepts that encompasses larger problems while industry practitioners (the subject matter experts) provide knowledge of real-world instantiations of those larger concepts. An additional benefit in this approach is that often times industry experiences disruption through technology changes and novel attacks first, this approach allows for greater sharing.

Delivery modes need reviewed. Content consumption in society has changed, academia has not adjusted. The face-to-face delivery as the primary method of teaching should be re-evaluated. Pedagogy favours face-to-face interactions, but andragogy favours "hands-on" exercises. Undergraduate students are young adults and may benefit from the introduction of andragogy techniques mixed with traditional pedagogy. Asynchronous concepts taught online and reinforced with applied laboratory exercises might produce better outcomes. Short-term bootcamp style learning can act as larger integrated lab exercises.

CIT in the Purdue School of Engineering and Technology, IUPUI has a program that combines theory and practice. The Living Lab (LL) allows students to apply networking, cybersecurity, database, website and application development concepts and techniques learned from prior CIT courses to service, internal and/or external projects. The Living Lab emulates an industry IT department in which students work on one or more projects as part of an IT team spending 200 working hours per semester equalling internship credit. The Living Lab provides students an experiential and service IT learning environment. By presenting students with a real production IT environment, enabling the students upon graduation to contribute immediately and directly to their employment workplace. Students solve real world IT and cybersecurity problems with little or no supervision while meeting business standards of managing, documenting and reporting their work via weekly status reports.

LL Statistics showed that from 2008-2020, 569 projects launched, thus, allowing 518 students to gain valuable real-world operational experience. "Hands-on" experience was obtained in the LL that could not otherwise be obtained in business due to risk concerns. The LL provides an example of industry partnership and academia working together to provide reinforced learning of broad academic concepts through real life examples. The LL allows for operational environment immersion, along with cyber-physical security workshops and training in support of IoT security at water/wastewater treatment facilities, and traditional risk assessments

At Boise State University, the Cybersecurity for All, and Cyber Operations, programs will go live in August 2020. The Cyber Informed Engineering program courses have been offered twice in the last two years reaching approximately 65 students representing a wide range of students, motivations and skillsets. The courses have been well-received, and most students have performed well, remarking on an appreciation for the course content and career field opportunity diversity.

However, a recent exercise revealed problems reflecting a lack of self-regulation, constrained mindsets and limited interpersonal skills. A penetration testing hands-on activity, using an air-gapped learning network for practicing lessons in a heterogeneous network consisting of wired and wireless components. Activities include WiFi penetration, network scanning and password cracking. Students were to leave a mark on the Raspberry Pi by using the default userid and password then access to the shadow password file. Leaving this network running for days allowed students to apply classroom lessons learned. However, someone changed the default user password and root password of the Raspberry Pi keeping other students from finishing the exercise, thus providing the educators a learning opportunity. One problem observed was the lack of understanding operating systems such as Linux makes the coursework challenging. For this reason, foundational courses that can be viewed as "how to" classes deserve consideration for cyber initiatives; however, providing the right amount of challenge and support on relevant learning tasks requires further investigation.

Cybersecurity observations from CIT@IUPUI lead to significant program changes. The Information Security and the Network Security Certificate has been taught since 2004., Echoing Boise State, IUPUI also noticed that the students' lack of Linux experience was a detriment in the learning process, particularly in advanced cybersecurity courses such as Advanced Network Security, Digital Forensics, and Wireless Security. In response to this problem IUPUI added Linux modules to earlier 100 and 200 level networking and security courses. Additionally, a 300-

level scripting/programming course using Python and Linux was also added. In all network, security and programming courses instructors are encouraged to standardize on the Linux operating system.

Recognizing other disciplines are becoming increasingly interweaved within cybersecurity, was an additional observation illuminating the need for the CIT program to update our curriculum (Sample & Justice 2018). The CIT program is working with other disciplines on campus to create certificates that will include other aspects of cybersecurity such as law, psychology, sociology, data science, international studies and other areas of study (Sample & Justice 2018). Of particular note, the inclusion of data science may force a marriage of disciplines since the best insights result from subject informed queries.

5. Conclusions

By examining two institutions of higher learning commonalities, differences were also discovered. Both programs continue to evolve in offering interdisciplinary courses that can more effectively address threat analysis and IoT in various environments. Both programs have maintained traditional cybersecurity courses that can be found at most universities and both observed the need for fundamental skills. Both programs offer cybersecurity for non-majors, that teaches basic cyber hygiene and may offer an opening for non-traditional students to ultimately pursue cybersecurity while bringing fresh perspectives that Borg (2018) deems necessary. This outreach can bring a multidisciplinary approach to cybersecurity and open the aperture to a diverse pool of talent, creative problem solvers and critical thinkers that will be an essential part of the solution

Cyber threats will continue to grow in creativity and sophistication as will the demand for a skilled cyber informed workforce capable of protecting and preventing such attacks. Furthermore, the introduction of artificial intelligence promises to disrupt the workplace. Without changes to university curriculum there is no viable way to sustainably address the workforce needs of today or the future. Significant intervention is necessary to build future talent pipelines, expand current capacity and enhance workforce development efforts with skilled professionals who are highly adaptive and capable of solving a wide variety of problems.

References

Blazer, C. 2008. Literature Review: Educational Technology. Research Services, Miami-Dade County Public Schools.

Boise State University Catalog [Online]. Available: https://www.boisestate.edu/registrar-catalog/ [Accessed].

Boise State University Course Catalog [Online]. Available: https://majors.boisestate.edu [Accessed].

Borg, S. 2018. Seven Overlapping Theses On Cyber-Security Education [Online]. Available: https://www.cerias.purdue.edu/nace/papers/Borg.pdf [Accessed].

Brown D. 2019. Your smartphone is 7 times dirtier than your toilet. Here's how to clean it. [online] USA TODAY. Available at: https://www.usatoday.com/story/tech/2019/02/26/your-smartphone-screen-probably-disgusting-heres-how-clean/2950106002/

Carey, M.J. and Jin, J., 2019. Tribe of Hackers: Cybersecurity Advice from the Best Hackers in the World. John Wiley & Sons.

Cary, M.J., and Jin, J., 2019. *Tribe of Hackers Red Team: Tribal Knowledge from the Best in Offensive Cybersecurity,* Wiley.

Corera, G., 2016. *Cyberspies: The Secret History of Surveillance, Hacking, and Digital Espionage*, Pegasus Books.

Cybercrime Magazine. (2019). Cybersecurity Jobs Report 2018-2021. [online] Available at: https://cybersecurityventures.com/jobs/, Visited: March 9, 2020.

Henry, A.P., 2017. Mastering the cyber security skills crisis: realigning educational outcomes to industry requirements (Vol. 4). ACCS Discussion paper Available at: https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/sites/accs/files/uploads/ACCS-Discussion-Paper-4-Web.pdf

Indiana University – Purdue University Indiana (IUPUI) Available: https://et.iupui.edu/departments/cigt/programs/cit/undergrad/bscit/plan

Isaak, J. and Hanna, M.J., 2018. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, *51*(8), pp.56-59.

Isc2.org. (2019). 2019 Cybersecurity Workforce Study. [online] Available at: https://www.isc2.org/Research/Workforce-Study

LeClair, J., et al. (2013). An Interdisciplinary Approach to Educating an Effective Cyber Security Workforce. Proceedings of the 2013 on InfoSecCD'13: Information Security Curriculum Development Conference, ACM.

Lencioni, P. (2002). The Five Dysfunctions of a Team: A Leadership Fable, Jossey-Bass.

Lencioni, P. (2016). The Ideal Team Player: How to Recognize and Cultivate the Three Essential Virtues, Jossey-Bass.

Lorenz, T., (2020).” ‘Zoombombing’: when video conferences go wrong”, *The New York Times.* [online] Available at: https://www.nytimes.com/2020/03/20/style/zoombombing-zoom-trolling.html

McChrystal, S., Collins, T., Silverman, D., Fussell, C., (2015). Team of Teams: New Rules of Engagement for a Complex World, Portfolio.

McKinsey & Company. (2019). Growing Opportunities in the Internet of Things. [online] Available at: https://www.mckinsey.com/industries/private-equity-and-principal-investors/our-insights/growing-opportunities-in-the-internet-of-things, Visited: March 9, 2020.

National Institute of Standards (2018). "New data show demand for cybersecurity professionals accelerating", Available at: https://www.nist.gov/news-events/news/2018/11/new-data-show-demand-cybersecurity-professionals-accelerating

Nisbett, R. 2004. *The Geography of Thought: How Asians and Westerners Think Differently and Why. Simon and Schuster*.

Roozenbeek, J. and Van Der Linden, S., 2019. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570-580.

Sample, C. and Justice, C., 2018. Suggestions for Addressing the Changing Needs of the Cyber Security Workforce. [online] Available at: https://www.cerias.purdue.edu/nace/papers/Sample.pdf

Segal, A., 2015. The Hacked World Order: how nations fight, trade, maneuver, and manipulate in the digital age., New York Public Affairs.

Staff, W. (2018). How Amazon Rebuilt Itself Around Artificial Intelligence. [online] WIRED. Available at: https://www.wired.com/story/amazon-artificial-intelligence-flywheel/

Trend Micro Report. (2019). Into the battlefield: A security guide to IoT botnets. Available at: https://www.trendmicro.com/vinfo/us/security/news/internet-of-things/into-the-battlefield-a-security-guide-to-iot-botnets

Zhang, C., Zhou, S. and Chain, B.M., 2015. Hybrid Epidemics—A Case Study on Computer Worm Conficker. *PloS one*, *10*(5).

Zhang, C., Zhou, S., Cox, I.J. and Chain, B.M., 2015. Optimizing Hybrid Spreading in Metapopulations. Scientific reports.

Towards a Methodology for the Classification of IoT Devices

Dimitrios Sarris[1], Konstantinos Xynos[2], Huw Read[2, 3] and Iain Sutherland[2]
[1]Digital14, Dubai, UAE
[2]Noroff University College, 4608 Kristiansand S, Vest-Agder, Norway
[3]Norwich University, Northfield, Vermont, USA
dimitrios.sarris@digital14.com
kxynos@mycenx.com
hread@norwich.edu
iain.sutherland@noroff.no

DOI: 10.34190/EWS.20.013

Abstract: Within the scope of the Internet of Things (IoT), there are multiple sensors connected to the Internet that have the ability to report on the status of a system; a state change, the surrounding environment and conduct automated functions based on sensory input. IoT devices commonly have a small form factor, with specific forms of connectivity and power requirements. The business drive to get a product to market and to set the dominant standard, can result in limited governance, as standards appear to be mainly focused on the communication and interactivity aspects of the technology, and less on the overall security or identification of the device, data transmission or quality of firmware.This paper explores possible foundations for the classification of IoT devices according to environmental usage, as a means to facilitate the required governance through regulation and standardisation, which in turn would enable for easier auditing when such devices are deployed, maintained and, finally, decommissioned. The authors recommend the development of a classification framework that would allow an end user to understand if an end product can be used within their required use case/domain.

Keywords*:* Internet of Things; IoT; classification; connected living; framework;

1. Introduction

The increasing use of embedded or smart technologies has been facilitated by a variety of factors, including evolving network connectivity enabling the exchange of disparate forms of information, and the increase in capability and decreasing cost of integrated components leading to the uptake of "smart" devices in both commercial and domestic environments. This offers the possibility of efficient savings in energy and has become one of the forces behind wider adoption of IoT. Security is crucial in the sustainability of new systems and networks. Lack of governance and regulations on the inception, architecture, production and use of the IoT and the Internet of Everything (IoE) has resulted in an environment where malicious actors can influence online systems including financial transactions or affect a victim's physical assets. The Mirai botnet, which affected many IoT devices "took the Internet by storm" (Antonakakis, 2017) when it overwhelmed targets with massively distributed denial-of-service-attacks (DDoS) in 2016. The domain name system (DNS) organisation DYN received a throughput reportedly up to 1.2Tbps, knocking high-profile websites offline on the East Coast of the United States including Netflix, Twitter, CNN and others (Guardian, 2016). Mirai targeted IoT devices, DVRs, cameras, routers, and printers, but the devices used were created by a handful of manufacturers (Antonakakis, 2017) with hard-coded passwords.

In the USA the term Cyber Physical Systems is also widely used alongside IoT, the NIST report by Greer et al. (2019) provides clarity by identifying the history of these neologisms and goes as far as describing a unified perspective to clarify the relationship between the two. These risks are increasingly being highlighted in academic journals (e.g. Awasthi et al. (2018), Ashrad et al. (2020) and Fafoutis et al.(2017)) and in the popular press via news articles (Turner (2017), Sadam et al. (2017)) with regards to the impact of cyber-attacks on an organisation's finances.

Research by Gartner (2018) shows a trend in increasing investment in relation to the security of IoT (Table 1). However, as organizations do not have full control over the software and hardware being used in IoT devices (Gartner (2018)), they state, "*We expect to see demand for tools and services aimed at improving discovery and asset management, software and hardware security assessment, and penetration testing*". Therefore the trend of increased spending in IoT security appears to be concentrated in the *consumer* market (i.e. organisations purchasing devices and securing them from harm) rather than in the *producer* market (i.e. manufacturers strengthening device security before they appear on the market), alluding to the "security as an afterthought" philosophy.

Table 1: Worldwide IoT Security Spending Forecast (Millions) (Gartner, 2018)

	2016	**2017**	**2018**	**2019**	**2020**	**2021**
Endpoint Security	240	302	373	459	541	631
Gateway Security	102	138	186	251	327	415
Professional Services	570	734	946	1,221	1,589	2,071
Total	**912**	**1,174**	**1,506**	**1,931**	**2,457**	**3,118**

The remainder of this paper is organised as follows; *2. Defining the Internet of Things* examines key literary sources to identify problem domains; *3. Related Work* discusses recent research around the area of defining and classifying IoT from the context of Cybersecurity; *4. Toward an IoT Classification Framework* details the perceived high level requirements and the proposed classification types for IoT devices; 5. The section on *Example Scenarios* provides instances of how the proposed IoT classification framework may be applied. Section 6 concludes the paper with a summary and proposed future work.

2. Defining the Internet of Things

The term IoT has changed significantly sine 1999 (Ashton, 2009) to incorporate an expanding range of devices, interactions, connectivity and new efficiencies and capabilities for existing equipment and environments. Research into publication trends conducted by Greer et al. (2019) into the terminology identified 3 temporal phases; low numbers and low growth in 2005-2009, increased IoT publications in 2010-2013, rapid growth after 2014. Their research into IoT definitions largely revealed terminology from the networking and IT communities.

451 Research (2015) defines IoT as "*... the apps and systems enabling the virtualization of the physical world*". This is a broad definition and could be argued that it encompasses a number of devices which would generally be considered to the part of IoT. The definition is broad enough to even include VR (Virtual Reality. Dedlic (2016) considers a similarly broad definition but focuses on scale and complexity with the following definition: "*At the very high level of abstraction, the Internet of Things (IoT) can be modelled as the hyper-scale, hyper-complex cyber-physical system*." Note the synonymous use of "*cyber-physical system"*; the work by Greer et al. (2019) conversely sought to define cyber-physical systems and IoT as two distinct entities.

A more environmental approach is taken by Yachir et al (2016) with a focus on value-added services as: "*The Internet of Things (IoT) envisions a world where smart objects connected to the Internet, share their data, exchange their services and cooperate together to provide value-added services that none of these objects could provide individually.*" The concept of individual IoT devices not being of much use on their own, but rather when working together, help create a distinction from traditional IT equipment. Gartner (2020) defines IoT as "t*he network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment.*"

It can be seen from the various definitions that the term IoT covers a broad range of devices used in numerous environments. A single connected sensor, to more complex devices that provide enhanced control or information by being connected to such devices. A single IoT device tends to perform a limited set of activities. Once they are connected and work as a collective their true flexibility is shown. These communication abilities further the complexity that follows such interconnected systems and increasing possible risk factors. Although an IoT appliance can be thought of as a single device, it is suggested that an 'IoT solution' is defined as the collection of IoT devices that are part of a deployment and that a deployment can include one or more IoT devices. This definition is proposed as a means to identify more complex systems that are made up of a number of IoT components. The reason behind this is the need to know when small systems are introduced into an environment and when they need updating or replacing. Knowing when a single IoT device within an IoT solution has an issue or a weakness (e.g. such as publicised common vulnerabilities and exposures (CVE), (MITRE, 2020)), to coin the old adage, 'a chain is only as strong as its weakest link', decisive action must be carried out to update, replace, or mitigate the Cyber vulnerability to strengthen the environment.

The growth in the use of IoT presents a number of challenges as the variety of devices supporting data processing and storage continue to evolve. Not limited to these factors, IoT and other technologies such as cloud computing allow for a wider adoption since devices may not need to process information, instead merely providing notification of state changes to remote servers.

These technologies range from fixed systems in smart homes to mobile and highly portable technology. Examples include wearable devices and clothing (Hexoskin, 2016), drinking containers (Disney Food Blog, 2016), (Glassify, 2016) and smart housing systems, including; automation (IFTTT, 2020), security systems (Verisure) and infotainment systems (Amazon, 2016). There are a number of possible ways to view IoT devices based on both capability and functionality.

Furthermore, there are several options available to manufacturers when deciding how the devices will communicate. These can include any one or a combination of the following: Ethernet, Wi-Fi, Bluetooth, Long-Term Evolution (LTE), GSM, Zigbee, Z-Wave, Ultra-Wideband (UWB), LoRa, CAN bus, and others. In particular, Zigbee and Z-Wave, have become prevalent in-home automation, both being tailored for typical IoT use cases; low power consumption, intermittent data transfer and wireless mesh networking.

3. Related Work

The huge increase in connected devices and spending on IoT validates the expected wide adoption of IoT technology (Gartner, 2018). Security researchers are increasingly concerned regarding the possible misuses that will occur. The security of IoT devices and systems need to be incorporated into any device from the design phase, up to its integration and life span, including when it gets decommissioned. The principles of Confidentiality, Integrity, Availability, and Non-repudiation are also of the utmost importance in the wider ecosystem of devices to avoid introducing further vulnerabilities into an environment. This is particularly important when maintaining a system with updates and if needed secure decommissioning when the device reaches the end of life.

The National Institute of Standards and Technology (NIST) provides considerations for managing IoT Cybersecurity and privacy risks (NIST, 2019). In particular, its focus is to provide support to federal agencies and other organisations in the USA to better understand and manage the risks throughout a devices' lifecycle. Three high-level issues that affect the management of cybersecurity of IoT devices (compared to traditional IT equipment) were identified;

1. IoT devices interact with the physical realm in ways traditional IT does not,
2. IoT equipment cannot be "*accessed, managed, or monitored*" using the same techniques as traditional IT, and
3. IoT devices' cybersecurity capabilities differ in their availability, efficiency and effectiveness when contrasted with traditional IT (NIST, 2019).

Furthermore, three mitigation goals of protecting device security, data security, and individuals' privacy are identified as the most significant risks an organisation will face when implementing IoT infrastructure.

More recently, NIST released documentation for IoT device manufacturers, "*foundational activities and core device cybersecurity capability baseline"* (NIST, 2020). This guidance (in its second draft at the time of writing) provides voluntary recommendations to manufacturers that they should consider performing before IoT devices reach the market. Six activities are identified, which all stem from identifying the product's customers and defining use cases early in development to determine which cybersecurity capabilities should be implemented. It does express the need of having unique identifiers, both physical and logical, for IoT devices. However, a general classification of such devices is outside the scope of their work and is left to the manufacturer to determine the most appropriate course of action.

In a similar vein, the European Telecommunications Standards Institute (ETSI) have released a technical specification, "*Cyber Security for Consumer Internet of Things".* It discusses a set of 13 cybersecurity provisions for implementation by device manufacturers; adherence providing assurance that companies "*can help in ensuring that [devices and services] are compliant with the General Data Protection Regulation*" (ETSI, 2019). Like NIST (2020), it also does not attempt to provide a classification or taxonomy of IoT devices.

It should be emphasised that both NIST (2019) and ETSI (2019) are voluntary implementations; the additional burden placed on an organisation (be it financial, delayed time to market, etc.) provide disincentives to adoption. However, the strength of the European General Data Protection Regulation (GDPR, 2016) will provide support for the uptake as, a) there are existing legal requirements, and b) the ETSI guidance is domain-specific within the realm of IoT. The United States, without any formal Federal equivalent of GDPR, delegates such decision-making to individual states. However, recent Californian legislature introduced in 2018 and coming into effect on January 1st 2020 (SB-327, 2018) which provides requests similar security mechanisms and controls as ETSI (2019) became the first "IoT law" in the USA related to strengthening Cyber on such devices. Given the strength of California as a state (e.g. having the highest GDP of all 50 states), one can hope that when vendors comply with local Californian laws, they will provide the same goods and services throughout the United States. It doesn't provide a classification or taxonomy of IoT devices, but relies on a more generic term of "connected device", which is defined as "*...any device, or other physical object that is capable of connecting to the Internet, directly or indirectly, and that is assigned an Internet Protocol address or Bluetooth address*" (SB-327, 2018).

ENISA (2017) have also provided a handbook on baseline security recommendations for IoT. It goes into detail starting from defining an IoT device, its communication, to threat analysis and attack scenarios. Our proposed classification framework would look at supplementing this very technical documentation but, simultaneously, try to avoid providing the end user with too much technical information. The handbook is oriented more towards the manufacturers and testers of of IoT devices. It would provide a good testing basis of identifiable issues within an IoT device.

Other frameworks have been proposed such as the IoT Security Compliance Framework (IoT Security Foundation, 2018). But these are very technical in nature, focussed mainly on providing guidance for managers developing products or services and on those developing IoT devices. Although a highly valuable resource, a simpler and more accessible domain-based system is worth exploring as a user focussed solution to IoT issues and how they evolve over time.

4. Toward a Classification of IoT Devices

The classification of IoT needs to include several points that provides the user assurance of the end-product. It should make it clear to the end user of what requirements have been met and if the item is fit for the intended purpose. The intention is to highlight the suitability of the IoT device for a domain, rather than to provide another classification system mapping to standards, regulation and/or code of practice (DCMS, 2018).

This ensures a level of trust can be associated with the device and in-line with the classification requirements. The classification will have to clearly define if a device has been built for one of the following domains: government, military/defence, industrial, healthcare, automotive, household, or open standards; it would be marked as such. The classification would also include no more than one subdomain, that would bring some specialisation to the top-level domain. Future work will look at setting out what requirements would need to be defined for each classification domain and sub-domain.

Classification domains of IoT devices would include details includin; how the device is constructed (quality of components, original components, etc.), operations, communication of security/bug issues to end user, and whether the firmware/software be maintained regularly. Classification would, initially, be up to the manufacturer of appointing a classification domain for each device created, in principle at the time of manufacturing, similar to the concept of allocating a unique MAC address for any network card. The classification process should be documented at all times.

Regulations, if created and wherever required, will mandate the "use by" and "use of" IoT, where classification is the means to identify at a high level the operating domain of a device, such that enforcement of the appropriate security requirements is possible in order to comply.

'Government' and 'Military/Defence' will include standards that need to be clearly defined, if not already, by local and/or international governments. This will depend on who is using the IoT device and its context. Devices can be classified differently by government bodies depending on the use case. Ideally a central government body should focus on governmental classification or reclassification. IoT devices classified under this category should, in principle, include the highest level of security baseline and trust, by utilizing current or future digital and physical non-repudiation technologies.

'Industrial' would include items that are specified for a wide area of industries. Industrial requirements would focus on the general needs of a manufacturing environment without many specialised requirements. Any special requirements could be written up within a subdomain. The authors appreciate there are a wide variety of industries with unique requirements and standards. Therefore it is suggested that more detailed information be included in a subdomain, like 'Industrial:Oil and Gas' or 'Industrial:Manufacturing' etc. Although manufacturers would be welcome to include the generalisation of the term showing that they have avoided including any special cases that would be included within a subdomain. Industrial features for consideration in IoT would normally include minimal downtime and the need to operate within inhospitable environments e.g. high temperature, humidity, high levels of particulate matter, electrical noise, or magnetic fields.

'Healthcare' IoT is one of the most sensitive levels of classification, especially when utilised for internal / critical applications. A high level of security baseline is expected on its creation and usage. IoT classified as 'Healthcare' may include: in-body solutions for patients, apparatus used in the diagnosis, treatment and monitoring of patients. The standards that define the 'Healthcare' IoT elements, would ideally be unanimously defined and have global reach, by organizations similar to World Health Organization (WHO), to enable strict governance and control, and minimize the potential of human casualties.

'Household' would include devices that ensure confidentiality is upheld at all times and provides confidence to the consumer when installing IoT device. This should also follow a standard baseline of security that is required and expected by a consumer device installed within a living environment.

'Automotive' IoT devices will have to reflect the strong cooperation of the automotive industry that will define their stringent requirements. This is irrespective of whether the items have been built internally or by their external automotive partners. A clear focus will be to ensure vehicle safety.

Lastly, 'Open' will determine that the device has been developed without a specific domain in mind. Although there might be a strong case to classify everything as 'open', manufacturers should take care as future regulation could disallow the usage of such devices in certain environments resulting in a limited market.

As can be seen in Figure 1, an IoT device or solution will go through a classification process. The item should be classified during manufacture. This clearly states what domain the manufacturer was envisaging. The classification can change during deployment, based on the characteristics of the device, how it is maintained, and what security or risks that may arise over the course of its lifetime. Any reclassification of the product will happen as a consequence of the manufacturer not adhering to the specified classification requirements and standards will make the product unfit for its intended classification. The end user could make an informed decision to decide to continue/discontinue with the product's placement/usage etc. In extreme cases regulation could even dictate a certain classification for an IoT device within a particular operating domain.

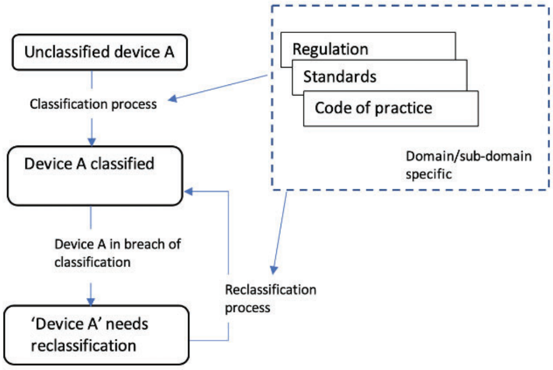


Figure 1: State diagram of proposed classification process

An open page or resource (e..g, public/private wiki) needs to be maintained and must be publicly accessible. It will list the IoT devices and/or IoT solutions and the current/historical classification of a device. It may also include an optional field detailing the reasoning for the classification, or re-classification.

In the future, a non-profit transparent committee could be setup (similar to a Working Group, IEEE, etc.) that would test and validate claims made by the manufacturers and set the date for retesting of devices or retest based on public evidence and reporting as such. This would allow for the devices to be observed and located in the event of any major issues that may be found with an IoT device.

The classification of solutions should automatically be set according to the most stringent classification of the IoT device(s) included in it, to eliminate the possibility of classification breaches and unauthorized access/use of such solutions.

5. Example Scenarios

Two example scenarios are presented below to demonstrate how to apply the proposed classification framework and illustrate the current challenges.

5.1 Scenario 1 - Embedded software issues

An organization makes use of sensors to gather room temperatures, using an inexpensive WiFi module widely found online. There is no known and non-intrusive way to check that the firmware does not have any malicious code loaded into it (e.g., manufacturing or delivery). It is not possible to check whether the firmware has been altered from the original one installed by the manufacturer. Firmware alterations could occur at the factory, in transit, or even after the module has been installed. Therefore, this device cannot be used within all the possible environments that might require the device to operate without any interruptions that might be caused by malicious code.

It may be argued that a temperature sensor presents a limited threat, except when the sensor is relied upon for monitoring critical systems and a chain reaction may be created. If the readings of the temperature sensor are used to regulate the temperature of an air-conditioned room containing heat sensitive servers, this could cause short-term or long-term issues. The incorrect readings could cause the air-conditioning system to not appropriately cool the servers and cause them to fail after incorrectly shutting down due to overheating.

If the item was classified as 'Manufacturing' then it would not be fit for purpose, since it's accuracy and reliability is not up to standard for that classification, as seen in Figure 2(a). Therefore, it would be re-classified as 'Household' or 'Open' depending on requirements. In the example presented in figure 2(a) the device is reclassified as 'Open'. With the inclusion of detailed documentation, when the reclassification occurs, it is possible to have the item under a less stringent classification. The consumer would still be able to understand why the item changed its classification and decide if it fits the use case.

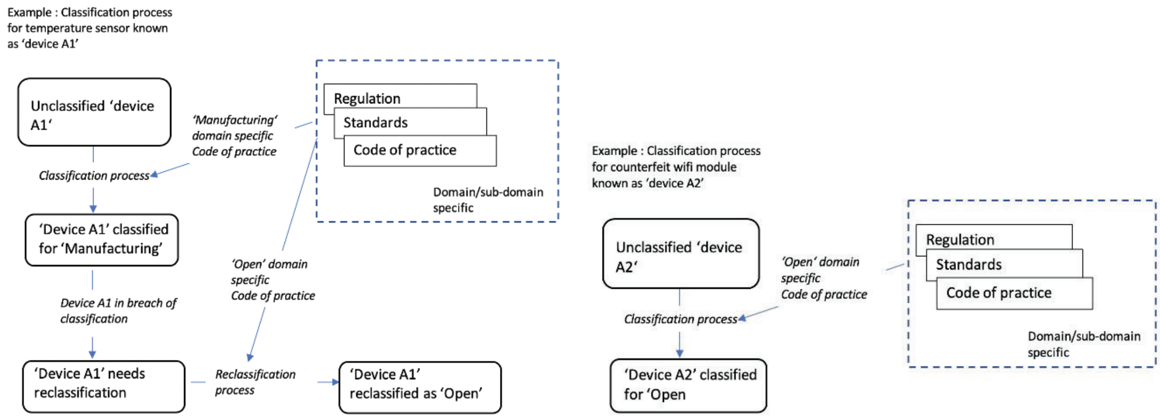

Example : Classification process for temperature sensor known as 'device A1'
Unclassified 'device A1'
'Manufacturing' domain specific Code of practice
Regulation
Standards
Code of practice
Domain/sub-domain specific
Classification process
'Device A1' classified for 'Manufacturing'
'Open' domain specific Code of practice
Device A1 in breach of classification
'Device A1' needs reclassification
Reclassification process
'Device A1' reclassified as 'Open'
Example : Classification process for counterfeit wifi module known as 'device A2'
Unclassified 'device A2'
'Open' domain specific Code of practice
Regulation
Standards
Code of practice
Domain/sub-domain specific
Classification process
'Device A2' classified for 'Open

Figure 2: a) Scenario 1: classification and reclassification process; b) Scenario 2: classification process

5.2 Scenario 2 - Counterfeit chipsets

In this case we will look at counterfeit chipsets, and one well documented case is that of Nordic's nRF24L01+ (Mahidharia, 2015). These are relatively cheap 2.4GHz modules that are ultra-low power 2 Mbps RF transceivers. Anecdotally, it is known that such devices are quite popular with tinkerers, hobbyists and prototype developers. There are a flood of cheap fakes/counterfeits/clones that can be found on Amazon, Alibaba and other such marketplaces (Tehranipoor et al, 2017). In the case of Nordic's nRF24L01+ counterfeits, they are known to have a different manufacturing technology of 350nm, instead of the 250nm found in the genuine ones. Mahidharia (2015) also reports that due to the technology and larger die, this could mean that the module would have higher power usage and lower sensitivities.

At first glance, these devices pose a very limited threat to the majority of hobbyists. The issue could become more prominent when the solutions are distributed to end users, who are not aware of the quality of hardware and software (e.g., firmware) operating the device.

The case is only made worse when the software and hardware is fully controlled by the vendor increasing the possibility of backdoors existing and/or being activated at will. The undeniable questions about the build quality also raises questions about the possibility of causing other issues like shorter battery life and even the possibility of a fire.

Based on the proposed IoT classification, IoT solutions (the product) including such devices should be classified as 'Open', the classification process can be seen in the state diagram figure 2(b). It would then be reported that the device uses a non-genuine component and one that the final IoT solution has chosen to include, possibly because it is a cheaper component. A wiki entry about the product's classification would be documented and the reasoning for the classification as well.

The consumer would then be able to verify the product (IoT solution) that includes the nRF24L01+ clone and decide on an acceptable level of risk. To determine on how to progress with installation, usage of solution, look at replacing the device or even decommission the device.

6. Conclusions and Future Work

The current state of cybersecurity within the specific domain of IoT has been discussed, with particular emphasis on the need for a formal classification model of devices, primarily based on their use case/domain of operation. The intention is not to produce further codes of practice or additional security criteria, but to develop an applied classification that considers the intended domain for the IoT device.

Two scenarios were presented to highlight the challenges associated with similar IoT devices/different domains, and explanations to support the formal creation of an IoT classification framework were defined.

This preliminary work can be used to move toward a formal classification, complementary to current Cyber frameworks such as NIST, ETSI and ENISA. It will further facilitate securing IoT systems used by organisations and individuals, and providing protection from the deployment of IoT devices.

Future work will include guidelines on the exact classification process and the trigger for reclassification; a process and methodology for the application of the relevant documents (regulations, standards and codes of practice) for each domain. This will detail how the proposed system will interact with the standards existing in these different areas, in particular those where there is rapid growth in the adoption of IoT devices such as industrial systems, automotive environments and home automation.

References

451 Research, (2015), "Explaining the Internet of Things Ecosystem and Taxonomy", [last accessed 20 Jan 2020], [online] :https://451research.com/images/Marketing/IoT/IoT_Taxonomy_12.1.15.pdf

Antonakakis et al, (2017), "Understanding the Mirai Botnet", 26th USENIX Security Symposium, 2017, August 16, 2017

Ashrad, J., Azad, M. A., Abdeltaif, M.M., Salah, K., (2020), "An intrusion detection framework for energy constrained IoT devices", Mechanical Systems and Signal Processing, Vol. 136, Feb 2020.

Ashton, K., (2009), "That Internet of Things thing", RFID Journal, 2009.

Awasthi, A., Read, H. O. L., Xynos, K., Sutherland, I., (2018), "Welcome pwn: Almond smart home hub forensics", In proceedings of the eighteenth annual DFRWS USA, Portland, Oregon, USA. [online]: https://doi.org/10.1016/j.diin.2018.04.014

DCMS (2018) Mapping of IoT Security Recommendations, Guidance and Standards to the UK's Code of Practice for Consumer IoT Security, UK Government, Department for Digital, Culture Media and Sport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/774438/Mapping_of_IoT__Security_Recommendations_Guidance_and_Standards_to_CoP_Oct_2018.pdf

ENISA, (2017), "Baseline Security Recommendations for IoT in the context of Critical Information Infrastructures", Nov 2017, [last accessed 20 Jan 2020], [online]: https://www.enisa.europa.eu/publications/baseline-security-recommendations-for-iot

ETSI, (2019), "CYBER; Cyber Security for Consumer Internet of Things", DTS/CYBER-0039, ETSI, February 2019, [last accessed 20 Jan 2020], [online]: https://www.etsi.org/deliver/etsi_ts/103600_103699/103645/01.01.01_60/ts_103645v010101p.pdf

Fafoutis, X, Marchegiani, L, Papadopoulos, GZ, Piechocki, R, Tryfonas, T & Oikonomou, G, (2017), "Privacy Leakage of Physical Activity Levels in Wireless Embedded Wearable Systems". *IEEE Signal Processing Letters*, vol 24., pp. 136-140

Gartner, (2018), "Gartner Says Worldwide IoT Security Spending Will Reach $1.5 Billion in 2018". [online] : https://www.gartner.com/en/newsroom/press-releases/2018-03-21-gartner-says-worldwide-iot-security-spending-will-reach-1-point-5-billion-in-2018

Gartner, (2020), "Gartner Glossary - Internet of Things (iot)", [last accessed 20 Jan 2020], [online]: https://www.gartner.com/en/information-technology/glossary/internet-of-things

GDPR, (2016), "General Data Protection Regulation", REGULATION (EU) 2016/679, European Parliament, 2016. [online] : https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Greer C., Burns M., Wollman D., and Griffor E., (2019) "Cyber-Physical Systems and Internet of Things", NIST Special Publication 1900-202, March 2019, https://doi.org/10.6028/NIST.SP.1900-202IDC, (2014), "Internet of Things (IoT) Taxonomy Map", [last accessed 20 Jan 2020], [online] : http://www.idc.com/downloads/IoT_Taxonomy_Map_V2_Nov2014.pdf

IFTTT, (2020), "IFTTT for Business", 2020. [last accessed 20 Jan 2020], [online]: https://platform.ifttt.com/

IoT Security Foundation (2018) IoT Security Compliance Framework v2, [last accessed 26 Jan 2020], [online] https://www.iotsecurityfoundation.org/wp-content/uploads/2018/12/IoTSF-IoT-Security-Compliance-Framework-Release-2.0-December-2018.pdf

Kemal A. Delic., (2016), "On Resilience of IoT Systems: The Internet of Things (Ubiquity symposium)". Ubiquity 2016, February, Article 1 (February 2016), 7 pages. http://dx.doi.org/10.1145/2822885

Lu, X., Qu, Z., Li, Q., and Hui, P., (2015), "Privacy information security classification for internet of things based on internet data", International Journal of Distributed Sensor Networks, 11(8). [online] : http://dsn.sagepub.com/content/11/8/932941.full

Mahidharia, A., (2015), "NORDIC NRF24L01+ – REAL VS FAKE", Hackaday, [last accessed 20 Jan 2020], [online] http://hackaday.com/2015/02/23/nordic-nrf24l01-real-vs-fake/

MITRE, 2020, "Common Vulnerabilities and Exposures", MITRE, 2020. https://cve.mitre.org/

Morgan, J, (2014) "A Simple Explanation Of 'The Internet Of Things'", [last accessed 20 Jan 2020], [online] : https://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/

NIST, (2019), "Internet of Things (IoT) Cybersecurity and Privacy Risks", NISTIR 8228, U.S. Department of Commerce, June 2019.

NIST, (2020), "Recommendations for IoT Device Manufacturers: Foundational Activities and Core Device Cybersecurity Capability Baseline", U.S. Department of Commerce, January 2020.

Sadam, R. and Dasgupta, S. (editting), (2017), "Mondelez 2nd-qtr revenue growth hit by global cyber attack", [last accessed 20 Jan 2020], [online]: https://www.reuters.com/article/us-cyber-attack-mondelez-intl-idUSKBN19R33R

SB-327, (2018), "A*n act to add Title 1.81.26 (commencing with Section 1798.91.04) to Part 4 of Division 3 of the Civil Code, relating to information privacy.",* Senate Bill 327, Ch. 886, September 2018. [last accessed 20 Jan 2020], [online]: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB327

Tehranipoor, M., Guin, U. and Bhunia S., (2017) "Invasion of the Hardware Snatchers: Cloned Electronics Pollute the Market", [online]: https://spectrum.ieee.org/computing/hardware/invasion-of-the-hardware-snatchers-cloned-electronics-pollute-the-market

Turner, G, Verbyany, V. , and Kravchenko, S, (2017), "New Cyberattack Goes Global, Hits WPP, Rosneft, Maersk" [last accessed 9 Jan 2020], [online]: https://www.bloomberg.com/news/articles/2017-06-27/ukraine-russia-report-ransomware-computer-virus-attacks

Woolf, N., (2016), "DDoS attack that disrupted internet was largest of its kind, experts say", The Guardian, October 2016, [last accessed 20 Jan 2020], [online]: https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet

Yachir, A., Amirat, Y., Chibani, Y. and Badache, N., (2016), "Event-Aware Framework for Dynamic Services Discovery and Selection in the Context of Ambient Intelligence and Internet of Things," in *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 85-102, Jan. 2016. doi: 10.1109/TASE.2015.2499792

Verisure, (2016), "Home alarm system", [last accessed 20 Jan 2020], [online]: https://www.verisure.no/

Zeptobars, (2015), “Nordic NRF24L01+ - real vs fake : weekend die-shot”, [last accessed 20 Jan 2020], [online]: https://zeptobars.com/en/read/Nordic-NRF24L01P-SI24R1-real-fake-copy

Cyber Security System of FPGA Platform for Wireless Sensor Networks

Gulfarida Tulemissova and Olimzhon Baimuratov
Suleyman Demirel University, Almaty, Kazakhstan
faridaertay@gmail.com
olimzhon.baimuratov@sdu.edu.kz

DOI: 10.34190/EWS.20.051

Abstract: This article hypothesizes that the presence of well-protected memory of sensor networks improves their performance as a multi-agent system. The agent's memory stores information about typical reactions to information signals from receptors, as well as information about the state of effectors and available resources. It also stores programs for processing input information into control signals supplied to effectors and the results of reactions to a particular external situation, the number and variety of knowledge and programs stored in it, the degree of development of the internal model of the external world, and reflection options that determine the complexity and nature of behavior agent. The main components of the memory block are filter systems that provide the most important information for the agent, the internal model of the outside world, and the model of the agent itself.

Purpose: to prove that using the FPGA platform guarantees programmable memory protection to maintain its level of autonomy and intelligence.

The object of study: cybersecurity of sensor networks with protected memory

Approaches: agent theory, machine learning

Objective: To prove that the use of the FPGA platform guarantees programmable memory protection to maintain its level of autonomy and intelligence.

Findings: sensor network security models

Keywords: wireless sensor network (WSN), information security, multi-agent system, machine learning, FPGA bit stream algorithms; clustering intrusion detection methods

1. Introduction

Well-known and frequently used wireless networks have long established themselves as productive technologies for many industries, agriculture, medicine and other business areas. It has become common to use WSN for manufacturing processes of various kinds. There is a lot of research and projects in these sectors, but now only one thing excites minds the minds of project developers - this is information security of networks in all aspects. So according to the data of 2018, in wireless networks are use 22 types of hacker attacks (Ahmed, 2018). The author (Chun-ta Li, 2010) argues that attackers can easily compromise the WSN by using the broadcast medium and the lack of protection against unauthorized access. This is all the more relevant in today's digital environment, cybercrime has become a disaster, for example, every 39 seconds there is a hacker attack, 300,000 new malware is created every day (Galov, 2020). In this connection, sharply there is a question on creation of the existing information security measures for many sectors of the developed industrial society.

Sensor networks themselves, such as sensors, wireless routers, and access points, are also subject to hacker attacks. The main goal in designing the architecture of wireless sensor networks is to protect data and ensure that the basic laws of information are met, namely, the integrity, availability, relevance of data, as well as their confidentiality and authentication. In addition to standard information security tasks, it is necessary to organize protection against many types of hacker attacks. The full list of attacks from the author (Basan et al 2017):

- Traffic analysis and listening to the communication channel by unauthorized persons
- Routing attacks
- Change route information.
- Selective mailing.
- Sinkhole Attack
- Sybil attack.
- Wormhole attack.
- HELLO flood attack
- Denial of service
- Node capture (node subversion)
- Node failure (malfunction)

- Simple node /failure
- Distortion of the message
- About the node
- Copying a network node
- Fake route information
- Selective forwarding
- Emergency attacks
- Wormholes
- HELLO flood attacks.

This is just a list of attacks carried out on WSN. However, for designing is need to take into account the statement of the problem, the conditions used for the functioning of the environment, the mathematical and theoretical apparatus, methods for studying problems and other issues. To study the problems of cybersecurity, we have chosen as the research base - the theory of agents, which is well suited for WSN, since agents inherently represent a flexible system that can receive information from the external environment and, most importantly, includes such properties as the ability to organize itself and forming a response to cyberattacks. Moreover, agents are mobile, which are able to migrate from one host to another (Borselius, 2014). Another significant advantage of using mobile agents is that they can overcome network delays (Chess et al 2005). Moreover, the solution of security problems for mobile agents is applicable to solve the security problems of any type of agent system (Jung et al 2012). Nevertheless, the security of these systems is not always are ensured. Again, the author (Borselius, 2014) concluded that there is no single solution to the security problems presented by mobile agents if there is no reliable equipment.

In recent years, many attempts were been made to identify and solve security problems in MAS. Wang W. et al. (2013) identified the following security vulnerabilities in MAS: malformed name services and mappings, insecure communication channels, insecure delegation of services, and lack of responsibility. The authors (Hedina et al 2015) propose systems for eliminating the aforementioned vulnerabilities, adding trust to MAS. Therefore, in our article we consider FPGA platforms as an option of reliable equipment for the central unit of information processing and storage. The FPGA platform memory filters all incoming information from the outside world and the model of the agent itself.

2. Rationale for using the Multi-Agent Systems

Today, when designing automated systems, it is important that the system react to external influences. Moreover, such a reaction should be instantaneous and must can recognize images that do not correspond to the allowed sets, organize itself and give an appropriate reaction. In this case, the system must be intelligent. These conditions have multi-agent systems that have properties that are very important to us, namely:

1. Activity, ability to organize and implement actions
2. Reactivity, the ability to perceive the state of the environment.
3. Autonomy, relative independence from the environment
4. Sociability arising from the need to solve their problems together with other agents and provide by developed communication protocols.
5. Purposefulness, assuming the availability of own sources of motivation - site https://www.phoenixcontact.com/online/portal/gb] fig. AAA. Moreover, multi-agent systems are artificial intelligence systems.

It is clear that autonomous agents and multi-agent systems represent a relatively new way to analyse, design, and implement complex software systems. Usually, open multi-agent systems were been considered as systems that exchange data via the Internet, allowing anyone to connect to the platform on which the agents work. This means, that in MAS is not global control of the system, and the information as a whole is highly decentralized. Nevertheless, the biggest danger of such open multi-agent systems is that they are accessible via the Internet for any hacker to connect to the platform on which the agents work (Cieszewski et al 2013).

Since the use of mobile agents causes a number of security problems, such systems should have their own special guaranteed security system. The specific attacks as a malicious host can commit were been described in works of authors (Hedina et al 2015), (Huffmire et al. 2008) and can summarized as follows: the particular attacks, that a malicious host were been described in Basan et al 2017) and can summarized as follows:

1. Observation of code, data and flow control,
2. Manipulation of code, data and flow control – including manipulating the route of an agent,
3. Incorrect execution of code – including re-execution,
4. Denial of execution – either in part or whole,
5. Disguise as another host,
6. Eavesdropping on agent communications,
7. Manipulation of agent communications,
8. False system call return values.

3. FPGA as a platform for MAS security

In our article, we focus on touch nodes based on programmable hardware devices FPGA. FPGA (Field Programmable Gate Arrays) is a platform with programmable logic blocks and programmable relationships. Such a platform can be used as a standalone or modular platform, or as a subsystem on a chip (Fresse et al 2014).

As article (Kadam et al 2014) have said, reconfigurable hardware becomes the best option for efficiently implementing complex and computationally expensive signal processing algorithms. Reconfigurable hardware takes advantage of the high computational efficiency of the hardware, as well as the flexibility of software implementation. The introduction of digital signal processing applications in FPGA demonstrates an improved result in terms of speed compared to the software implementation and previously stated architecture. And the algorithms used in signal processing applications require significant computational resources of a programmable field array (FPGA)(Van Court et al 2006).

FPGA-based design creates a special modeling environment and uses three basic design rules. These are algorithm refinement, modularity, and systematic search for the best compromise between management efficiency and architectural constraints (Fresse et al 2014). The authors are interested in the fact that an FPGA its possible programmed once or programmed several times to implement new projects. You can partially use the FPGA processor. Its posiible to program them using an external device by loading the configuration stream in programming mode. The SRAM or EPROM in the FPGA stores all configurations that are easy to delete.

The main advantage of such platforms is their ability to operate at low power – so compared to the SENTIO32 microcontroller (AVR32), which consumes 78.45 mW, the FPGA consumes only 6.77 mW (Compton et al 2002). Although compared to systems on a chip, the power consumption of an FPGA is high, but in practice the problem is solved by using a flash FPGA, which saves the entire project when the power is turned off until it is turned on again.

The second advantage is the integration of this platform with CPLD, ASIC, FPAA and SOPC, MPSOC, MPPSOC as a system on a single chip. This means that depending on the needs of the architecture, you can use FPGA with various additional devices or on a single chip.

The third advantage is the high parallelism of tasks, because the most important thing in the formation of the sensor network is to build the correct algorithm of processes for processing information. The author speaks about such algorithms used in signal and image processing applications and requiring significant computational resources (Cieszewski et al 2013). Hardware implementation of these algorithms requires the use of digital signal processing technology and devices that support this process. It is a programmable FPGA platform that provides a high level of parallelism of computational tasks in algorithms. Algorithms used in signal and image processing applications require significant computational resources. And the practice of using such algorithms has proved that it is best to use FPGA (Hemnath et al 2013), (Drimer, 2008).

The fourth advantage is the flexibility of the embedded software, which can include various tasks such as strict cryptography, data compression, encoding and decoding tasks, etc.

The fifth advantage is that FPGAs are useful for large sensor systems and complex computing-intensive tasks with high performance (Huffmire et al 2008).

The seventh advantage is the confidentiality of the project, which have not easily accessible.

The eighth advantage is the separation of projects from different clients in a single configurable space. "The upper half of the FPGA can be allocated to one client, and the lower half to the second," says (Jonas Krautter et al 2018) as another ITEC member. This usage scenario is highly desirable for cloud services where tasks are associated, for example, with AI applications such as machine learning or financial applications, which must to run databases.

Many modern FPGAs can dynamically change part of their configuration during the bitstream.

From a security point of view, FPGAs are increasingly used, but their capabilities have not yet been properly explored (Piltan et al 2012).

4. Pipeline cybersecurity and Management Solution

In the field of monitoring and ensuring safety in the oil and gas industry, there are a number of solutions in the form of modules for detecting leaks, thefts, line breaks, monitoring tightness and voltage throughout the life of the pipeline. The combination of products, solutions and services for complete pipeline management meets operational, environmental, environmental and regulatory requirements. Regardless of the purpose of pipelines for oil, or gas, water, or products from the chemical or any other industry, there are comprehensive solutions tailored to the specific needs of use.

At the site Phoenix Contact company (https://www.phoenixcontact.com/online/portal/gb] fig. AAA) has the Pipeline Management Solutions for designs and develops innovative technologies for secure communication in manufacturing and process automation and monitoring Figure 1) , which applies for the process industry as well and certainly for a data transmission solution for pipeline monitoring and leakage detection systems (Nasheed et al 2018).

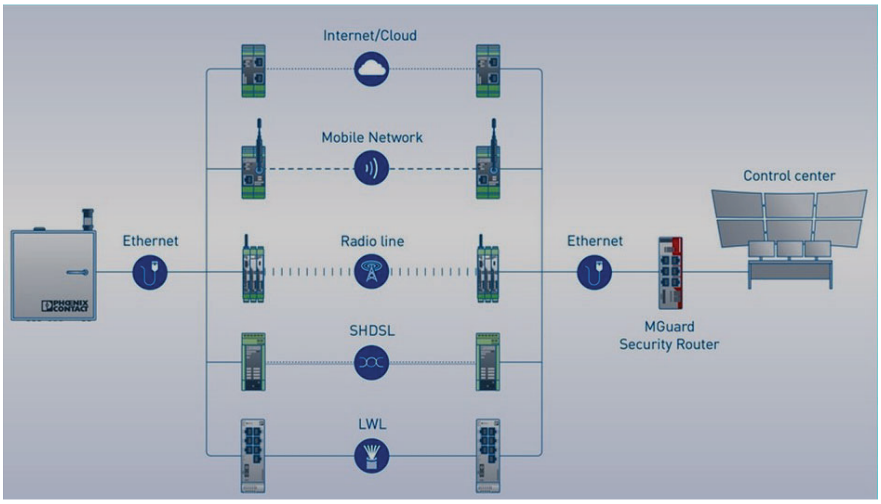


Figure 1: Connectivity options between measuring points and control center (Nasheed et al 2018)

Pipeline systems suffers from hacker attacks and information leaks, and according to statistics, according to Juniper research for 2019, the damage was 2 trillion dollars. For example, in 2018, a series of cyber intrusions were been on natural gas pipelines in the United States, which led to the shutdown of systems in four major national companies (Azubogu et al 2013).

In this connection, sharply there is a question on creation of the existing information security measures for many sectors of the developed industrial society.

Sensor Networks (SN) refers to a group of spatially dispersed and dedicated sensors for monitoring and recording the physical conditions/parameters of pipeline environment, and organizing the collected data at a central location. Research in Security SNs is gaining interest due their developing methods of Attacks to applications and elements of system that using the constrained resources of sensor nodes bring on the field. In this type of ad-hoc network, routing of data to the base station with secure transmission is of prime concerns. In article (Kumar, 2013) author says about possible improvements in SN routing and security through the employment of concepts coming from Artificial Intelligence (AI) area, such as swarm intelligence, artificial immune systems and artificial neural networks. Since SNs are distributed computing networks, the use of AI makes the network "cognitive" toward solving problems these networks are often prone too.

As presented in the work (Kumar, 2013) is improve WSN Routing and Security is possible with an Artificial Intelligence approach and the defence mechanism module as shown in Figure 2. This module is done been designed for built every sensor node, which can combat against such attacks, with minimal help from Base Station.

Is assumed, that an adversary to attack any victim node by sending malicious packets to hack the information in the victim node.

In addition, intruder node is assume as mobile with its attack packet bit energy varying, because that the intruder's mobility and bit energy of attack packet is highly random in nature. Even, the attacker's strategy is random. The module consists of an ANN, whose inputs are distance, bit energy, attack count and attacker strategy.

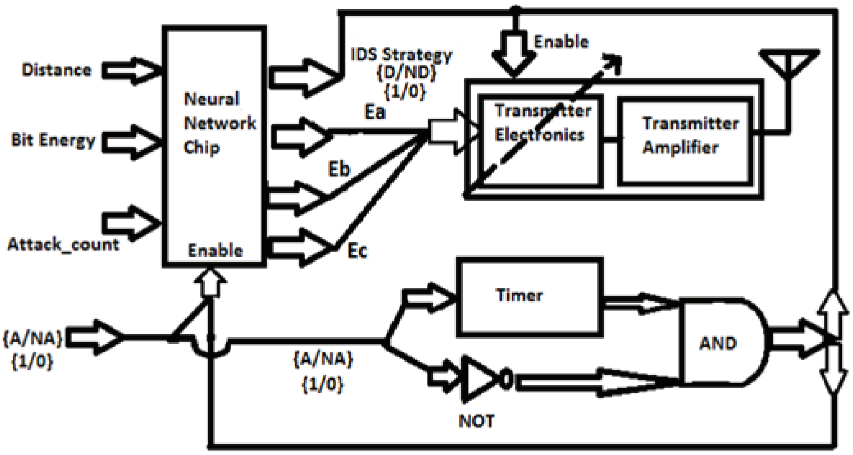


Figure 2: Intruder Defense Mechanism [26]

The need to develop a comprehensive approach to protecting wireless sensor networks is to provide protection against most existing attacks, which is achieved only when integrating machine learning solutions into the system or regularly updating the knowledge base. Therefore, there is a need to conduct new research in this direction.

5. Proposed models and management methods

Research and development of models and methods for integrated management sensor networks can improve the parameters that determine the quality of functioning the entire system as a whole. The objects in the works (Abramov et al 2013) are often set static, and the implementation of the proposed models depends on the tasks set and the limitations of system resources. Let us look at the main works on ensuring the security of wireless sensor networks using machine learning, as well as methods and solutions.

In paper (Chander et al 2018) authors are formulate an Online Locally Weighted Projection Regression (OLWPR) for anomaly detection in Wireless Sensor Network. Linear Weighted Projection Regression methods are nonparametric and the current predictions are perform by local functions that use only the subset of data. In work (Alotaibi, 2019) an efficient approach called Hamming residue method (HRM) is present to mitigate the malicious attacks.

Wireless sensor networks monitor dynamic environments that change rapidly over time. This dynamic behavior is either cause by external factors or initiated by the system designers themselves. To adapt to such conditions, sensor networks often adopt machine-learning techniques to eliminate the need for unnecessary redesign. Machine learning is also inspiring many designers of practical solutions that maximize resource utilization and prolong the lifespan of the network (Alsheik et al 2014). Knows, to maintain the level of autonomy and intelligence, it is necessary to provide a high level of reliability and security. The technical characteristics of existing systems have a significant impact on the data processing speed, quantitative and qualitative characteristics, but mathematically the calculation and optimization tools make it possible to improve these parameters (Singh et al 2017)

In process developing models and methods for controlling sensor networks, prefer to choose reliable, "approved" security mechanisms based on established protocols and algorithms. In this position, cost-effectiveness may be important, but not at the expense of security and reliability. In some cases, older, more Mature and, therefore, more reliable integrated circuits are used. In addition, as known to use the latest technologies and equipment that are more resistant to aggressive attacks and increase the entry threshold of potential opponents due to the higher cost (Chun-ta Li, 2010).

Analyze works and modern solutions; we can will consider the main concepts and description of the design process. Very interesting is the implementation of systems with FPGA, which significantly affect the main quality indicators. The mathematical model of a system depends on the structure, elements, and relationships between them that determine its complexity. Consider the main works where FPGA is part of the system. Figure 3.

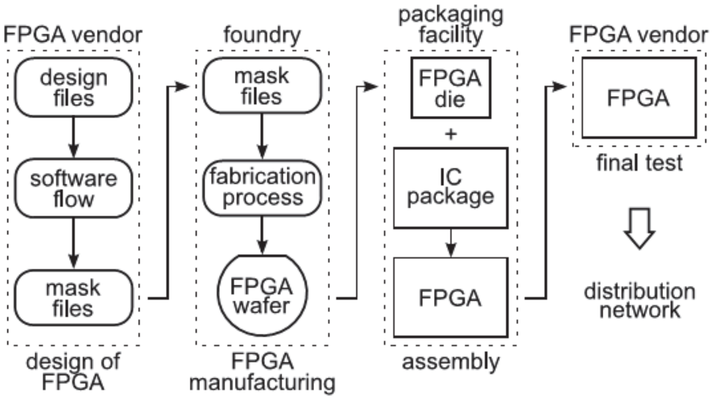


Figure 3: Simplified depiction of the FPGA design, manufacturing, packaging, and testing processes (Chun-ta Li, 2010).

In work (Fresse et al, 2014), the feasibility to identify mathematical models for exploring the NoC on FPGA devices have shown. The number of FPGA resources (LUT, MLUT and FF) can to be estimate using these models and without experiments.

NoC are communication structures proposed as a solution for the communication challenge. NoC architecture is composed of several basic elements depicted in Figure 4:

- Network Interface (NI): it enables PE (Processing Element) to communicate with routing node they are connects too.
- Routing node (or switch): according to the routing algorithm, the switch sends packets to the appropriate link in the network.
- Links: connect the routing switches together or switches to NI.
- Processing Element: these units correspond to various modules of SoC, such as IP blocks, memories, processors.

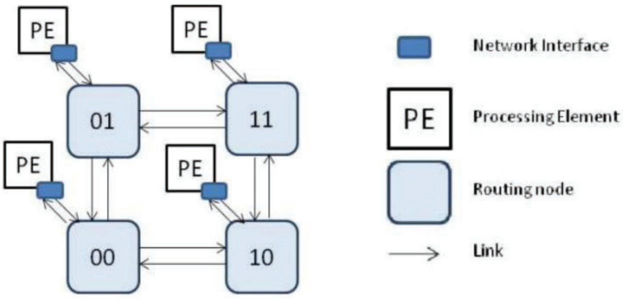


Figure 4. Basic elements for the NoC structure (Fresse et al, 2014)

The mesh topology is the most appropriate for FPGA devices (as depicted in Figure 4) and for NoC. The following work uses mesh-based NoC but the framework can be extend to others. Systems On-Chip (SoC) using Network on Chip (NoC) are the most appropriate systems for real time embedded applications. NoCs are emerging communication structures as they provide high bandwidth and high scalability with low power. The mathematical models obtained can estimate the number of resource with the lowest error rate.

The work (Chander et al 2018) identifies and presents the main advantages of NoC for designing modular and scalable communication architectures in which different IP cores were connect to a network based on a router using the appropriate network interface.

FPGA A "Spartan2E"there is module is use to host the algorithms that is needed for the autonomy of the agent. In addition, the FPGA provides a re-configurable processing solution where algorithms can be tested and updated, when required. This provides flexibility to the system in regards to the development and optimization of algorithms (Murphy et al 2007)

The hardware in work (Sahu et al 2013) was design to be versatile since its wireless core system can utilize custom interfaces and so can used in a host of ad-hoc networks and applications/scenarios. The design of the Wireless Sensor Networks (WSN) hardware supports autonomous formats and uses wireless units that are design to collect data and transmit to a central host (or distributed hosts). The unit is make up of a modular system (Figure 5) of hardware components that include resources for computation, communications and sensor implementation for its system. Thus, the module is adaptable to various configurations due to its flexible and generic design capabilities.

Refer to this research (Sahu et al 2013) a position FPGA-based mathematical error-based tuning sliding mode controller (MTSMC) is propose for PUMA robot manipulator. Pure sliding mode controller with saturation function and fuzzy sliding mode controller with saturation function have difficulty in handling unstructured model uncertainties. It is possible to solve this problem by combining fuzzy sliding mode controller and mathematical error-based tuning.

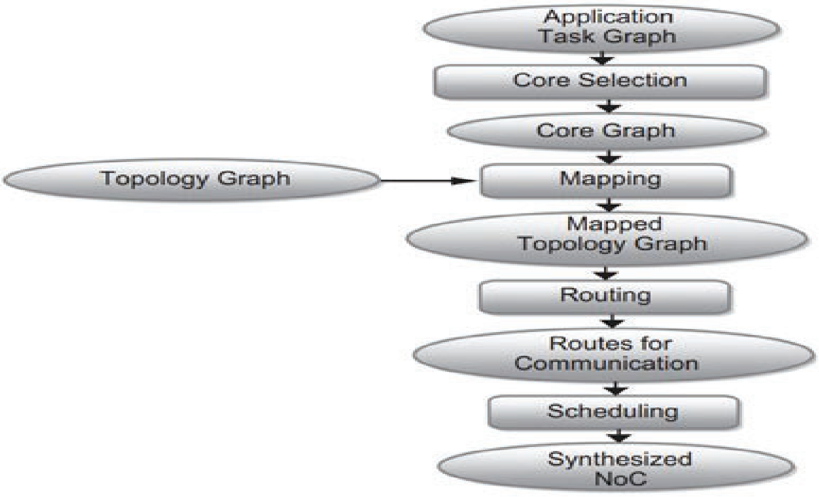


Figure 5: Application-specific NoC design flow

The aim of the research in (Basan et al 2017) is to develop a method that, will effectively to detect active attacks by an attacker based on the analysis of network traffic and physical parameters of wireless sensor network (WSS). Nodes, based on the use of probabilistic functions, calculating the confidence interval and the probability of deviation of current indicators from the confidence interval. The novelty of this method lies in the fact that when a malicious node is detected, the distribution function of the current node is not compared with the reference distribution, but it is estimated that the current value of the function falls within the confidence interval. The advantage is that in mobile WSS there is a high probability of not only network attacks, but also attacks aimed at disrupting the physical activity of the node (violation of the movement of nodes, exhaustion of energy resources, noise transmission channel).

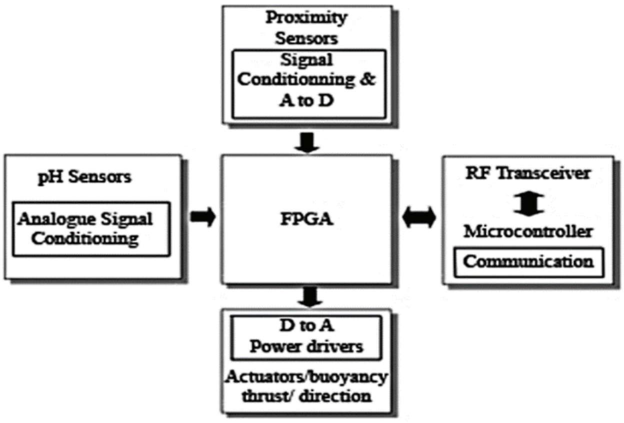

Proximity Sensors
Signal Conditionning & A to D
pH Sensors
Analogue Signal Conditioning
FPGA
RF Transceiver
Microcontroller
Communication
D to A Power drivers
Actuators/buoyancy thrust/ direction

Figure 6: A modular description of the CAA [44]

FGPA accelerators are entering the main stream of high performance computing. Researchers have demonstrated impressive performance gains using FPGA acceleration, 100-1000× in some applications. These applications have generally been hand coded by skilled logic designers, and tend be to build as point solutions to particular problems or families of problems. In the work (Van Court et al 2006) is present a general technique based on the knowledge of the application's multidimensional indexing.

The use of FPGA in multi-agent systems has a different purpose: in articles (Kuandykov et al 2015), (Uskenbayeva et al 2013), (Zhou et al 2016), (Aitimov et al 2017), (Drimer, 2008), which confirms the relevance of FPGA and its significance.

In this way (Zhou et al 2016), a multi-agent simulation method based on FPGA for urban land layout is propose in this paper. The evolution rule of urban eco-land explores by combination of cellular automata, dynamic reconfiguration and multi agent methods.

Findings of existing techniques used to improve the security in WSNs vulnerable to different attacks in site - https://www.tsa.gov/sites/default/files/pipeline_security_guidelines.pdf. One of the major security issues in WSNs is eavesdropping which is an attack that captures information transmitted over a network by other devices. Eavesdropping attacks are insidious, and it is difficult to know they are occurring. Once you connected to a network, users may unwittingly feed sensitive information like passwords, account numbers, surfing, the content of email messages, etc.

Therefore, security enhancement communication of wireless networks is in demand (Chander et al 2018). Solutions with FPGA and machine learning for security are presented in the works (Nasheed et al 2018), (Sahu et al 2013**), (**Mohamed et al 2015).

Relying on platform of MATLAB software and dynamic reconfiguration of FPGA logic resources, a regional ANN-CA-Agent model for evolution and prediction model of urban land layout is established.

Figure 7 illustrates a BBB network manager/gateway system containing field-side radio modules for wireless communication, a core system processor, and a communication coprocessor with an internal (soft) embedded processor implemented in an FPGA. FPGAs provide a flexible platform that offloads the task of communication over the plant network, freeing up the core system processor to handle the countless tasks of managing the network, ensuring secure communication, and more.

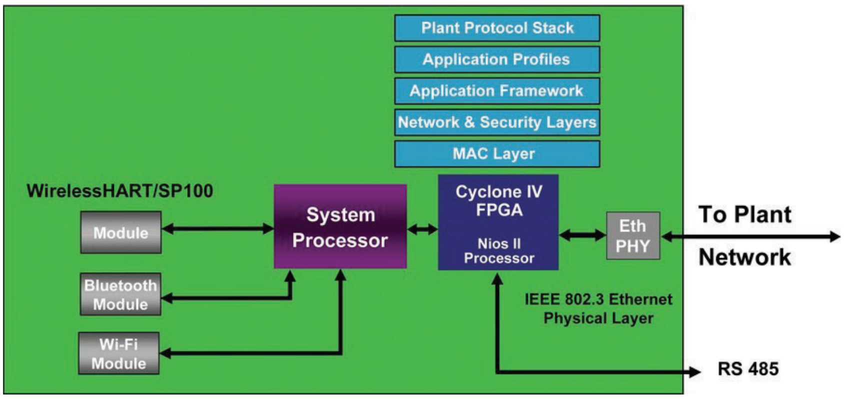


Figure 7. BBB: A wireless network manager/gateway system includes radio modules, a core system processor, and a communication coprocessor with an embedded processor in an FPGA (http://industrial.embedded-computing.com/article-id/?4409=)

In this paper, we propose the idea of a response from the knowledge base to types of security attacks based on FPGAs using multi-agent systems (MAS). As agents, we use the sensors themselves, which sense the types of attacks. All information is send to the processor and the knowledge base in memory (Flash memory). In this case, the amount of memory is not a critical value, although the most unprotected component of the platform is memory, especially since it is often embedded memory. To ensure fast response of the system, we suggest using a key code. But if the information on attacks is growing rapidly in the system, then here we use the idea of machine learning, when the response to an attack is self-learning and the response (reaction) to an attack is fast, because the FPGA passes millions of bit streams in parallel and thanks to a formalized knowledge base. After a new instant response is generates, the session memory is reset to zero. Training have been base on the clustering method.

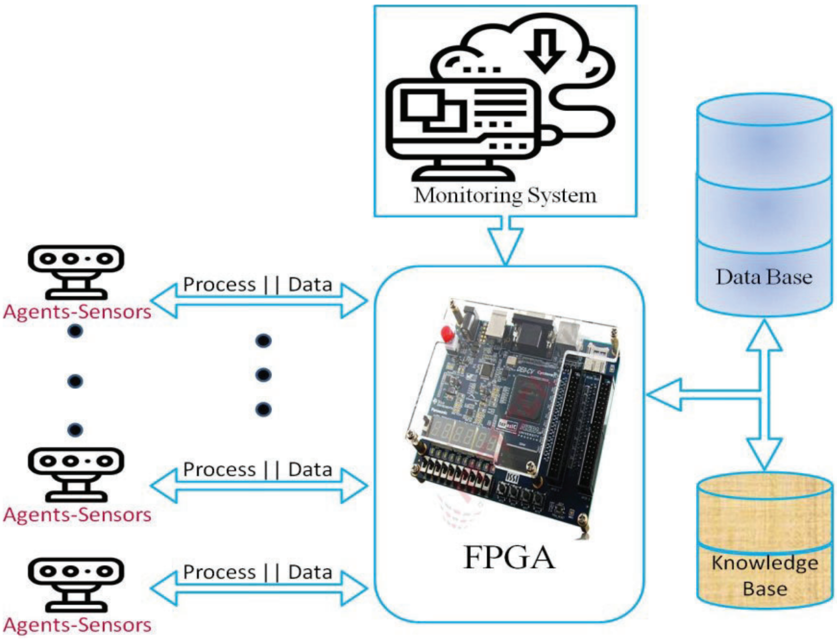


Figure 8: Generalized system model with FPGA

The advantage of system is achieve due to the speed of the monitoring system for the attack response, which was identify in knowledge base. The monitoring system detects violations of processes or attacks, and identifies them based on the unique code of process execution sequence that was generate in FPGA and stored in the knowledge base.

FPGAs in security modules often include:

- Cryptographic module specification
- Cryptographic module ports and interfaces

- Physical security
- Operational environment
- Cryptographic key management
- Electromagnetic interference/electromagnetic compatibility
- Self-tests, etc.

In the presented model (Figure 8), a high reaction rate is achieve due to not only the key code, but also the code corresponding to the protection action.

Further development of the system includes machine learning and a knowledge base to ensure high-speed decision-making for a complex response.

This model can be to apply in pipeline Cybersecurity and in ensuring the reliability of liquid and gas delivery systems (-https://www.tsa.gov/sites/default/files/pipeline_security_guidelines.pdf).

6. Conclusion

Modern security requirements are increasing and existing solutions need to be improve, so companies providing solutions and services are moving on to developing corporate cybersecurity programs or creating a cybersecurity framework, which is the rationale for using knowledge-based machine learning.

Because of the analysis of the cited works, it was determined that machine learning provides a set of methods to increase the ability of a wireless sensor network to adapt to the dynamic behaviour of its environment. In particular, wireless sensor network developers have in fact agreed with machine learning paradigms to address the widespread system-security challenges.

References

Ahmed, S. (2018) Sensor Networks Attacks Classifications and Mitigation, Annals of Emerging Technologies in Computing (AETiC), Vol. 2, No. 4, pp. 28-43, International Association of Educators and Researchers (IAER).

Abramov E.S. and Basan E.S. (2013) Development of secure cluster based wireless sensor network model, Journal Engineering Sciences, Vol. 12, No.149, pp 48-56.

Aitimov, M.Z. et al. (2017) Analysis of the structure and calculation of time for the environmental monitoring system with multi-parameter sensors, Journal ИЗВЕСТИЯ НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК РЕСПУБЛИКИ КАЗАХСТАН, Национальная академия наук Республики Казахстан ,Алматы, Vol. 1, No. 422, pp 149-156

Alsheikh M., Lin S, Niyato D. and Tan H. (2014) Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications, [online], https://arxiv.org/pdf/1405.4463.pdf].

Alotaibi, M. (2019) Security to wireless sensor networks against malicious attacks using Hamming residue method, Journal Wireless Com Network, Vol. 8, [online], https://doi.org/10.1186/s13638-018-1337-5, https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-018-1337-5

Azubogu, A.S., Idigo V. E. and Simon O. S. (2013), Wireless Sensor Networks for Long Distance Pipeline Monitoring, Engineering and Technology International Journal of Electronics and Communication Engineering. Vol.7, No.3.

Basan, A.S., Basan E.S. and Makarevich O.B. (2017) The method of resistance to active attacks in wireless sensor networks, Engineering Sciences, DOI 10.23683/2311-3103-2017-5-16-25.

http://industrial.embedded-computing.com/article-id/?4409= (2010) Building an FPGA-based solution for industrial wireless sensor networks, Industrial Embedded Systems

Borselius N., Security in Multi-Agent Systems, 2014, University of London, UK.

Cardoso M.P., Diniz P. C. (2010) Compiling for reconfigurable computing: A survey, ACM Computing Surveys, Vol. 42, No. 4, Article 13.

Chess, D., Harrison, C. and Kershenbaum, A. (2005) Mobile agents: Are they a good idea? Springer, Lecture Notes in Computer Science book series (LNCS, Vol. 1222), pp 25-45.

Cieszewski R. et al. (2013) Review of parallel computing methods and tools, Institute of Electronic Systems and University of Technology, Warsaw, Poland.

Compton K., Hauck, S. (2002) Reconfigurable Computing: A Survey of Systems and Software, ACM Computing Surveys, Vol. 34, No. 2, pp 171–210.

Chun-ta Li (2010), Security of Wireless Sensor Networks: Current Status and Key Issues, Smart Wireless Sensor Networks, London, UK

Drimer S. (2008) Volatile FPGA design security – a survey [Online]. http://www.cl.cam.ac.uk/~sd410/papers/fpga security.pdf

Fresse, V., Combes, C. and Belhassen, H. (2014), Mathematical models applied to on-chip network on FPGA for resource estimation, 21st IEEE International Conference on Electronics Circuits and Systems, Chengdu, China

Galov N. (2020) 40 Worrisome Hacking Statistics that Concern Us All in 2020, [online], https://hostingtribunal.com/blog/h

Hedina, Y., Moradiana, E. (2015) Security in Multi-Agent Systems, 19th International Conference on Knowledge Based and Intelligent Information and Engineering, Procedia Computer Science, V. 60, pp 1604 – 1612, Elsevier

Hemnath, P. and Prabhu V. (2013) Compression of FPGA Bitstreams Using Improved RLE Algorithm, International Journal of Scientific & Engineering Research, Vol. 4, No, 6.

Hohl, F. (1998) A model of attacks of malicious hosts against mobile agents, Proceedings of the ECOOP Workshop on Distributed Object Security and 4th Workshop on Mobile Object Systems: Secure Internet Mobile Computations, [online], http://mole.informatik.uni-stuttgart.de/simc98.ps.gz.

Huffmire, N. et al. (2008) Managing Security in FPGA-Based Embedded Systems, IEEE Design and Test of Computers, Vol. 25, No 6, pp 590-598.

Jung, Y. et al. (2012) A survey of security issue in multi-agent systems, Journal 'Artificial Intelligence Review, V.37, pp 239-260, Springer.

Kadam, M. and Sawarkar, K. (2014) Overview of Reconfigurable Hardware for Efficient Implementation of DSP Algorithms, IOSR Journal of Engineering (IOSRJEN), Vol. 04, No. 02, pp 34-43, Mumbai University, India**.**

Krautter, J., Gnad, D.R.E. and Tahoori, M.B. (2018) FPGA hammer: Remote Voltage Fault Attacks on Shared FPGAs, suitable for DFA on AES, IACR: Transactions on Cryptographic Hardware and Embedded Systems, No 8, pp 44-68.

Kuandykov, A. et al. (2015) Multi-Agent Based Anti-Locust Territory Protection System, Procedia Computer Science, Vol. 56, pp 477-483.

Kumar S.E. (2013) Improving WSN Routing and Security with an Artificial Intelligence approach Vol.1334, Karnataka, India

Kuandykov, A.A. et al. (2015) Design and construction a model of remote control software mobile robot for MAS, IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan

Majzoobi, M., Koushanfar, F. and Potkonjak, M. (2009) Techniques for Design and Implementation of Secure Reconfigurable PUFs, ACM Transactions on Reconfigurable Technology and Systems (TRETS), No. 5

Mohamed, A., Hamdi, M. S., and Tahar, S. (2015) Machine Learning Approach for Big Data in Oil and Gas Pipelines, International Conference on Future Internet of Things and Cloud, Vol.54, IEEE

Murphy, F. et al. (2007) Development of a Wireless Sensor Network for Collaborative Agents to Treat Scale Formation in Oil Pipes, Springer, Vol. 4373, pp 179 – 194, Berlin.

Nasheed, F. et al. (2018) Design of Oil Pipeline Monitoring System based on Wireless Sensor Network, Iraqi Journal of Computers, Communications, Control & Systems Engineering (IJCCCE), Vol. 18, No. 2, pp 427.

Piltan F. et al, (2012) Methodology of FPGA-Based Mathematical Error-Based Tuning Sliding Mode Controller, International Journal of Control and Automation, Vol. 5, No. 1.

Sahu, P. K. and Chattopadhyay, S. (2013) A survey on application mapping strategies for Network-on-Chip design, Journal of Systems Architecture: the EUROMICRO, Vol. 59, No.1.

C. Singh, R., and Kaur, M. (2017) Review of security enhancement techniques for Wireless Sensor Network, International Journal of Electronics Engineering Research. Vol. 9, No 8, pp. 1185-1196

Chander, B., Kumar, P. and Kumaravelan, B. (2018) Analysis of Machine Learning in Wireless Sensor Network, International Journal of Engineering & Technology, Vol. 7, pp 185-192,

Trimberger S., Conn R. O. (2007) Remote field upgrading of programmable logic device configuration data via adapter connected to target memory socket, United States Patent Office, [Online], http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=7269724

Uskenbayeva, R.K. et al. (2013) Organization of computing processes in the large heterogeneous distributed systems, Robotics (ISR 44th International Symposium on Digital Object Identifier, Vol. 10, pp. 1- 4

Uskenbayeva, R.K. et al. (2013) Models and Methods of Joint Work Management of Group of Unmanned Vehicles, 13th International Conference on Control, Automation and Systems (ICCAS), Gwangju, Korea

Van Court, T., Herbordt M. (2006) APPLICATION-SPECIFIC MEMORY INTERLEAVING FOR FPGA-BASED GRID COMPUTATIONS: A GENERAL DESIGN TECHNIQUE, International Conference on Field Programmable Logic and Applications, Madrid, Spain.

Wang W. et al. (2013) Security in Wireless Sensor Network's, Wireless Network Security. Springer, Berlin, Heidelberg, [online], https://link.springer.com/chapter/10.1007/978-3-642-36511-9_7

Zhou X. and Fu W. (2016) A Multi-Agent Simulation Method of Urban Land Layout Structure Based on FPGA, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLI-B8, Prague, Czech Republic

https://www.phoenixcontact.com/online/portal/gb] fig. AAA

https://www.tsa.gov/sites/default/files/pipeline_security_guidelines.pdf]

Security of Systems on Chip

Eda Tumer and Leandros Maglaras
De Montfort University, School of Computer Science and Informatics, Leicester, UK
ed.tumer@gmail.com
leandros.maglaras@dmu.ac.uk

DOI: 10.34190/EWS.20.005

Abstract: In recent years, technology has started to evolve to become more power efficient, powerful in terms of processors and smaller in size. This evolution of electronics has led microprocessors and other components to be merged to form a circuit called System-on-Chip. If we are to make a vast and cursory comparison between SoC and microcontrollers, microprocessors and CPUs; we would come to the conclusion of SoCs being a single chip, doing all the things the other components can do yet without needing any external parts. So SoCs are computers just by themselves. Furthermore, SoCs have more memory than microcontroller in general. Being a computer just by themselves allows them also to become servers. Nowadays, an SoC may be regarded also as a Server-on-Chip (Davis, 2012).

Keywords: Cyber Security; System on Chip; Microcontroller Security; Hardware Security

1. Introduction

Briefly, an SoC is an integrated circuit(IC) viz. a chip (Rajesvari et al, 2013) which is, in other words, a wide-ranging integration of various components such as processor cores, accelerators, peripheral controllers, memory controllers, intellectual properties(IP) along with on-chip and off-chip memory components on only one IC for specialised uses (Davis, 2012)(Wang et al, 2008). As long as Moore's Law (Gelsinger, 2006) continues on, the chips and its parts will shrink further enabling more transistors to be put on the same part of the silicon (Davis, 2012).

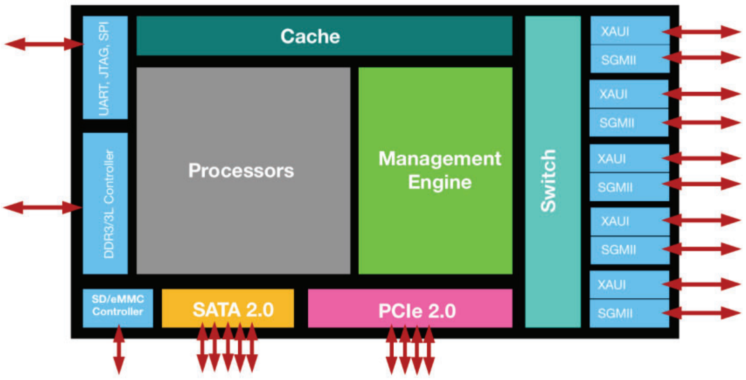


Figure 1: System on Chip Properties

The Internet of Things is the interconnection between intelligent devices and systems, and data collected by physical sensors, actuators and other physical objects (Nolan, n.d.)(GSM Association, 2014). These devices process and report data to the network and some of them provide outputs which control other devices or systems, either autonomously or by network commands (Nolan, n.d.).

In simpler words, the IoT is a system of devices that are interconnected. Furthermore they can control each other to mitigate changes in the environment which are provided by sensors. The controllers of these devices can connect to the internet in order to provide distant control to users.

2. Role of SoC and Microcontrollers in IoT Systems

After the invention of the first silicon chip in 1958 by Jack Kilby, the chips diminished in size. When a whole system was integrated on a single chip in 1970, it led to huge downsizing of physical systems. With these advancements, implementing a system started to require less power and space. It also grants higher performance due to increased number of circuits on the chip, and further reliability (Palomar Technologies, n.d.). Subsequent to these advantages, it was inevitable that many systems started to use SoC, though some continue to use microcontrollers still.

As the needs and demands for systems get complicated, so do the systems' operations and structures. The operations and functions of these systems naturally depend on the signals of the components and calculations or decisions made by the brains of this system, namely microcontrollers and SoCs. In a case that these components are cornerstones of the system, they may create single-point of failures if not protected and handled with redundancy. Since these MCUs and SoCs are interconnected, they also expose vulnerability risks in the whole system. Therefore they have to be protected against possible cyber attacks.

Due to the reason that the concept of IoT is growing exponentially (as seen on Figure 2), the need for protection of these systems become more and more crucial with the same rate (Maglaras et al, 2018). The full range of potential applications of IoT devices has not yet been perceived and its application range includes infrastructure, personal, medical, industrial and more (Nolan, n.d.). In order to provide the highest level of integration and save space, many IoT devices are thought to have a single SoC (Nolan, n.d.).

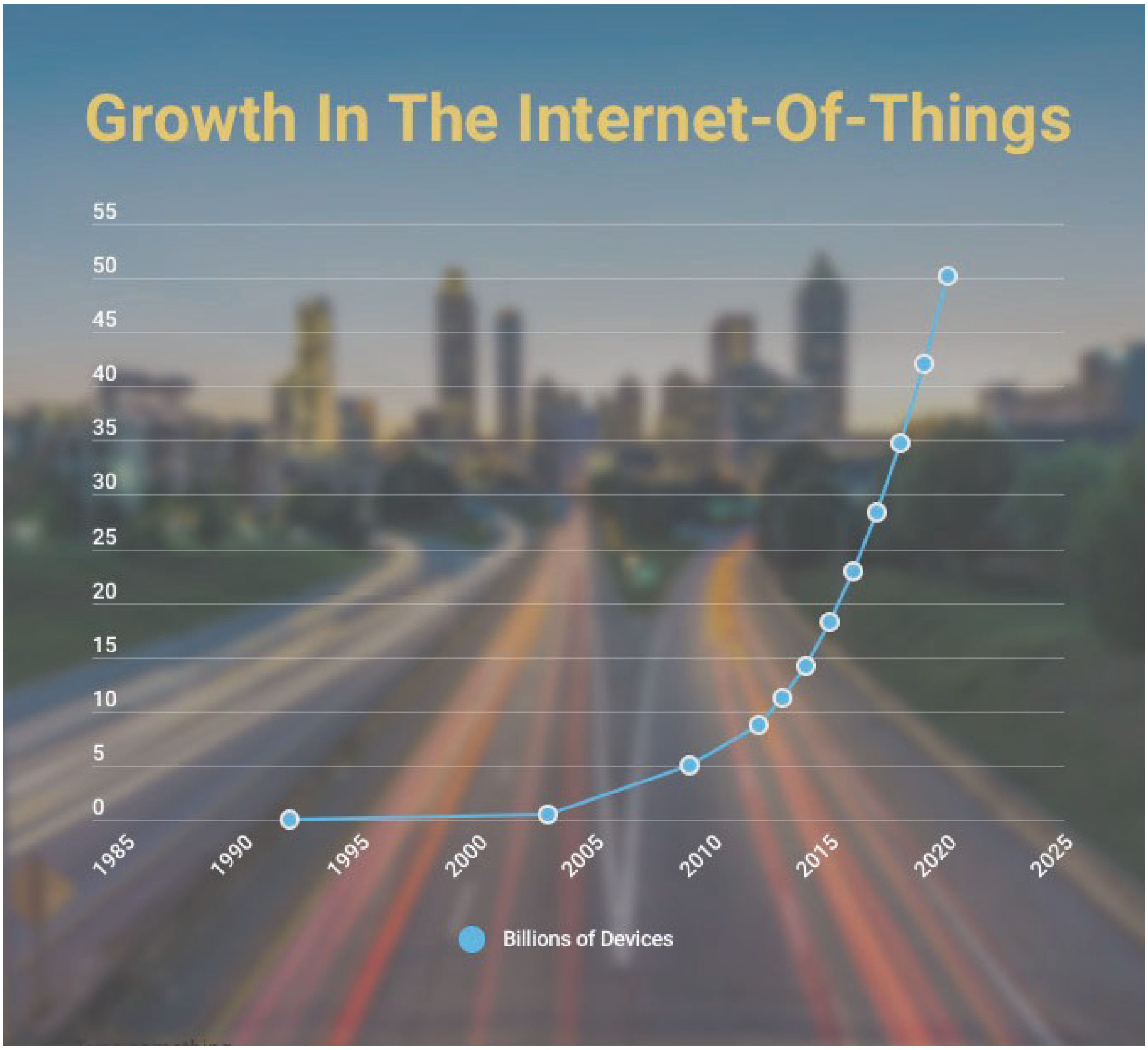


Figure 2: Growth of IoT (NCTA, 2015)

Because IoT systems are a network of many connected devices, the process of securing the whole system also becomes complex and challenging. There can be malicious hardware or software which may compromise the security of this ecosystem. Security issues of IoT systems are to be handled with even more care because the data transmitted in the network between the nodes may be sensitive, ranging from personal and private data to secrets of an organisation (Kumar et al, 2017).

To implement the system in a secure way, the organisations have to consider potential adversaries. It is described to be a very creative process by (Ray and Jin, 2015) because the team has to think of every possible scenario and considerations of immoral actions and leakages. The following adversary examples are self-explanatory hence only their titles are mentioned.

- Unprivileged Software Adversary, System Software Adversary, Software Covert Channel Adversary (when a side-channel or covert-channel has access to non-functional characteristics of the system) (Ray and Jin, 2015)
- Naive Hardware Adversary (when attackers manage to get into hardware devices), Hardware Reverse-Engineering Adversary, Malicious Hardware Intrusion Adversary (Ray and Jin, 2015)

3. Hardware Security of SoC

As a consequence of the rapid growth of IoT (NCTA, 2015)(Evans, 2011), many edge devices have to be given access to a large number of security assets which are to be protected from unauthorised or malicious access (Nath et al, 2018). The edge nodes are the devices which gather data from sensors and interact with physical objects in the ecosystem (Kumar et al, 2017).

SoC design consists of interactive firmware and hardware. In order to verify the security of these modules, a method called instruction-level abstraction (ILA) is used, which addresses verification of properties across firmware and hardware by "Raising the level of abstraction of hardware modules to be similar to that of instructions in firmware" (Malik and Subramanyan, 2016). Hence, this technique helps to ensure that hardware implementations and the firmware of the chip do not contradict with each other. Moreover, there is hardware support for security called Trusted Execution Environment (TEE) for microcontrollers and SoCs used in sensitive applications (Kumar et al, 2017).

Security operations such as encryption-decryption and access blocking are handled at the hardware level (Nichols, 2019). In order to raise security of a chip, cryptography is used. There are two crypto systems, symmetric and asymmetric (Elkeelany and Olabisi, 2006). Symmetric (Schubert and Anheier, 1999) systems use only one key to encrypt and decrypt messages. Asymmetric systems on the other hand, use two that are called public and private keys.They use the public key to encrypt and the private key to decrypt messages (Elkeelany and Olabisi, 2006).

Hitherto, patching has been used only on the software (or firmware) side of the system, in order to strengthen or to cover a discovered security hole. "Hardware patching" means hardware implementation of security requirements that permit seamless post-silicon adaptation (Nath et al., 2018). Recently, it is firmly suggested that also the hardware should be patchable on devices which are to become part of the IoT (Ray et al, 2017). The reason given for this is because soon it will not be possible to fix every single security vulnerability only by altering the software (Ray et al, 2017). In addition to that, implementing a functionality in hardware reduces the energy consumption greatly which is an immense advantage (Ray et al, 2017). Secure hardware systems demand a very high level of verification due to the fact that "patching" is not an option (Bartley, 2017). This can indicate that when patching is made available, enforcements to verify hardware systems may be reduced.

In order to increase patchability of the system, Nath et al.2018, introduces a new framework with a centralised Reconfigurable Security Policy Engine (RSPE), smart security wrappers, and Design-for-Debug (DFD) infrastructure interface. Design-for-test (DFT) infrastructure is used to tackle hardware trojan problems with minimum hardware insertion into SoCs (Backer et al, 2014). The DFT infrastructure enables SoC integrators to gain access and have control over IP hardware during post-silicon testing, and after the testing, the DFT infrastructure remains unused (Backer et al, 2014).

The integration of these chips involves connecting third party IP blocks along with DFT hardware pieces, and verification and validation processes of the overall system design (Saleh et al, 2006). These third party hardware IP may be acquired from untrusted vendors and in that case they may have a different level of integrity and security issues (Ray et al, 2018)(Gundabolu and Wang, 2018). These third party vendors may compromise the integrity and security of an SoC design by inserting malicious hardware trojans in their IP (Ray et al, 2018), so the risk of the built-in hardware trojans emerges (Ray et al, 2018)(Li et al, 2016). Unlike software trojans, hardware trojans cannot simply be removed and these trojans may hinder fundamental functionalities of the hardware (Kim and Villasenor, 2015)(Jin and Phatak, 2015). The detection of these trojans occurs by going through post-deployment functional verifications (Jin and Phatak, 2015).

3.1 Functional Verification Examples

Property Specification Language (PSL) Assertions: They are used to aid verification of complex systems and they may be used to monitor various signals for legal and illegal transitions (Montador, n.d.).

Transaction based testbench-simulations: Traditional behavioural testbenches may become bulky when it comes to testing complex systems (Montador, n.d.) (e.g. going through all the states of a system would be burdensome and cost too much time).

Verification IP: It is used to ensure IP interoperability and system behaviour (Mentor, n.d.). They can also support rapid testbench development (Sondrel, 2016). Usage of IP in a testbench may include functionalities such as keeping track of the traffic between interfaces, and reporting on any violations with respect to existing standards or specifications (Mantador, n.d.).

Trace-Based Debugging: This technique allows SoC integrators to monitor internal signals of IP cores embedded in the SoC (Backer et al, 2016). Its uniqueness is that there is no need to stop the execution in order to capture the values of the signals (Backer et al, 2016).

These are only a few examples of how a hardware can be verified. If we are to talk about standards, the Universal Verification Methodology (ACCELLERA, 2009) is widely used for integrated circuits.

4. Software Security Architecture

Naturally, protecting a chip from hardware faults and producing them without bugs and malfunctions is not solely enough. The SoC has to be protected also from possible threats coming from faulty firmware or the OS kernel and applications (Ray and Jin, 2015).

Simple software designs are commonly used in most security systems because when the system is simple, there is a lower risk of bugs which are a security vulnerability (ARM, 2009). Because multi-threading introduces another layer of software complexity -which is also very hard to test due to timing sensitivity-, many Secure world software implementations may choose to implement single-processor (ARM, 2009). Secure world, also known as TrustZone and Trusted Execution Environment, is a sand-boxed execution environment that has higher privileges than the normal kernel (Vita Development Wiki, n.d.).

Unauthorised access to sensitive assets proposes a critical issue to SoC designs. And modern SoC designs contain numerous assets of this type which must be secure against unauthorised access (Ray and Jin, 2015). The authentication mechanisms that protect these assets are governed by security policies which are also referred to as security architectures in an SoC design (Ray and Jin, 2015).

In order to make justified design choices concerning what and how much to protect, the threats must be well defined in the scope of their risks and the definition of security for that system (ARM, 2009).

In order to protect the system from malicious software, there are some policies introduced as follows:

Access Control: This policy defines which “agent” has access to which asset at different points of the system execution (Ray and Jin, 2015)(Goguen and Meseguer, 1982).

Information Flow: This technique attempts to analyse all the possible interactions and interferences (Goguen and Meseguer, 1982). Naturally, this policy is implemented along with access control policies and even with additional constraints if necessary (Ray and Jin, 2015). (An agent can be a hardware or a software component of an IP)

Liveness: To meet availability requirements, the system has to perform without any pause or “stagnation” throughout its execution which also prevents getting the system caught in deadlocks or livelocks (Basak, 2015). This policy is meant to ensure protection against denial-of-service attacks (Ray and Jin, 2015).

Time of Check vs. Time of Use (TOCTOU): the security objective of an SoC has to be that at any time, the device may only run authenticated firmware(Krstic et al, 2014). In other words, any agent that wants to gain access to a resource has to be authenticated (Basak, 2015) and it is ensured that the authenticated agent is really accessing the resource (Ray and Jin, 2015).

Message Immutability: Messages sent between components have to be exactly how they are expected to be (Ray and Jin, 2015).

Redirection and Masquerade Prevention: A message sent from a component to another cannot be delivered to any other component which masquerades itself as the destination component, or the message cannot be duplicated and sent to another component (Ray and Jin, 2015).

Non-Observability: A private message between two components should never be accessible by another IP during its travel (Ray and Jin, 2015).

Furthermore, besides those policies, there are several characteristics introduced within (Ray and Jin, 2015) such as: Boot Confidentiality which instructs that no IP may access internal registers during boot, Firmware Integrity, that any firmware running on an IP has to be signed or authenticated beforehand which is closely related to the TOCTOU policy.

5. Weaknesses / Vulnerabilities

As mentioned earlier, the IoT is drawing much attention and it is becoming highly prevalently used by numerous companies all around the world. There are already projects such as cloud based smart cities and perhaps in a few decades these projects will be live. Yet even now there are known failures of some IoT systems. Several of them, which are not deliberate violations, are listed below.

- The vulnerability, or more like a leakage, of emergency broadcast systems produced by Acoustic Technology Inc. (ATI) was found by Bastille Security that command packet broadcasts could be captured over the air, even modified and replayed (Sanders, 2018).
- Belkin WeMo, included digital keys in their firmware and the leakage of these keys was a great weakness for hackers to be able to take control of lights and home appliances (Heideman, 2016)(Goodin, 2014).
- Wink sold hubs that could connect to IoT devices. When the installed security certificate expired the whole IoT system malfunctioned/stopped working (Heideman, 2016), or as described in (Barrett, 2015), lobotomised.

These and many more IoT system failures are attributed to the lack of experience and expertise according to Cisco's Australian CTO Kevin Bloch as he mentioned in an interview (Reichert, 2017). It is credibly true for the technical part of failures of IoT systems, when the system is developed by using appropriate designs, standards and policies. But besides the technicality of a project, it should not be forgotten that the organisation itself plays a big role for a project which also affects communications in teams. Of course communication may be considered at an individual level but the culture inside an organisation has a very important impact on it. Thirdly, available tools may have diverse influences such as the familiarity of the team members about particular tools.

6. Novelty

There are numerous publications which discuss many aspects of chips' software and hardware security as referenced throughout the paper. Yet all of these references look at only singular and specific aspect of security. The purpose of this publication is to summarise how securing a chip is approached on both software and hardware during the whole process of making the chip and maintaining it in the course of its lifetime.

References	Year	Title	Hardware/Software-Firmware	Specification
Kumar, S. *et al*	2017	Security Enhancements to System on Chip Devices for IoT Perception Layer	Both	Cursory solutions and no verifications / focus on IoT devices
Malik, S. et al	2016	Specification and Modeling for Systems-on-Chip Security Verification	Hardware	Hardware Security Verification

References	Year	Title	Hardware/Software-Firmware	Specification
Ray, S. *et al*	2017	To Secure the Internet of Things, We Must Build It Out of "Patchable" Hardware	Hardware	Hardware Patchability
Nath, A. *et al*	2018	System-on-Chip Security Architecture and CAD Framework for Hardware Patch	Hardware	Hardware Patchability
Kim, L. *et al*	2015	Dynamic Function Verification for System on Chip Security Against Hardware-Based Attacks	Hardware	Hardware Attacks
Bartley, M.	2017	Hardware Security Challenges and Solutions	Hardware	Hardware Protection
Jin, Y. *et al*	2015	Introduction to Hardware Security	Hardware	Hardware Protection
Schubert, A. *et al*	1999	Efficient VLSI Implementation of Modern Symmetric Block Ciphers	Both	Ciphering data
Montador, A.	n.d.	Verification Methodology for Standards-based IP & SOC	Hardware	Hardware Verification
Sondrel	2016	Functional Verification Techniques for your SoC Design	Hardware	Hardware Verification
Rajesvari, M. *et al*	2013	System-on-Chip (SoC) for Telecommand System Design	Hardware	Chip Design
Krstic, S. *et al*	2014	Security of SoC Firmware Load Protocol	Firmware	Firmware Protection
Mentor	n.d.	Mentor Verification IP: Comprehensive verification IP built using advanced methodologies for fastest time to verification sign-off	Both	Verification
Ray, S. *et al*	2015	Security policy enforcement in modern SoC designs	Hardware	Policies
Backer, J. *et al*	2016	Secure and Flexible Trace-Based Debugging of Systems-on-Chip	Hardware/Firmware	Debugging
Ray, S. *et al*	2018	System-on-Chip Platform Security Assurance: Architecture and Validation	Both	Design and Verification
Wang, L. *et al*	2008	System-on-ChipTest Architectures:Nanometer Design For Testability	Hardware/Firmware	Testing Hardware/Firmware

References	Year	Title	Hardware/Software-Firmware	Specification
Gundabolu, S. *et al*	2018	On-chip Data Security Against Untrustworthy Software and Hardware IPs in Embedded Systems	Both	Malicious IPs
Elkeelany, O. *et al*	2006	Gaining Extra Crypto-Security using System on Chip Model for RC5	Both	Cryptography
Backer, J. *et al*	2015	Reusing The IEEE 1500 Design for Test Infrastructure For Security Monitoring of Systems-on-Chip	Both	Design and Architecture for Security
Li, H. *et al*	2016	A survey of hardware Trojan threat and defence	Hardware	Hardware Trojans
ARM®	2019	Building a Secure System using TrustZone® Technology	Hardware/Firmware	Chip Design
ACCEL-LERA	2011	Universal Verification Methodology (UVM) 1.1 User's Guide	Hardware/Firmware	Verification
Basak, A. *et al*	2015	A Flexible Architecture for Systematic Implementation of SoC Security Policies	Hardware/Firmware	Architecture and Policies

7. Conclusion

In conclusion it is very important for an SoC system in general to have security characteristics and to follow appropriate and correct standards and policies. It is always and has always been a good practice to follow globally accepted standards and define which policies to proceed with. The design process of security, and the decisions taken accordingly, highly depend on the imagination of the organisation. Some of the things may be obvious and underestimated or simply missed due to human factors.

It is extremely important to secure microcontrollers and SoCs (whichever is used in the system) because they are the inevitable edge nodes of IoT (Kumar et al, 2017). By all means, protecting these chips mean protecting the whole system.

Acknowledgements

We would like to thank Dominic Fennell for his proof reading.

References

ACCELLERA (2011) Universal Verification Methodology (UVM) 1.1 User's Guide.

ARM® (2009) Building a Secure System using TrustZone® Technology. ARM Security Technology.

Backer, J., Hely, D., Karri, R. (2014) Reusing The IEEE 1500 Design for Test Infrastructure For Security Monitoring of Systems-on-Chip. 2014 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT). Amsterdam: IEEE, pp. 52-56.

Backer, J., Hely, D., Karri, R. (2016) Secure and Flexible Trace-Based Debugging of Systems-on-Chip. ACM Transactions on Design Automation of Electronic Systems, 22 (2), Article-31.

Barrett, B. (2015) Wink's Outage Shows Us How Frustrating Smart Homes Could Be. [Online] Available at: https://www.wired.com/2015/04/smart-home-headaches/ [Accessed: 03/06/2019]

Bartley, M. (2017) Hardware Security Challenges and Solutions. TVS. [Online] Available at: https://www.testandverification.com/wp-content/uploads/TVS-Hardware-Security-Challenges-and-Solutions.pdf [Accessed: 21/05/2019]

Basak, A., Bhunia, S., Ray, S. (2015) A Flexible Architecture for Systematic Implementation of SoC Security Policies. 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). Austin, TX, USA: IEEE, pp. 536-543.

Davis, M. (2012) What is an SoC? Hint: the "S" stands for Server. ARM SERVERS [Online] Arm Ltd. Available from: https://armservers.com/2012/05/14/what-is-an-soc-hint-the-s-stands-for-server/ [Accessed: 29/03/2019]

Elkeelany, O., Olabisi, A. (2006) Gaining Extra Crypto-Security using System on Chip Model for RC5. Proceedings of the 38th Southeastern Symposium on System Theory. Tennessee Technological University, TN, USA.

Evans, D. (2011) The internet of things - how the next evolution of the internet is changing everything. White Paper. Cisco Internet Business Solutions Group (IBSG), 2011.

Gelsinger, P. (2006) Moore's Law - The Genius Lives On. *IEEE solid-state circuits society newsletter* September 2006. Available from:http://web.archive.org/web/20070713083830/http://www.ieee.org/portal/site/sscs/menuitem.f07ee9e3b2a01d06bb9305765bac26c8/index.jsp?&pName=sscs_level1_article&TheCat=2165&path=sscs/06Sept&file=Gelsinger.xml [Accessed: 29/03/2019]

Goguen, J., Meseguer, J. (1982) Security Policies and Security Models. IEEE Symposium on Security and Privacy, pp. 11–20.

Goodin, D. (2014) Password leak in WeMo devices makes home appliances susceptible to hijacks (updated). [Online] Available at: https://arstechnica.com/information-technology/2014/02/password-leak-in-wemo-devices-makes-home-appliances-susceptible-to-hijacks/ [Accessed: 03/06/2019]

GSM Association (2014) Understanding the Internet of Things (IoT). [Online] Available from: https://www.gsma.com/iot/wp-content/uploads/2014/08/cl_iot_wp_07_14.pdf [Accessed: 12/04/2019]

Gundabolu, S., Wang, X. (2018) On-chip Data Security Against Untrustworthy Software and Hardware IPs in Embedded Systems. 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 644-649.

Heideman, P. (2016) Early IoT Missteps - 10 Examples of Failure. [Online] Available at: https://www.omniresources.com/blog/early-iot-missteps-10-examples-of-failure [Accessed: 03/06/2019]

Jin, Y., Phatak, D.(ed) (2015) Introduction to Hardware Security. Electronics. Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, USA.

Li, H., Liu, Q., Zhang, J. (2016) A survey of hardware Trojan threat and defence. INTEGRATION, the VLSI journal, 55 (2016) , pp. 426–437.

Kim, L., Villasenor, J. (2015) Dynamic Function Verification for System on Chip Security Against Hardware-Based Attacks. IEEE Transactions on Reliability, 64 (4), pp. 1229-1242.

Krstic, S., Yang, J., Palmer, D., Osborne, R., Talmor, E. (2014) Security of SoC Firmware Load Protocol. IEEE HOST.

Kumar, S., Sahoo, S., Mahapatra, A., Swain, A., Mahapatra, K. (2017) Security Enhancements to System on Chip Devices for IoT Perception Layer. 2017 IEEE International Symposium on Nanoelectronic and Information Systems. Rourkela.

Maglaras, L. A., Kim, K. H., Janicke, H., Ferrag, M. A., Rallis, S., Fragkou, P., ... & Cruz, T. J. (2018). Cyber security of critical infrastructures. ICT Express, 4(1), 42-45.

Malik, S., Subramanyan, P. (2016) Specification and Modeling for Systems-on-Chip Security Verification. Department of Electrical Engineering, Princeton University, Texas, USA.

Mentor (n.d.) Mentor Verification IP: Comprehensive verification IP built using advanced methodologies for fastest time to verification sign-off. A Siemens Business. [Online] Available at: https://www.mentor.com/products/fv/verification-ip [Accessed: 22/05/2019]

Montador, A. (n.d.) Verification Methodology for Standards-based IP & SOC. Cadence Design Systems Inc, Livingston, Scotland. [Online] Available at: https://www.design-reuse.com/articles/13229/verification-methodology-for-standards-based-ip-soc.html [Accessed: 22/05/2019]

Nath, A., Ray, S., Basak, A., Bhunia, S. (2018) System-on-Chip Security Architecture and CAD Framework for Hardware Patch. IEEE. USA.

National Cable & Telecommunications Association - NCTA (2015) Behind The Numbers: Growth in the Internet of Things [Online] Available at: https://www.ncta.com/whats-new/behind-the-numbers-growth-in-the-internet-of-things [Accessed: 12/05/2019]

Nichols, M. (2019) Is System-on-Chip a viable solution for IoT security? [Online] Born2Invest. Available from: https://born2invest.com/articles/system-chip-viable-solution-iot-security/ [Accessed: 31/05/2019]

Nolan, S. (n.d.) Power Management for Internet of Things (IoT) System on a Chip (SoC) Development. Vidatronic, Inc. [Online] Available at: https://www.design-reuse.com/articles/42705/power-management-for-iot-soc-development.html [Accessed: 18/05/2019]

Palomar Technologies (n.d.) System on a Chip (SoC) [Online] Available at: https://www.palomartechnologies.com/applications/system-on-a-chip [Accessed: 11/05/2019]

Rajesvari, R., Manoj, G., Angelin Ponrani, M. (2013) System-on-Chip (SoC) for Telecommand System Design. International Journal of Advanced Research in Computer and Communication Engineering, 2 (3), pp. 1580-1585

Ray, S., Jin, Y. (2015) Security policy enforcement in modern SoC designs. 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). Austin, TX, USA: IEEE.

Ray, S., Basak, A., Bhunia, S. (2017) To Secure the Internet of Things, We Must Build It Out of "Patchable" Hardware. IEEE Spectrum. [Online] Available at:https://spectrum.ieee.org/telecom/security/to-secure-the-internet-of-things-we-must-build-it-out-of-patchable-hardware [Accessed: 12/05/2019]

Ray, S., Peeters, E., Tehranipoor, M., Bhunia, S. (2018) System-on-Chip Platform Security Assurance: Architecture and Validation. *Proceedings of The IEEE,* 106 (1), pp. 21-37.

Reichert, C. (2017) Cisco: Most IoT projects are failing due to lack of experience and security. [Online] Available at: https://www.zdnet.com/article/cisco-most-iot-projects-are-failing-due-to-lack-of-experience-and-security/ [Accessed: 03/06/2019]

Saleh, R., Wilton, S., Mirabbasi, S., HU, A., Greenstreet, M., Lemieux, G., Pande, P., Grecu, C., Ivanov, A. (2006) System-on-Chip: Reuse and Integration. Proceedings of the IEEE, 94 (6), pp. 1050-1069

Sanders, J. (2018) 5 Biggest IoT Failures in 2018. [Online] Available at: https://www.techrepublic.com/article/5-biggest-iot-security-failures-of-2018/ [Accessed: 03/06/2019]

Schubert, A., Anheier, W. (1999) Efficient VLSI Implementation of Modern Symmetric Block Ciphers. Proceedings of ICECS'99, Pafos, Cyprus 1999.

Sondrel (2016) Functional Verification Techniques for your SoC Design. [Online] Available at: http://blog.sondrel.com/functional-verification [Accessed : 22/05/2019]

Vita Development Wiki (2016) Secure World. [Online] Available at: https://wiki.henkaku.xyz/vita/Secure_World [Accessed: 02/06/2019]

Wang, L., Stroud, C., Touba, N. (2008) System-on-Chip Test Architectures:Nanometer Design For Testability. Burlington, USA: Morgan Kaufmann Publishers.

Influence of Attack Vectors on Generic Artificial Intelligence – assisted Smart Building Feedback Loop System

Petri Vähäkainu, Martti Lehto and Antti Kariluoto
University of Jyväskylä, Finland
petri.vahakainu@jyu.fi
martti.lehto@jyu.fi
anjuedka@jyu.fi

DOI: 10.34190/EWS.20.046

Abstract A Cyber-Physical system utilizes real-world measurements collected via IoT sensors. Using data on measurements and outside weather, it is possible to train a machine-learning model that can automatically adjust HVAC and other heating devices of smart buildings. With traditional systems, it is not possible to attain the required energy efficiency values. Instead, artificial intelligence could minimize these values. However, in case the perpetrator can influence the AI governing the feedback system, the cost may increase significantly. For example, the cost of power spikes has a long payback, which can be multiple years. In this article, we define the concepts of cybersecurity and machine learning, do a literary review on machine learning models, recognition of the relevant attack vectors to a feedback system of a smart building and countering of those attacks, and discuss the generic feedback system implementation.

Keywords: artificial intelligence-based applications, artificial intelligence, cloud service, data platform, attack vectors, adversarial attacks

1. Introduction

Cyber-physical systems (CPS), such as smart, intelligent buildings, utilizing artificial intelligence (AI) are becoming more common and widespread in the future. IoT smart sensors that can be automatically used to measure usage, functions, and variables describing the state of a smart building in the cyber-physical system, can provide a crucial dataset for creating machine learning (ML) models. Gathering relevant data following the security model, CIA (Confidentiality, Integrity, and Availability), is essential for the development of these models. The data gathered needs to be protected from unauthorized access, deletion, or modification from any unauthorized party, and the data must be available to use. Lacking or absent controls lead to insecure cyberspace. Properly designed and implemented predictive ML models can be productized and used in feedback systems to control the HVAC functions of a smart building.

Artificial intelligence can be used to detect threats and other malicious activities, while machine learning can process data inconspicuous to humans and be used to adjust hyperparameters of the AI model to yield desirable results. Companies specializing in cybersecurity are teaching AI systems to detect malware and viruses by using sophisticated algorithms and utilizing pattern recognition. Utilizing AI can also be beneficial in identifying behaviors of ransomware and malware attacks and function as a defense against cyber-attacks.

Patterns intrinsic both in the data and the model might expose models to the threat of adversarial attacks. An adversarial attack is an attack vector created using artificial intelligence. These attacks are adversarial disruptions constructed purposely by the attacker, which are imperceptible in the human eye but generally adversely impact to neural network models. These days, adversarial attacks towards machine learning models are becoming more and more common, bringing out noticeable security concerns. For example, in the context of smart building (CPS), an attacker may have a chance to deceive the ML model into causing harm, such as to create conditions for consumption spikes, when attacking the heating system guided by artificial intelligence-based feedback system.

In this paper, our purpose is to do a literature review of relevant machine learning models, present a simulated generic artificial intelligence-based feedback system and attack vectors impacting the system concerned. Our method, besides the literary review, is to discuss our findings as well as the strengths and weaknesses of our presented solution to tackle the threats towards the feedback system.

This paper is structured as follows: The first chapter is the introduction, where we presented common adversarial attack vectors towards the feedback system in the smart building's context. In chapter 2, we explain cybersecurity and cyber threats. In chapter 3, we present the background of artificial intelligence and its subset machine learning and continue presenting the review results of machine learning models. In chapter 4, we introduce a generic simulated feedback system case. In chapter 5, we talk about relevant attack vectors impacting the feedback system introduced. In conclusion, in chapter 6, we discuss our findings and conclude our paper.

2. Cybersecurity and cyber threats

Defining cybersecurity is challenging, and there does not exist a widely adopted definition of cybersecurity, even though the term itself is extensively utilized. On the one hand, Gartner defined cybersecurity as the combination of people, policies, processes, and technologies employed by an organization to protect its cyber assets. Cybersecurity is optimized to levels that business leaders define, balancing the resources required with usability/manageability and the amount of risk offset. Subsets of cybersecurity include IT security, IoT security, information security, and OT security. (Gartner) On the other hand, cybersecurity is a collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance, and technologies that can be used in protecting the organization's assets (Solms & Niekerk, 2013). Cybersecurity can be defined as a range of actions taken in defense against cyberattacks and their consequences and includes implementing the required countermeasures.

In the field of cybersecurity, there is a need to define the threat, vulnerability, and risk, which are not separate concepts, but intertwined. The risk lies in the intersection of assets, threats, and vulnerabilities, and it can be seen as a function of threats exploiting vulnerabilities to obtain, damage, or destroy assets, which can be critical resources. When conducting risk management, the familiar formula to determine the risk is A (asset) + threat (T) + vulnerability (V) = risk (R). A threat is anything that can exploit the vulnerability, accidentally or intentionally, to obtain, damage, or destroy an asset. Vulnerabilities mean weaknesses or gaps in the security program that can be exploited by threats to gain unauthorized access to an asset. Examples of intentional threats are, for example, malware, espionage, or privacy violation. (Flores et al., 2017)

Cybersecurity is built on the threat analysis of an organization or an institution. The structure and elements of an organization's cybersecurity strategy and its implementation program is based on the estimated threats and risk analyses. (Lehto, 2015, 3 - 29.) Organizations greatly benefit from well-made targeted cybersecurity strategies, and in many cases, they are necessary. A threat analysis identifies and models the relevant threats against assets, and a risk assessment classifies the impact and likelihood associated with each threat. Threats can be divided into natural and human-made threats, which can be further divided into accidental and intentional threats. Human error and software and hardware faults can be counted into accidental threats and outsider and insider threats into intentional threats.

Organizations need to develop and implement reliable and comprehensive defenses to face current and forthcoming cyber-attacks. Various kinds of research in the field of cybersecurity have been conducted, but the threat environment is constantly changing, causing endless and consuming race between defending organizations and perpetrators trying to penetrate the organization's information system's security. Therefore, all necessary preparations are important to be made to counter these threats and to protect, e.g., organizations from them. To address cyber threats efficiently, organizations may utilize cyber threat intelligence (CTI) to help them to identify, detect, and respond to threats. CTI enhances situational awareness and strategic responses helping to decrease potential risk towards organizations. Organizations can also increase common knowledge concerning cyber threats, improving operational capability, and maintaining security (Limnéll, Majewski & Salminen, 2014, 107.). A comprehensive cyber breach response plan helps minimize the impact of a cyber-attack after identifying the breach. Technologies, such as machine learning, are proven to be useful in detecting cyber threats based on analyzing data and then identifying a potential threat before the threat takes advantage of a vulnerability in the organization's information system.

3. Machine learning and artificial intelligence

The artificial intelligence algorithm is a function estimator. It can be expressed mathematically as f(x): Rn -> Rm, where f(x) is the function to be modeled, Rn represents the real input values, and Rm represents the possible real output values. (Vähäkainu, Lehto & Kariluoto, 2019) Artificial intelligence is an umbrella term for various other algorithmic models, such as decision trees, support vector machines (SVM), and neural networks, used to model the connections within data mathematically. These algorithmic models can be used for classification and regression, depending on the available data and the objective function to be estimated. Machine learning is a sub-domain of artificial intelligence, which aims to make AI systems more generalizable. ML is needed to adjust the AI system to deal with unexpected situations, such as handling data from a new input domain. Therefore, some algorithmic models used with ML are re-trained after a certain number of calculations or events.

These days, the neural networks model, also called an "artificial neural network" (ANN), is widely used as the trainable model. Commonly used ANN algorithms are perceptron, back-propagation, Hopfield Network, and Radial Basis Function Network (RBFN). ANNs require less statistical training and can detect complex nonlinear relationships between the predictor and response variables. Disadvantages are black-box nature, heavy computational burden, and tendency to overfitting. (Qolomany, Al-Fuqaha, Benhaddou & Qadir, 2019)

The ANN model is constructed from organized and connected layers, which are made of interconnected neurons (nodes). Straightforwardly expressed, the output of a neuron is a function of the weighted sum of the inputs plus a bias. In a neural network, the activation function, such as Rectified Linear Unit (ReLu), is responsible for transforming the summed weighted input from the node into the activation of the node or output for that input. (Brownlee, 2019) The complete neural network function is a computation of the outputs of all the neurons. ANN model is considered trained when no significant improvement is recorded between the output values of the model and the target values. Some other performance metrics used to define the learning algorithms might also be employed. ANN's have been used in smart home context, e.g., to measure energy efficiency in pattern recognition. Badlani and Bhanot (2011) developed a smart home system in which they utilized the recurrent neural network (RNN) to capture human behavior patterns and an ANN for security applications to be used in smart homes. ANNs have also been used to calculate the probability of occupation for different areas of the building and comparing the probability with the current situation. (Qolomany et al., 2019)

Long Short-Term Memory (LSTM) neural networks are a variation of recurrent neural networks able to learn from experience to classify, process, and predict, for example, stock market or cyber-physical system time series. The ability of LSTM to learn patterns in data over long sequences makes them suitable for time series forecasting (Jain, 2019). LSTMs possess the ability to remember information for a long time due to their inner cell's ability to carry information unchanged. LSTMs have been able to model temporal sequences and their long-range dependencies more accurately than an ordinary RNNs. The network completely controls the cell state and is capable of adding, edit, or remove information in the cell using gates. According to Gasmi, Laval & Bouras (2019), LSTMs were developed by Hochreiter and Schmidhuber in 1997 (Hochreiter & Schmidhuber, 1997) to address the problem of long-term dependency learning, and to explore vanishing and exploding gradient problems of an ordinary RNNs. Conventional RNNs perform well with short sequences, but long sequences cause challenges to training RNNs, especially if the number of model parameters becomes too high. At the time, the training is improbable to converge. (Gasmi, Laval & Bouras, 2019)

LSTM and conventional RNN are the most noteworthy deep neural network models, which provide a feedback loop from the past inputs. According to Nugaliyadde, Somaratne & Wong (2019) research recurrent neural networks mentioned are proven to perform better than other deep neural networks (DNN) models lacking feedback loops, such as feed-forward neural networks. Some of the other DNN networks are complex and computationally demanding. In the case of, e.g., forecasting electricity consumption of homes or office buildings, traditional models, such as ANN, SVM, and Fuzzy logic, provides means mainly for short term forecasting. Recurrent neural networks, such as RNN or LSTM, are shown to perform better and can provide forecasting for short-term, mid-term, and long-term forecasting with decent accuracy. LSTM can also be used to predict general weather variables precisely. In the future, it may replace humans and traditional methods in weather forecasting.

Clustering is a well-known unsupervised machine learning method, where data is automatically grouped based on the similarity of the data points and dissimilar to the data points in other groups. In clustering methods, data points can be grouped into clusters based on their distance to the center of the cluster. In the case of data points located outside of a certain cluster group, it can be assumed to be an anomaly (Atat, Liu, Wu, Li, Ye & Yi, 2018). Clustering can be considered to be a kind of a compilation of objects based on similarity or dissimilarity between them. In the unsupervised learning process, references from datasets are drawn in order to find a meaningful structure. The most common clustering algorithm to carry out the benchmarking is K-means, iteratively looking for a local maximum. According to Lin, Lin, Liu & Tseng (2019), K-means have been utilized in smart home context, e.g., in solving optimization problems, such as optimal control of HVAC power consumption or the prediction of energy consumption through generating of energy consumption models and improving HVAC power consumption. In order to improve HVAC power consumption prediction, clustering the data gathered from IoT sensors and combined utilization with naïve Bayes classifier to classify the data have been conducted in Lin et al. study. According to the study, enormous potentialities of energy savings, even with basic energy retrofit actions, can be achieved.

Decision trees, belonging to the family of supervised learning, are one of the most efficient and popular tools to execute, i.e., pattern recognition, classification, and prediction. The target of using decision trees is to generate a training model to be used in predicting the class value of the target variables. The process includes learning simple decision rules concluded based on previous training data. Chauhan (2019) The three resembles the structure of a flowchart in which "tests" on attributes are represented by nodes, class labels representing numerical data by leaves, and the outcome of the test by branches. Decision trees are used to learn logical descriptions of activities from complex sensor readings (Pal & Mather, 2001; Stankovski & Trnkoczy, 2006). According to Sánzhez & Skeie (2019), they have also proved to be useful in human activity recognition (HAM) modeling in smart building environments resulting in an accuracy rate of 88.02 % in the field tests.

Support vector machines (SVM) are one of the best and highly robust supervised machine learning models. SVM's have extensively been enacted to pattern classification problems and nonlinear regressions (Chu, Jin & Wang, 2005). According to Seyedzadeh, Rahimian, Glensk & Roper (2018), in some instances, when there are no gathered historical data available, SVMs can be used because they can be trained with only a few numbers of data samples. SVM's have been used in building context to predict cooling and heating loads, electricity consumption, and classification of energy usage of buildings. For example, in Singapore, SVM's have been used to predict monthly electricity usage by using datasets gathered within three years, including input parameters, such as temperature, humidity, and solar radiation. The trained model based on data gathered resulted in excellent accuracy in predicting the electronic loads with a meager error rate of only 4 %. Sretenovic, Jovanović, Novakovic& Nord's (2018) study states that the SVM is generally understated in solving prediction problems. The SVM is comparable to other even more innovative ML models, and in many cases, it can outperform them.

Genetic algorithms (GA), developed by John J. Holland, are one of the most generally utilized optimization algorithms applicable to cyber-physical systems, such as smart buildings. This lead to Holland's book "Adaption in Natural and Artificial Systems," published in 1975. The algorithm is designed for solving both unconstrained and constrained optimization problems based on natural selection, analog to Charles Darwin's evolution theory. The algorithm carries out the iterative process in which the most suitable individuals of the population are chosen to produce offspring to create the new generation. The genetic algorithm randomly selects individuals for the population concerned at each iteration. Then, individuals selected will become parents producing children for the next generation. After the generations, the population will reach the optimal solution. (Mathworks, 2019)

Building operation optimization utilizing machine learning models can decrease building's energy consumption, provide cost savings, and also amend indoor environment quality. According to the research, buildings are estimated to consume up to 40 % of primary energy in the USA and China, and around 30 % in the rest of the world. Building energy consumption could be decreased by 30 % - 40 %, but it is challenging due to dependent factors such as weather, and inhabitant behavior. (Li, Shao, Zuo & Huang, 2017) According to the Shaik, Nor, Nallagownden, Elamvazuthi& Ibrahim (2016) research, one of the challenges is how to make the indoor environment of a building comfortable while improving energy efficiency. Genetic algorithms, such as multi-objective genetic algorithm (MOGA) and hybrid multi-objective genetic algorithm (HMOGA), have been used as optimization algorithms to provide substantial enhancements in energy efficiency and indoor comfort in smart buildings. The utilization of the HMOGA technique provided an energy efficiency of 31.6 %, with an 8.1 % improved comfort index.

According to Parra, Tauler, Veng, Ligeza, Gonzalez & Aguilo (2015), Bayesian network (BN) can be considered as an established framework for managing uncertainty in artificial intelligence taking advantage of graph and probability theories in describing the relationship between nodes concerned. BN modeling can extract probabilistic influences from big data sets. Bayesian networks are capable methods to be utilized in uncertain and labile situations. BSs are useful as they are adaptable; therefore, they can self-modify their modeling structure, specific instances, or parameters. Constructing a preliminary network with scarce initial information is a possible scenario. The network can be augmented when gaining new information. Learning from experience is the principal capability that BN networks can provide. In addition, to be able to deal with uncertainties, they are potent methods clinching complex diagnostic and prediction issues in incidents where the information is transferred from multiple various kinds of sources. In order to reach that, Bayesian Networks can merge preceding information with an empirical one. (Ferreiro, Sierra, Gorritxategi & Irigoien, 2010) Bayesian networks have also been used in the field of smart buildings. For example, Rohan (2015) conducted research, which covered utilizing Bayesian networks for forecasting heat load in a district heating system, reaching an average accuracy of 77.97% with just 10% of the data for a forecast time window of an hour. Utilized parameters were supply temperature, return temperature, flow rate, the difference between supply and return temperature, outdoor temperature forecast, behavioral parameters, and current heat load consumption.

4. Simulated ML–based generic feedback system

On an abstract level, cyber-physical systems can be divided into the computing-based guiding part and the physical part that includes the structure, sensors, and actuators. In our context, the structure refers to smart building itself, sensors are physical measurement units, and actuators are physical devices, such as motors or switches. Depending on the calculating power and energy consumption tolerance, as well as the frequency of measurements, the cyber part of the CP system may be calculated in local physical units of the system or the cloud. Neural networks that are often used in the field of artificial intelligence can be used in both.

In our simulated case, we focus on artificial intelligence-based control of smart ice hockey and sports building complex for the reduction of energy spikes. The massive building has multiple temperature sensors within different sections of the complex, and interface for altering the system parameters on a computer. The heating of the complex is done by utilizing the heat grid and collecting used hot water from the showers, while the main heating comes from the cooling process of the indoor ice hockey rink. The sports building is open for the public between 7 am and 11 pm, and collects information, such as CO2 measurements within the building's confinement and social media data from volunteers.

Our generic feedback loop for the smart building control is shown in Figure 1. Internal data is collected from the local measurement units. Data is enriched with external data that is extracted from open source weather service and the social media accounts of the volunteers. Data are firstly cleaned and merged. Before and after the merger, data goes through a cybersecurity AI model, which checks for any malicious inputs, such as adversarial attacks, and tries to recognize and mitigate possible DDOS attacks by refusing connections. Previous data and commands are added into data, and data is sent to the guiding ML model. The guiding model should be an ensemble model that utilizes, therefore varying levels of information for making decisions. For example, an LSTM neural network could be used to calculate the previous prediction errors of temperature, and a type of deep-learning model could be used to recognize certain situations that are interesting from the building utilization standpoint, such as sudden spikes in CO2 levels of smaller rooms. At this point, the model may use additional information to calculate the prediction error. A retraining event is emitted if the error is higher than some preset value. Many enough triggered events cause the model to begin a separate process in which the retraining of a new ML model is done. This allows our model to learn by adjusting its hyperparameters or structure, but learning can happen sparingly. It is necessary to retrain our model because buildings are rarely exactly similar and used identically.

The same ensemble model calculates new predictions in the next phase of our model, namely, calculating new predictions. Based on the predictions, new commands are calculated, which are then logged into a database. After this, new commands are sent to the actuators.

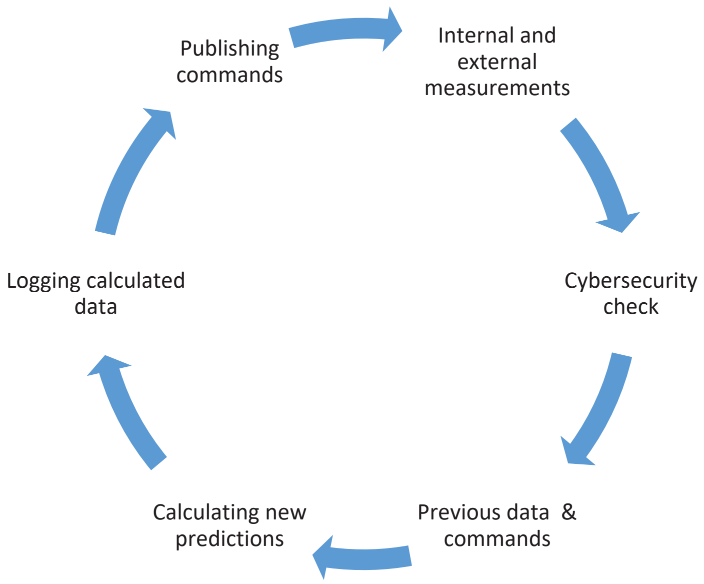


Figure 1: Generic artificial intelligence control loop in smart building's context

5. Attack vectors and defense methods impacting the feedback system

In these days, deep learning and neural networks is a hot topic around the world. LSTMs are considered to be more efficient in classifying sequential or time-series data. LSTMs, as well as other recurrent and convolutional neural networks, are vulnerable to adversarial examples. Adversarial examples are misclassified samples and are generally unnoticeable to human eyes. Those misclassified samples can be generated, e.g., by adding minor perturbations, such as flawed pixels, on the images to form an image classification problem and to be used to deceive sophisticated deep neural networks in the testing or deploying stage. The vulnerability of adversarial examples is ample and ever-growing risk especially when the field of critical infrastructure is concerned. Fooling the predictive deep neural network used to adjust the HVAC system of a cyber-physical system can cause challenging situations in the form of energy consumption spikes causing increasing costs.

Classification-oriented LSTM utilized in implementing a generic feedback system can be attacked, for instance, by using the Jacobian-based Saliency Map Approach (JSMA). JSMA, proposed by Papernot, McDaniel, Wu, Jha & Swami (2016), is an iterative method for targeted misclassification. JSMA can be seen as a matrix of forward derivatives for every feature of a sample with respect to the output classes. The method calculates the forward derivates of the neural network output with respect to the input example (Wang, Li, Wen, Nepal & Yang, 2015). It then perturbs the input in the desired direction to selectively make the model misclassify to an appropriate output class (Anderson, Bartolo & Tandon, 2016). Deep neural networks can be fooled by adding even minor perturbations to the input vector. For example, in the case of image perturbation, the method can effectively produce samples that are correctly classified by humans but successfully misclassified into a specific target class by a DNN.

The JSMA attack can cause the predictive model to output more erroneous predictions, which can, eventually, make the controlling model either complacent or too reactive. Both choices could be monetarily crippling. For example, Papernot, McDaniel, Ananthram, Swami & Harang (2016) were able to perturb both categorical and sequential RNNs with JSMA adversarial attack. Therefore, the chance exists that the perpetrator could, if given enough time and resources, afflict damage to both AI models, namely the cybersecurity AI model and the controlling AI model.

Defensive distillation is a successful adversarial defense method against relatively well-known adversarial attacks, such as Fast Gradient Sign Method (FGSM) and the Jacobian-based Saliency Map Approach (JSMA). Hinton, Vinyals & Dean (2015) originally introduced the defensive distillation in "Distilling the Knowledge in a Neural Network as a technique for model compression." The defensive distillation can be considered as a strategy, which includes training the model to output probabilities of different classes. Probabilities are then fed by the previous model that is trained on the same task and using hard class labels. (Goodfellow, Papernot, Huang, Duan, Abbeel & Clark, 2017) The model adds flexibility to an algorithm's classification process in order to make the model more robust to perturbations and less vulnerable to exploitation. Utilization of defensive distillation significantly decreases the success rate of the adversarial crafting process, making it a remarkably effective model against adversarial attacks, such as JSMA.

However, according to Soll, Hinz, Magg & Webmter (2019), defensive distillation solely has a minor effect on text classifying neural networks, and it does not provide means to improve their robustness against adversarial examples. These days, there exist attack strategies on how to initiate a successful attack towards ML models. One of the strategies to launch a working attack is to utilize the so-called black-box attack in which the adversary trains a substitute model mimicking the defended model and taking advantage of the substitute model in generating adversarial examples that transfer back to the distilled model. The attack method concerned has a high success rate due to the wide transferability of adversarial examples between models trained and designed to be used in clinching the same ML problem. Defense distillation can also be evaded with optimization attacks, utilizing the phenomenon of gradient masking. This method affects the heuristics of FGSM or JSMA attacks by obliterating gradients used by these attacks. (Papernot & McDaniel, 2017)

Adversarial training is one of the most distinguished defense methods to counter adversarial attacks, and it has reached the de-facto standard in creating robust models (Stutz, Hein & Schiele, 2019). According to Samangouei, Kabkab & Chellappa (2018), adversarial examples with their correct labels, which include one or more chosen attack models, can be added to the training set in order to create the augmented training set. The training process is an iterative process running through multiple iterations. Then, adversarial examples are crafted via, e.g., FGSM or another method to be added to the dataset, and the next iteration begins. Augmented training set increases robustness in case of the generated augmented training set is the same as a perpetrator uses. If a perpetrator uses a different kind of attack strategy, the efficiency of the adversarial training will decrease. Due to gradient masking, the generated model affects more to white-box than black-box –based attacks. From the perspective of traditional security, adversarial training can be seen as a type of penetration testing that helps to detect vulnerabilities in a model, which are then fixed by utilizing the training process (Kurakin, Goodfellow & Bengio, 2018).

The adversarial training method is an efficient process against FGSM attack vectors, but not against iterative vectors, such as JSMA. Therefore, original single-step adversarial examples cannot defend against iterative adversarial examples, but adversarial training with iterative adversarial examples can defend against iterative adversarial examples, but computational requirements are too steep, and it is not scalable. One well-known defense method against JSMA is input gradient regularization, which uses double backpropagation, which trains neural networks by minimizing the energy of the network and rate of change of the energy with respect to the input features (Ross & Doshi-Velez, 2017).

6. Discussion and conclusion

Our generic feedback loop capable of supporting AI models was designed with the avoidance of energy spikes in mind. Within the control loop structure, there exist multiple direct and indirect use-cases for ML models, e.g., the collection of inputs, cybersecurity control, prediction of missing values, forecasting horizon calculations for different features, and event-based machine learning.

In order to protect artificial intelligence controls of the smart building against cybersecurity threats besides insisting on the use of privileged nodes and encrypted communications, it might be beneficial to separate the cybersecurity AI model and the controlling AI model. Retraining both models based on events, such as after a significantly grown prediction errors, might also help even though retraining can be time and resource consuming. Adversarial training with adversarial examples, adversarial distillation, and adversarial noise removal can make it more difficult to attack the smart building's artificial intelligence-based control loop. Using ensembles of ML models trained with the techniques as mentioned earlier together with the rotation of the protective ensembles increases the difficulty level of finding the control models hyperparameter values. That said, the adversarial training method is an efficient process against FGSM attack vectors, but not against iterative vectors, such as JSMA.

References

Anderson, M., Bartolo, A. & Pulkit, T., 2016. Crafting Adversarial Attacks on Recurrent Neural Networks. Cornell University, arXiv:1604.08275v1 [cs.CR].

Atat, R., Liu, L., Wu, J., Li, G., Ye, C. & Yang, Y., 2018. Big Data Meet Cyber-Physical Systems: A Panoramic Survey. IEEE Access, 6, pp. 73603-73636. doi:10.1109/ACCESS.2018.2878681

Badlani, A., Bhanot, S., 2011. Smart home system design based on artificial neural networks. Proc. World Congr. Eng. Comput. Sci., pp. 106-111, Oct. 2011.

Brownlee, J., 2019. A Gentle Introduction to the Rectified Linear Unit (ReLu). Accessed 6.1.2020 https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks.

Chauhan. 2019. Decision Tree Algorithm. Accessed 29.12.2019 https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4.

Chu, F., Jin, G. & Wang, L., 2005. Cancer Diagnosis and Protein Structure Prediction Using Support Vector Machines. StudFuzz 177, 343 – 363, Springer-Verlag Berlin, Heidelberg, Germany.

Ferreiro, S., Sierra, B., Gorritxategi, E. &Irigoien, I., 2010. Bayesian Network for Quality Control in the Drilling Process. IFAC Proceedings Volumes (IFAC-PapersOnline), Vol 43, Issue 4, 264 - 269.DOI: 10.3182/20100701-2-PT-4011.00046.

Flores, C., Guasco, T. & Leon-Acurio, J., 2017. A Diagnosis of Threat Vulnerability and Risk as IT Related to the Use of Social Media Sites When Utilized by Adolescent Students Enrolled at the Urban Center of Canton Canar. International Conference on Technology Trends, 199 – 214.

Gasmi, H, Laval, J. &BourasAbdelaziz., 2019. Information Extraction of Cybersecurity Concepts: An LSTM Approach. Journal of Applied Sciences, 9(19), 3945. DOI: 10.3390/app9193945.

Gartner. Cybersecurity. Accessed 11.12.2019 https://www.gartner.com/en/information-technology/glossary/cybersecurity.

Goodfellow, I., Papernot, N. Huang, S., Duan, R., Abbeel, P. & Clark, J., 2017. Attacking Machine Learning with Adversarial Examples. Accessed 20.1.2019 https://openai.com/blog/adversarial-example-research

Hinton, G., Vinyals, O. & Dean, J., 2015. Distilling the Knowledge in a Neural Network. Cornell University, https://arxiv.org/abs/1503.02531.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

Jain, V., 2019. Doing Multivariate Time Series Forecasting with Recurrent Neural Networks Using Keras' Implementation of Long-Short Term Memory (LSTM) for Series Forecasting. Accessed 14.1.2019 https://databricks.com/blog/2019/09/10/doing-multivariate-time-series-forecasting-with-recurrent-neural-networks.html.

Kurakin, A., Goodfellow, I., J. & Bengio, S. (xxxx) Artificial Intelligence Safety and Security. Adversarial Examples in the Physical World. Artificial Intelligence and Robotics Series, Chapman & Hall/CRC.

Lehto,M., 2015. Phenomena in the Cyber World. M. Lehto & P. Neittaanmäki (Edit.) Cyber Security: Analytics, Technology and Automation. Berlin: Springer.

Li, T., Shao, G., Zuo, W. & Huang, S., 2017. Genetic Algorithm for Building Optimization: State-of-the-Art Survey. ICMLC 2017: Proceedings of the 9th International Conference on Machine Learning and Computing. DOI: 10.1145/3055635.3056591.

Limnéll, J., Majewski, K. &Salminen, M., 2014. Kyberturvallisuus. Saarijärvi: Docendo.

Lin, C-M., Lin, S-F, Liu, H-Y. & Tseng, K-Y. Applying Naïve Bayes Classifier to HVAC Energy Prediction Using Hourly Data. Microsystems Technologies, 1 – 15, Springer, Berlin, Heidelberg. DOI: 10.1007/s00542-019-04479-z.

Mathworks. 2019. What is the Genetic Algorithm? Accessed 6.1.2019 https://www.mathworks.com/help/gads/what-is-the-genetic-algorithm.html.

Nugaliyadde, A., Somaratne, U. & Wong, K., W., 2019. Applied Energy, vol. 212, 372 – 385. Predicting Electricity Consumption Using Deep Recurrent Neural Networks. DOI: 10.1016/j.apenergy.2017.12.051.

Pal, M. & Mather, P., M., 2001. Decision Tree Based Classification of Remotely Sensed Data. Proceedings of 22nd Asian Conference on Remote Sensing, 5 – 9 November 2001, Singapore.

Papernot, N. & McDaniel, P., 2017. Extending Defensive Distillation. Cornell University, arXiv:1705.05264v1 [cs.LG].

Papernot, N., McDaniel, P., Swami, A., Harang, R., 2016. Crafting Adversarial Input Sequences for Recurrent Neural Networks. In Military Communications Conference, MILCOM 2016-2016 IEEE. IEEE,49–54.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A., 2016. Distillation as a defense to adversarial perturbations against deep neural networks. Symposium on Security and Privacy, 582–597.

Parra, P., Tauler, P., Veng, M., Ligeza, A., Gonzalez, A., & Aguilo, A., 2015. Bayesian Network modeling:A case study of an epidemiologic system analysis of cardiovascular risk. Elsevier. DOI:10.1016/j.cmpb.2015.12.010.

Qolomany, B., Al-Fuqaha, A., Benhaddou, D. & Qadir. J., 2019. Leveraging Machine Learning and Big Data for Smart Buildings: A Comprehensive Survey. IEEE Access 7, 90316 – 90356. DOI: 10.1109/ACCESS.2019.2926642

Rohan, N., 2015. A Bayesian Approach for Forecasting Heat Load in a District Heating System. Master's thesis. Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology.

Ross, A., S. & Doshi-Velez, F., 2017. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients. Cornell University, arXiv:1711.09404 [cs.LG].

Samangouei, P., Kabkab, M. and Chellappa, R., 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605.

Sánzhez, V. G. & Skeie, N-O., 2018. Decision Trees for Human Activity Recognition Modelling in Smart House Environments. Proceedings of the 59th Conference on Simulation and Modelling (SIMS 59), 26 – 28 September 2018, Oslo Metropolitan University, Norway.

Seyedzadeh, S., Rahimian, F., P., Glensk, I. & Roper, M., 2018. Machine Learning for Estimation of Building Energy Consumption and Performance: a Review. Visualization in Engineering, 6(5). DOI: 10.1186/s40327-018-0064-7.

Shaikh, P., H, Nor, B., M., Nallagownden, P., Elamvazuthi., I. Ibrahim, T., 2016. Intelligent multi-objective control and management for smart energy efficient buildings. Journal of Electical Power and Energy Systems, 74, 403 – 409. DOI: 10.1016/j.jksues.2016.03.001.

Soll, M, Hinz, T., Magg, S. & Wermter, S., 2019. Evaluating Defensive Distillation for Defending Text Processing Neural Networks Against Adversarial Examples. ICANN 2019: Image Processing. International Conference on Artificial Neural Networks, Artificial Neural Networks and Machine Learning, vol 11729, 685 - 696.

Solms, V., R., Niekerk, V., J., 2013. From information security to cyber security. computers & security, 38, pp.97-102.

Stankovski, V. and Trnkoczy, J., 2006. Application of decision trees to smart homes. In Designing smart homes (pp. 132-145). Springer, Berlin, Heidelberg.

Stretenogic, A., Jovanović, R, Novakovic, V. & Nord, N., 2018. Support Vector Machine for the Prediction of Heating Energy Use. Thermal Science, 2018, 126 – 126. DOI: 10.2298/TSCI170526126S.

Stutz, D., Hein, M. & Schiele, B., 2019. Confidence-Calibrated Adversarial Training and Detection: More Robust Models Genralizing Beyond the Attack Used During Training. University of Cornell, arXiv: 1910.06259v2 [cs.LG].

Vähäkainu P, Lehto M, Kariluoto A., 2019. IoT –based Adversarial Attack's Effect on Cloud Data Platform Service in Smart Building's Context. ICCWS 2020.

Wang, D., Li, C., Wen, S., Nepal, S. & Xiang, Y., 2015. Defending Against Adversarial Attack Towards Deep Neural Networks via Collaborative Multi-task Training. Cornell University, arXiv:1803.05123 [cs.LG].

Assessment of the French and Dutch Perspectives on International Law and Cyber-Operations

Brett van Niekerk[1], Trishana Ramluckan[1] and Daniel Ventre[2]
[1]University of KwaZulu-Natal, South Africa
[2]CESDIP, CNRS, France
vanniekerkb@ukzn.ac.za
ramluckant@ukzn.ac.za
daniel.ventre@cesdip.fr

DOI: 10.34190/EWS.20.029

Abstract: Cyber-operations have altered the nature of war and the applicability of international law to this new form of conflict has been widely debated; however, no clear consensus in this regard has been reached. Challenges that contribute to the uncertainty of international law's applicability to cyber-operations include the difficulty of attribution of cyber-attacks, and translating state sovereignty into Cyber-space. Academic and advocacy groups have provided publications with proposals for international law and cyber-operations; however, these documents are non-binding. In 2019, both France and the Netherlands released their official perspectives on international law and cyber-operations. This paper compares and assesses these two national documents using the NVivo software to conduct qualitative document analysis. The analysis highlights challenges in applying international law to cyber-operations, and illustrates the similarities and contradictions in the national perspectives as compared to each other and the guidelines set out by previous authoritative documents.

Keywords: Attribution, cyber-law, cyber-operations, international humanitarian law, sovereignty

1. Introduction

Warfare and conflict have evolved through the use of cyber-operations, and there has been a marked increase in possible state-backed operations in cyberspace since the mid-2010's, including allegations of the Ukrainian power grid being affected (Greenberg, 2017), interference in the 2016 US presidential elections (DNI, 2017), reported airstrikes in retaliation for persistent cyber-attacks (Fingas, 2019), and reported cyber-attacks in retaliation for physical attacks (Doffman, 2019). The International Institute for Strategic Studies (2018: 6) states that:

> *Cyber capability should now be seen as a key aspect of some states' coercive power, giving them the chance to wage covert digital campaigns. This might be an adjunct to military power, or employed in its place, in order to accomplish traditional objectives. This has driven some European states to re-examine their industrial, political, social and economic vulnerabilities, influence operations and information warfare, as well as more traditional areas of military power.*

The growing prevalence of cyber-operations in international relations gives rise to the need of assessing the applicability of international law to this new form of conflict. Whilst the relevance of international law to cyber-attacks and cyber-operations has been widely debated, there is yet to be an official agreement on this matter. Examples of challenges include the need for attribution which is difficult in cyber-space, and uncertainty of how concepts such sovereignty and use of force apply in a virtual context. The most in-depth study resulted in two publications that provide guidelines and suggestions for the application of international law in cyber-space: the Tallinn Manual on the International Law Applicable to Cyber Warfare (Schmitt, 2013) and the Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Schmitt, 2017). In 2015 a United Nations document examined the impact on growing information technology on national security (United Nations, 2015). In 2019, France and the Netherlands released their official national stance on this matter, International Law Applied to Operations in Cyberspace (Ministère des Armées, 2019) and International Law in Cyberspace (Netherlands Parliament, 2019b), respectively. These are the first national stances that have been published, illustrating foreign perspectives on cyber-operations.

The two national documents have been the subject of numerous readings and comments (in the media, on the Internet, on specialized websites) but academic works are necessary to methodically analyse the postures of the States. This paper aims to: (1) to identify the importance placed on various concepts of international law by the two national documents; (2) to investigate the relationship amongst the various legal aspects; (3) to identify challenges raised by the international documents, and any measure to mitigate these; and (4) to assess the alignment of the national documents to existing authoritative (but compliance is not mandatory) documents. To achieve these aims, the NVivo software is used to conduct qualitative document analysis.

The paper continues in Section 2, which presents a background to the previous efforts assessing international law and cyber-operations. The methodology is described in Section 3, and the results of the document analysis is presented in Section 4. A summary of the results is provided in Section 5, followed by the conclusion in Section 6.

2. Background

There has been extensive academic and professional debate regarding the application of international law in cyber-space. A primarily academic enquiry soliciting input from a panel of experts produced two outputs, being the two *Tallinn Manuals* referred to in Section 1 (Schmitt, 2013; Schmitt, 2017). These guidelines often indicate the views of the experts, and differences of interpretation or opinions that exist, indicting ongoing uncertainty in how international laws can be translated into a virtual world. In November 2018 the *Paris Call for Trust and Security in Cyberspace* (2018) was released, however this aligns more to the combatting of cyber-crime, but does refer to the Budapest Convention which sets out mechanisms for law enforcement operations in cyber-space that could occur across national boundaries (Council of Europe, 2001). In December 2018, the Global Commission on the Stability of Cyberspace released the *Singapore Norm Package*, which focusses on the responsibilities of states to ensure a safe, secure and available Internet (GCSC, 2018).

These documents and related legal concepts have undergone commentary and further academic discussion. Kilovaty (2014) further investigated the challenges with *Jus ad Bellum* in cyber-space with reference to the first *Tallinn Manual*; Student n°2222171 (2016) considered this in terms of the attempts at regulation and possible humanitarian impact. Eilstrup-Sangiovanni (2018) motivates for a dedicated international convention on cyber-warfare. Lindsay (2015) specifically focusses on attribution and deterrence in cyber-space, and Lotrionte (2012) considers the concept of state sovereignty and self-defence online.

The Netherlands national perspective that is considered is an appendix to a letter dated 5 July 2019, sent to the Netherlands Parliament by their Foreign Minister (Netherlands Parliament, 2019a; Netherlands Parliament, 2019b). The letter and the appendix were in response to requests regarding initiatives to strengthen international law in cyber-space and emerging from an inquiry into espionage attributed to Russia (Netherlands Parliament, 2019a). The Dutch document was published publically on the 26 September 2019, after the French published the document *International Law Applied to Operations in Cyberspace* (*Droit international appliqué aux operations dans le cyberespace*) on the 9 September (Ministère des Armées, 2019; Roguski, 2019). Roguski (2019) considers this document a consolidation of the French perspective exhibited in three documents related to defence and cyber-security from 2013, 2017 and 2018: the *White Book on Defense and National Security* (Livre blanc sur la défense et la sécurité nationale), *International Cyber Strategy* (Stratégie internationale de la France pour le numérique), and the *Strategic Review of Cyberdefense* (Revue stratégique de cyberdéfense), respectively.

3. Methodology

Qualitative document analysis of the two national documents is conducted, in particular content analysis and thematic analysis of the documents using word frequencies, coding, and clustering. The software NVivo was used to conduct the analysis as it provides the functionality to perform all the analytic methods.

For the coding, a total of thirteen codes (also called nodes in NVivo) were used. Two codes are related to the challenges and uncertainty, and a further nine conceptual codes (relating to major themes of international law) were identified from the previous documents, including: armed conflict; attribution; human rights; non-intervention; operations outside of conflict; response/retaliation; responsibility and due diligence; sovereignty; and use of force. A further two conceptual codes (child nodes) were unidentified whilst coding the documents: cyber-weapon, and pre-emptive response. Only content of the documents directly related to the application of international law to cyber-operations was considered; background information and challenges regarding international law in general were excluded from the coding. Clustering and Pearson's Correlation were used to determine relationships amongst the codes, and relationships between the national documents and the previous relevant documents. The previous documents considered include the two Tallinn Manuals (Schmitt, 2013; Schmitt, 2017), the Singapore Norms (GCSC, 2018), and the Paris Call (2018). The latter two documents are considered as they were released within 12 months of the two national documents, therefore will have considered current events. Word frequencies, visualised by word clouds, are used to illustrate major concepts in the various documents and coding. In these instances, common words and obvious words (international, law, cyber) were excluded.

4. Analysis of the National Documents

This section presents the results of the analysis. The section is broken into four sub-sections: Section 4.1 provides a summary of the analysis of the national documents; Section 4.2 focuses on the conceptual codes; Section 4.3 considers the uncertainty and challenges; and Section 4.4 provides an assessment of the alignment between the two national documents and previous authoritative documents.

4.1 Summary of the Analysis of the National Documents

A high-level summary of the two national documents is provided in Table 1. As is evident, the French document is considerably larger (twice that of the Dutch document), which corresponds to the number of text references to the two documents. The percentage coded and the number of codes (nodes coding) are similar for the two documents; the two extra coding nodes in the French document are the two identified during analysis (the pre-emptive response and cyber weapons). The Pearson Correlation between the two documents in terms of word similarity is 0.36, indicating some similarity does exist. It should be noted that the French document has a focus on cyber-operations specifically, whereas the Dutch document has a broader focus on cyber-space, as indicated by their respective titles.

Table 1: Summary of Documents

Document	Pages	Words	Paragraphs	Nodes Coding Source	Coded Percentage	Text References
France	20	11568	323	13	0.2840	157
Netherlands	9	6459	121	11	0.2644	67

Table 2 provides a comparison of the coding per source. Due to the French document being larger, there is a greater degree of coding in that document; however, the coding references and coded words for sovereignty and non-intervention in the Dutch document is greater, and the coding (references and words) is close between the documents for use of force and uncertainty. The two codes unique to the French document are again illustrated. The focus of the French document on cyber-operations and in particular cyber-warfare is evidenced by the coding for armed conflict.

Table 2: Comparison of Coding per Source

	France		Netherlands	
	Coding References	Words Coded	Coding References	Words Coded
Armed conflict	49	3619	4	245
Cyber-weapon	5	292	0	0
Attribution	13	780	9	684
Challenges	7	352	2	38
Human rights	13	994	4	314
Non-intervention	2	100	4	239
Operations outside of conflict	2	86	2	45
Response, retaliation	26	1524	12	1032
Pre-emptive	1	65	0	0
Responsibility and due diligence	12	715	3	228
Sovereignty	11	585	13	638
Uncertainty	3	91	3	92
Use of Force	13	833	12	784

Figure 1 provides a comparison of the word frequencies of the two documents, visualised using word clouds. The prevalence of the words align to the theme of the document titles. The focus on cyber-operations and cyberwarfare in the French document is evident as the word ‘armed’ is the most prevalent, with words such as ‘military’, ‘cyberattack’, ‘conflict’ and ‘effects’ appearing relatively frequently. In comparison, the Dutch document has ‘states’ as the most frequently occurring word (also prevalent in the French document), with ‘sovereignty’, ‘government’, ‘human’, and ‘right(s)’ appearing relatively frequently. These words infer the broader application of international law to cyberspace as the title suggests.

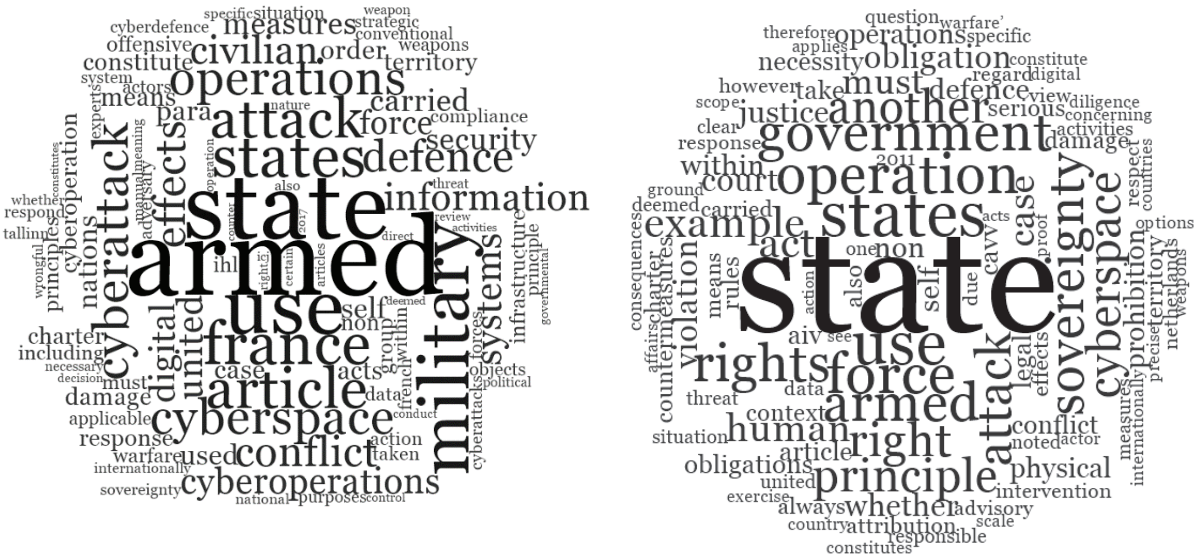


Figure 1: Word Cloud Comparing the French (left) and Dutch (right) Documents

4.2 Conceptual Codes

Table 3 provides a breakdown of the references and coded words per conceptual code for both national documents combined. As can be seen, armed conflict, response/retaliation, use of force, sovereignty, and attribution are the top five (in that order) based on references; however, for the number of coded words human rights exceeds sovereignty. These themes align to major concepts that are currently the focus of debate when applying international law to cyber-space.

Table 3: Summary of Conceptual Codes

Node	Number of Sources	Coding References	Words Coded	Paragraphs Coded
Armed conflict	2	53	3,864	106
Armed conflict\Cyber-weapon	1	5	292	7
Attribution	2	22	1,464	30
Human rights	2	17	1,308	37
Non-intervention	2	6	339	6
Operations outside of conflict	2	4	131	5
Response, retaliation	2	38	2,556	69
Response, retaliation\Pre-emptive	1	1	65	3
Responsibility and due diligence	2	15	943	27
Sovereignty	2	24	1,223	34
Use of Force	2	25	1,617	37

Figure 2 presents the word frequency of the coded content, visualised using a word cloud. The predominate words align to those presented in Figure 1, indicating the coded portions of the document (limited to the content explicitly considering cyber-space) is representative of the documents themselves.

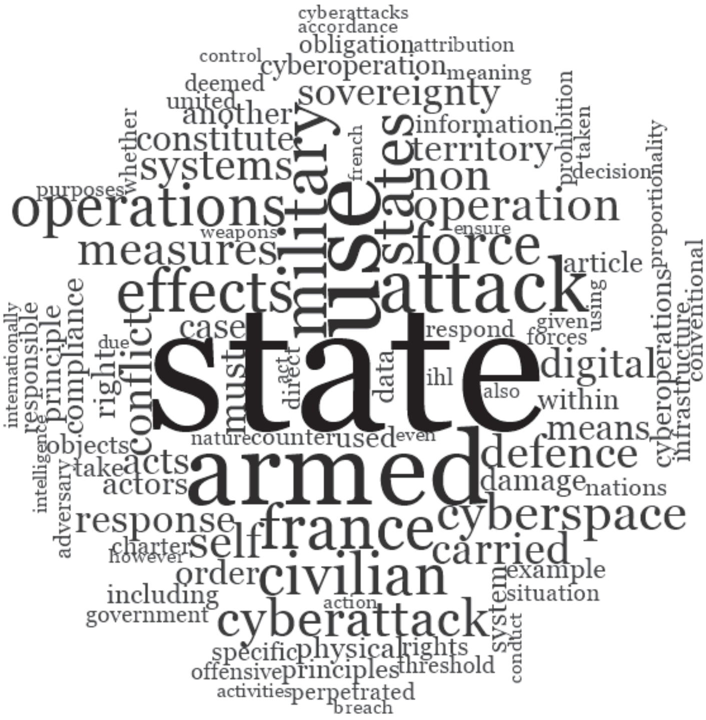


Figure 2: Word Cloud for Conceptual Codes

A cluster diagram based on word similarity, illustrating the similarities of the codes, is presented in Figure 3; both the conceptual codes and the challenges and uncertainty codes are presented in this figure for convenience. There are eight main clusters: (1) pre-emptive response and use of force; (2) uncertainty; (3) attribution, response/retaliation and responsibility and due diligence; (4) non-intervention and sovereignty; (5) operations outside of conflict; (6) challenges; (7) cyber-weapon; and (8) armed conflict and human rights. Clusters (1) to (4) are grouped together, and clusters (5)-(8) are grouped together. Table 4 presents the Pearson Correlations of the conceptual codes. This is formatted as a heat map, where dark green of the strongest correlation and dark red is the weakest. This provides a view of the strength of the various relationships amongst the conceptual codes. There are three pairs with the highest correlation (0.66): armed conflict and human rights; and responsibility and due diligence with both attribution and response/retaliation. This is followed closely by two pairs with a correlation of 0.65: response/retaliation with both attribution and use of force. Other notable correlations include sovereignty with responsibility and due diligence (0.57) and non-intervention (0.53) and armed conflict with response/retaliation (0.54).

Response/retaliation has four correlations greater than 0.5, followed by responsibility and due diligence with three. A correlation of greater than 0.5 implies a strong positive relationship. This indicates these two legal concepts are taking precedence in the discussion. This follows as international law is based on the responsibility of nations, however their legal ability to respond to international incidents needs to be considered with reference to the other concepts considered. Armed conflict, attribution and sovereignty all have two correlations above 0.5; this further illustrates these concepts have some significance.

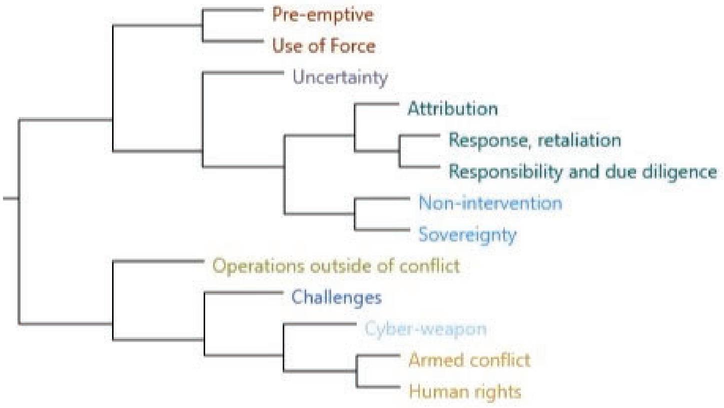


Figure 3: Codes Clustered based on Word Similarity

Table 4: Correlation of Conceptual Codes

	Armed Conflict	Cyber weapon	Attribution	Human rights	Non-intervention	Operations outside of conflict	Response/retaliation	Pre-emptive	Responsibility and due diligence	Sovereignty	Use of force
Armed Conflict	■	0.5	0.4	0.66	0.26	0.28	0.54	0.12	0.44	0.3	0.47
Cyber weapon	0.5	■	0.09	0.3	0.1	0.11	0.15	0.01	0.13	0.13	0.24
Attribution	0.4	0.09	■	0.12	0.35	0.12	0.65	0.12	0.66	0.42	0.39
Human rights	0.66	0.3	0.12	■	0.13	0.16	0.22	0.01	0.19	0.16	0.22
Non-intervention	0.26	0.1	0.35	0.13	■	0.15	0.42	0.1	0.38	0.53	0.48
Operations outside of conflict	0.28	0.11	0.12	0.16	0.15	■	0.18	0.03	0.06	0.27	0.19
Response/retaliation	0.54	0.15	0.65	0.22	0.42	0.18	■	0.29	0.66	0.48	0.65
Pre-emptive	0.12	0.01	0.12	0.01	0.1	0.03	0.29	■	0.16	0.06	0.39
Responsibility and due diligence	0.44	0.13	0.66	0.19	0.38	0.06	0.66	0.16	■	0.57	0.43
Sovereignty	0.3	0.13	0.42	0.16	0.53	0.27	0.48	0.06	0.57	■	0.42
Use of force	0.47	0.24	0.39	0.22	0.48	0.19	0.65	0.39	0.43	0.42	■

4.3 Challenges and Uncertainty

This section focuses on the codes for challenges and uncertainty, which are summarised in Table 5. These two codes are considered separately to provide specific word frequency analysis on what the areas of uncertainty or challenges. As is evident, both documents raised challenges and areas of uncertainty, however more instances of challenges were indicated.

Table 5: Summary of Challenges and Uncertainty Codes

Node	Number of Sources	Coding References	Words Coded	Paragraphs Coded
Challenges	2	9	390	10
Uncertainty	2	6	183	7

Figure 4 presents the word frequencies of the two codes as a word cloud. As can be seen, cyberspace is predominant; however, words such as “application”, “defined”, “actors”, and “effects” are noticeable. This implies that there are challenges/uncertainty regarding the application of the laws or their definition. In addition, there is uncertainty in determining the actors (i.e. attribution) or the effects of cyber-operations.

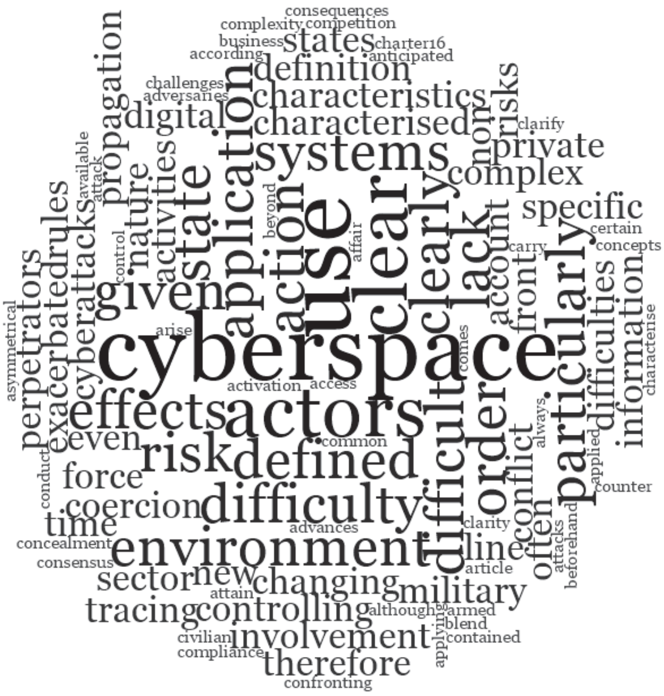


Figure 4: Word Cloud for Challenges and Uncertainty Codes

Table 6 presents the correlation amongst the challenges and uncertainty codes and conceptual codes. The strongest correlation is between uncertainty and use of force (0.34), indicating that determining what is considered a use of force is still unclear. A correlation of 0.29 is recorded for uncertainty and non-intervention as well as challenges and armed conflict, again implying problems in defining these concepts. Challenges and uncertainty have a low correlation (0.2), indicating the coding of these two was not identical.

Table 6: Correlation of Challenges and Uncertainty Codes with Conceptual Codes

	Armed Conflict	Cyber weapon	Attribution	Human rights	Non-intervention	Operations outside of conflict	Response/retaliation	Pre-emptive	Responsibility and due diligence	Sovereignty	Use of force	Challenges	Uncertainty
Challenges	0.29	0.20	0.18	0.24	0.12	0.04	0.15	0.02	0.20	0.18	0.15		0.20
Uncertainty	0.24	0.07	0.19	0.11	0.29	0.07	0.23	0.13	0.21	0.16	0.34	0.20	

The Netherlands document (2019: 1) suggests “international debate on ways to clarify the application of international law in cyberspace” to resolve the uncertainty. The document also indicates: “The government endorses the generally accepted position that each case must be examined individually to establish whether the ‘scale and effects’ are such that an operation may be deemed a violation of the prohibition of use of force” (p. 4). This is supported in a similar manner by the French document (Ministère des Armées, 2019: 8): “France reaffirms that a cyberattack may constitute an armed attack within the meaning of Article 51 of the United Nations Charter resulting from the use of physical force, if it is of a scale and severity comparable to those .”

In addition, when considering the uncertainty of the effects of cyber-weapons, that “these risks may be contained by the development of specific cyber weapons whose use is decided according to the desired effects” (Ministère des Armées, 2019: 16).

4.4 Alignment with Other Documents

This section assesses the alignment of the two national documents with previous authoritative documents relating to international law and cyber-space. The clustering of the documents is presented in Figure 5. Three main clusters can be seen: one contains the two *Tallinn Manual* documents, another contains the two national documents, and the third contains the Paris Call and Singapore Norm Package documents (although these last two can also be considered as separate clusters). The clustering of the two *Tallinn Manual* documents is unsurprising. The clustering of the two national documents shows their consistency.

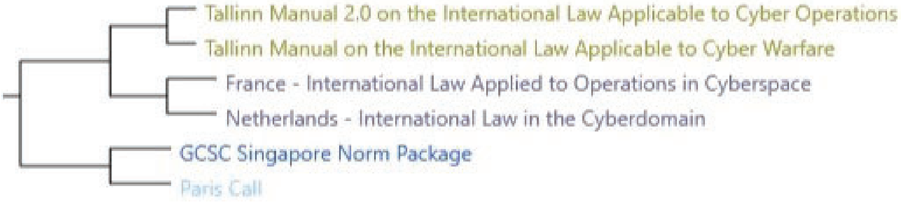

Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations
Tallinn Manual on the International Law Applicable to Cyber Warfare
France - International Law Applied to Operations in Cyberspace
Netherlands - International Law in the Cyberdomain
GCSC Singapore Norm Package
Paris Call

Figure 5: Sources Clustered based on Word Similarity

To further investigate the relationship of the national documents to the other documents, the Pearson Correlation is provided in Table 5, however limited to the two national documents correlated with the others and themselves (the correlations of the other documents amongst themselves is excluded). The strongest correlations (0.76) is between the Dutch document and the Tallinn Manual 2.0, followed by the French document and the first Tallinn Manual. Two correlations of 0.65 are present: the French document with Tallinn Manual 2.0, and the Dutch document with the first Tallinn Manual. These correlations are not surprising as the French document has a stronger focus on cyber-warfare (as does the first Tallinn Manual), whereas the Dutch document considers broader aspects of international law in cyber-space (and the Tallinn Manual 2.0 focuses on cyber-operations).

Table 5: Correlation of National Documents with Other Previous Documents

	GCSC Singapore Norms	Paris Call	Tallinn Manual	Tallinn Manual 2.0	France	Netherlands
France	0.43	0.37	0.72	0.65		0.36
Netherlands	0.45	0.36	0.65	0.76	0.36	

A comparison of the word frequencies for the national documents and the other documents is provided in Figure 6. There appears to be some consistency between the two sets of documents, and there is also consistency with Figures 1 and 2.

5. Summary of Results

The study has four objectives: (1) to identify the importance placed on various concepts of international law by the two national documents; (2) to investigate the relationship amongst the various legal aspects; (3) to identify challenges raised by the international documents, and any measure to mitigate these; and (4) to assess the alignment of the national documents to existing authoritative documents.

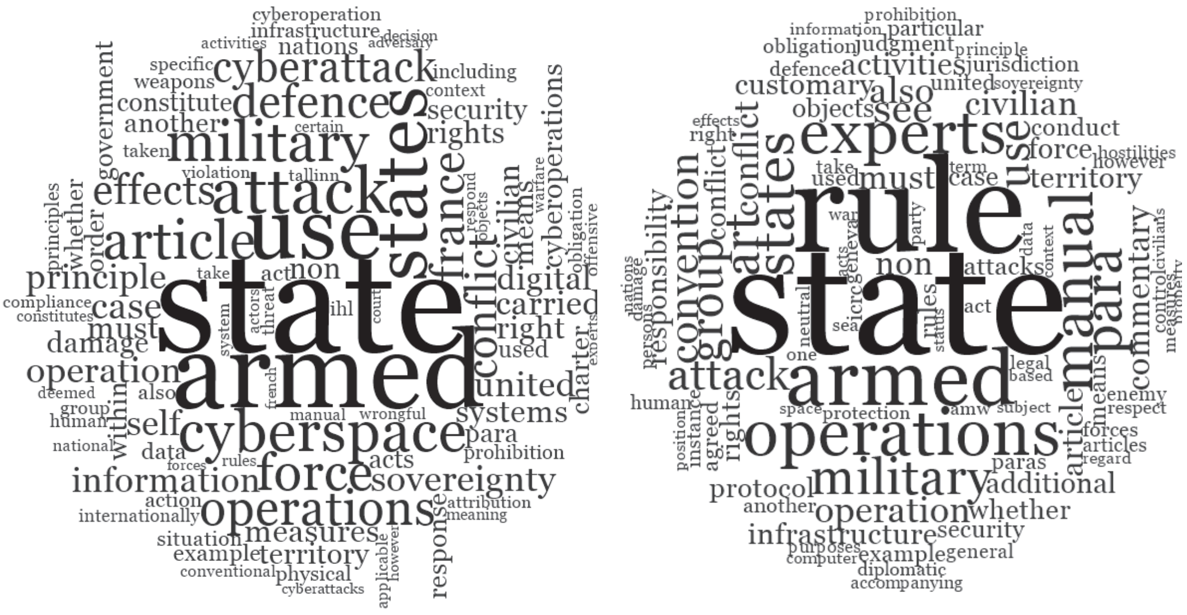


Figure 6: Word Clouds Comparing the National Documents (left) with Previous Legal Texts (right)

From Sections 4.1 and 4.2, the major concepts based on the number of references and coded words are armed conflict, response/retaliation, use of force, sovereignty, attribution, and human rights. These themes are consistent with the major areas of debate regarding cyber-operations and international law. In Section 4.2 it was shown that response/retaliation has the strongest correlations (four), followed by responsibility and due diligence with three. Other terms with strong correlation include armed conflict, use of force, human rights, attribution and sovereignty. This aligns to the codes with the most references, reinforcing the importance of these concepts. This also illustrates the linkages amongst the major concepts and therefore the complexities of cyber-operations and international law.

As illustrated in Section 4.3, the strongest correlations are between uncertainty and use of force, uncertainty and non-intervention, and challenges and armed conflict. This implies that there is ongoing concerns regarding the application of international law to cyber-operations, and the key themes identified above may still be subject to interpretation. Proposed measures to mitigate these areas of uncertainty include further international debate and equating the effects of cyber-attacks to those of physical attacks.

There is strong correlation between the Dutch document and the Tallinn Manual 2.0, as well as the French document and the first Tallinn Manual. In the documents, the Dutch explicitly state that the document follows the advice provided in the Tallinn Manuals, whereas in the French document it is stated that there is deviation from the Tallinn Manuals on some points. Despite this, Schmitt (2019) feels that the French position does align with that of the Tallinn Manuals in some respects, such as the definition of attack. Roguski (2019), however, considers that the French position does deviate in places from the Tallinn Manuals and other views. Regardless, the two Tallinn Manual documents can be considered to have strongly influenced the two national documents.

The limitation of the qualitative document analysis is that it does not give sufficient understanding of political or strategic thinking or intent. Future research can combine the document analysis with additional political science analysis. Future research can provide analysis of the two national documents in relation to a broader set of existing documents and commentary, and additional comparisons with any future national perspectives on the applicability of international law on cyber-space and cyber-operations.

6. Conclusion

The prevalence of cyber-attacks in international relations has resulted in renewed focus on the applicability of international law related to cyber-space and cyber-operations. Two countries, France and the Netherlands, released national perspectives on this matter. This paper assessed the two national documents and identified the major themes that the documents focus on, which is in line with existing academic and international debate. There is correlation amongst a number of concepts, further indicating the importance thereof, whilst illustrating the complexity of the topic. There are still elements of uncertainty highlighted by the documents, however in order to mitigate this there is an attempt to equate the effects of cyber-operations to physical attacks. There is a call for further international debate to aid in reaching consensus regarding the lack of clarity. The study found that there was a strong influence on the national documents by the two *Tallinn Manuals*, even if the documents stated that they deviate from the views in the manuals. There is scope to expand this research as additional commentary and national perspective become available.

References

Council of Europe. (2001) Convention on Cybercrime, Budapest, European Treaty Series - No. 185.

DNI, United States Director of National Intelligence, (2017) "Intelligence Community Assessment: Assessing Russian Activities and Intentions in Recent US Elections", 7 January, [online], accessed 12 April 12 2018, https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Doffman, Z., (2019) "U.S. Attacks Iran With Cyber Not Missiles -- A Game Changer, Not A Backtrack", *Forbes*, 23 June, [online], accessed 12 July, https://www.forbes.com/sites/zakdoffman/2019/06/23/u-s-attacks-iran-with-cyber-not-missiles-a-game-changer-not-a-backtrack/#3970a285753f

Eilstrup-Sangiovanni, M. (2018) "Why the World Needs an International Cyberwar Convention", *Philosophy & Technology* 31, 379-407.

Fingas, J., (2019) "Israel is the first to respond to a cyberattack with immediate force", *Endgadget*, 5 May, [online], accessed 6 May, https://www.engadget.com/2019/05/05/israel-responds-to-cyberattack-with-airstrike/

Global Commission on the Stability of Cyberspace. (2018) *Norm Package Singapore*, November, [online], accessed 30 January 2019, https://cyberstability.org/wp-content/uploads/2018/11/GCSC-Singapore-Norm-Package-3MB.pdf.

Greenberg, A. (2017) "How an Entire Nation Became Russia's Test Lab for Cyberwar," *Wired,* 25(7), [online], accessed 18 January 2019, https://www.wired.com/story/russian-hackers-attack-ukraine/.

Kilovaty, I. (2014) "Cyber Warfare and the Jus Ad Bellum Challenges: Evaluation in the Light of the Tallinn Manual on the International Law Applicable to Cyber Warfare", *American University National Security Law Brief* 5(1), 91-124.

Lindsay, J.R. (2015) "Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack", *Journal of Cybersecurity*, 1(1), 2015, 53–67.

Lotrionte, C. (2012) "State Sovereignty and Self-Defense in Cyberspace: A Normative Framework for Balancing Legal Rights", Emory International Law Review, 26(2), [online], accessed 16 January 2020, http://law.emory.edu/eilr/content/volume-26/issue-2/symposium%20/state-sovereignty-self-defense-in-cyberspace.html

Ministére des Armées, (2019) *International Law Applied to Operations in Cyberspace*, [online], accessed 18 January, https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf

Paris Call for Trust and Security in Cyberspace (2018), 12 November, [online], accessed 18 January 2019, https://www.diplomatie.gouv.fr/IMG/pdf/paris_call_cyber_cle443433.pdf

Netherlands Parliament, (2019a) Letter of 5 July 2019 from the Minister of Foreign Affairs to the President of the House of Representatives on the international legal order in cyberspace.

Netherlands Parliament, (2019b) *Appendix: International law in cyberspace*, 26 September, [online], accessed 8 January 2020, https://www.government.nl/binaries/government/documents/parliamentary-documents/2019/09/26/letter-to-the-parliament-on-the-international-legal-order-in-cyberspace/International+Law+in+the+Cyberdomain+-+Netherlands.pdf

Roguski, P. (2019) "France's Declaration on International Law in Cyberspace: The Law of Peacetime Cyber Operations, Part I", *OpinioJuris*, 24 September, [online], accessed 26 September, https://opiniojuris.org/2019/09/24/frances-declaration-on-international-law-in-cyberspace-the-law-of-peacetime-cyber-operations-part-i/

Schmitt, M.N. (2013) *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge: Cambridge University Press.

Schmitt, M.N. (2017) *Tallinn Manual 2.0: On The International Law Applicable to Cyber Operations*, Cambridge: Cambridge University Press.

Schmitt, M.N. (2019) "France Speaks Out on IHL and Cyber Operations: Part II", *EjilTalk*, 1 October, [online], accessed 16 October, https://www.ejiltalk.org/france-speaks-out-on-ihl-and-cyber-operations-part-ii/

Student n°2222171 (2016) *The Challenges of Cyber Warfare to the Laws of Armed Conflict: Humanitarian Impact and Regulation Attempts*, Master dissertation, University of Glasgow.

The International Institute for Strategic Studies, (2018) "Editor's Introduction: Western technology edge erodes further," *The Military Balance,* 118(1), pp. 5-6.

United Nations. (2015) Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/70/174, 22 July, [online], accessed 27 January 2020, https://dig.watch/sites/default/files/N1522835.pdf

Signing Your Name Away: An Overview on the Risks of Electronic Signatures

Suné von Solms
University of Johannesburg, South Africa
svonsolms@uj.ac.za

DOI: 10.34190/EWS.20.009

Abstract: Many companies are joining the worldwide drive toward a paperless environment. The utilisation of digital documents, contracts or approval forms can save organisations large costs and time as documents can be electronically signed and transmitted to their destination within minutes. Electronic signatures can bring many advantages, but a review of the current situation shows many cybersecurity threats around the signing of documents in a digital form. Many companies are still in the early phases of utilising electronic signatures, with many potential opportunities for manipulation. This paper discusses the various types of electronic signatures, how they can be exploited and recommends general security measures individuals and organisations can follow in order to use electronic signatures more securely.

Keywords: Advanced electronic signatures, Cybersecurity, Digital signatures, Electronic signatures, Security.

1. Introduction

Many large companies across the globe are joining the worldwide drive to go paperless. Considerations for greater levels of digitisation does not only include a smaller carbon footprint and the preservation of the environment. Additional advantages for a paperless environments includes the reduction in the need for storage facilities for legal documents, contracts and other documents which must be preserved, the courier costs and time spent sending documents from signatory to signatory as well as the clumsy process of print, sign and scan (Finjan, 2017). One of the activities which accompanies paper-less processes is utilisation of electronic signatures. Electronic signatures are considered one of the last barriers toward the achievement of digital end-to-end solutions, coined the "final frontier" in the journey toward the ultimate destination of the paperless organisations (Pulse, 2015).

The use of signatures is so engrained in our modern way of life and work and are universally understood to carry a specific meaning, both legally and in business (Heyink, 2014). Therefore, the move toward the digitisation of the signature comes with numerous challenges, apprehension and ignorance (Heyink, 2014; Pulse, 2015). The utilisation of electronic signatures can bring with it a multitude of cybersecurity threats. As with any digital document or digital information, it must be ensured that the reliability of the signature itself and the integrity of the electronic information is preserved (Heyink, 2014). As the use of electronic signatures are only starting to become practice within companies, many practices within organisations are insecure and open to manipulation and fraud.

This paper considers the various methods which can be used to electronically sign a document and reviews the cybersecurity threats related to the various methods of electronically signing a document. The outline of the paper is as follows: Section 2 provides an overview of the functions of signatures in manuscripts and electronic documents as well as the various functionalities and implementation of electronic signatures. Section 3 contains a discussion on the cybersecurity threats related to the utilisation of electronic signatures where Section 4 concludes this paper.

2. Overview of Electronic Signatures

This section considers the various concepts around electronic signatures. In informal conversation, the terms "digital signature", "electronic signature" or "e-signature" can be used interchangeably and considered to be the same. Although these terms may be considered to mean something similar, the technologies are different and, in many countries, carry a different legal weight. This section provides an overview on the various types of signatures which exist as well as the advantages and uses for them, also comparing them to the traditional manuscript signatures.

2.1 Publications per year of African countries

The signature has been part of the written communication for thousands of years, where the implication of the signature carries a certain meaning. A signature mainly has three primary functions which are recognised within organisations, law and practice (Heyink, 2014; Michaelsons, 2017):

1. Identity: provides evidence of the identity of the signatory.
2. Intent: provides evidence that the signatory intended to sign the document.
3. Approval: provides evidence that the document on which the signature is made is adopted or approved by the signatory.

When a signature is seen on a document, this is considered evidence that the abovementioned functions were intended to be achieved by the person signing the document. Many practices have developed over the years to mitigate possible cases of signature manipulation. For example, when a paper document is signed, the ink becomes part of the paper and cannot be separated and the practice to initial all pages of a contract ensures that content cannot be changed or added. (Heyink, 2014).

Past court rulings relating to signatures have stated that a signature can be considered a valid signature if the three abovementioned functions are satisfied. This means that a plain name written down, a thumb print or a mark made can be considered a signature if it is the name or mark of the person signing, if the person made the mark themselves and if the person intended to make the sign in approval of the contents of the document. Therefore, considering the functions of a signature, an electronic signature can be considered a legal signature as it fulfils the functions of a signature (Christensen, 2003; Michaelsons, 2017). The Electronic Communications and Transactions Act No 25 of 2002 (the ETC act) section 13(2) in South Africa states that an electronic signature is not without legal force and effect simply because it is in electronic form (Adobe, 2017; ETC act, 2002). The Electronic Identification and Authentication Service Regulation (910/2014/EC) states that an electronic signature shall not be denied legal effect simply because it is in electronic form (Adobe, 2015). Many other electronic signature laws across the globe, although all unique, follows the same general argument that the intent of the signature is considered above the form of it. Therefore, a signature, whether it be digital, manuscript or a name at the bottom of an email, will be regarded by the courts as evidence that these functions were intended by the signatory, absent from evidence proving the contrary (Heyink, 2014).

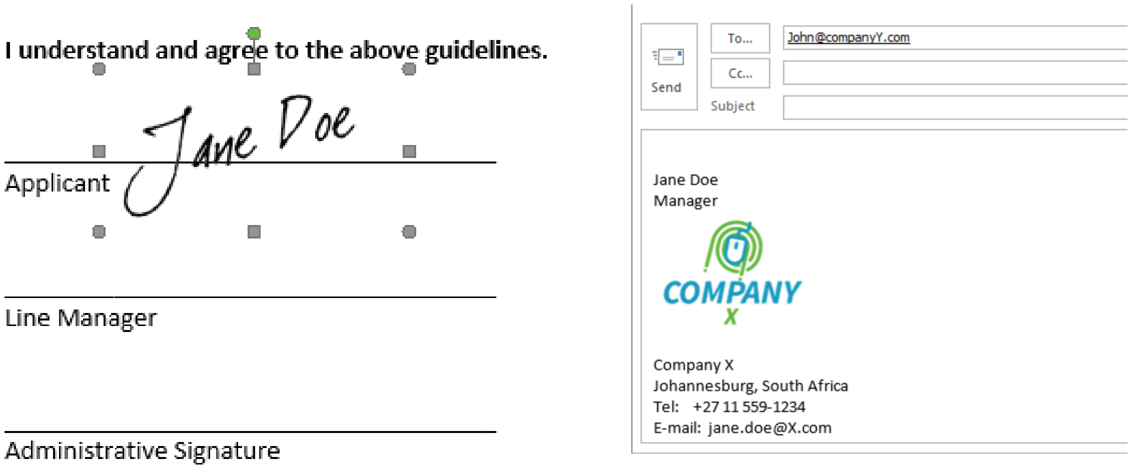
The ETC act defines an "electronic signature" as "data attached to, incorporated in, or logically associated with other data which is intended by the user to serve as a signature" (ETC act, 2002). The American Bar Association states in their Guidelines to Digital Signatures that digitally signed writing serve the following purposes (ABA, 1996):

1. Evidence: a signature authenticates a writing by identifying the signatory with the signed document.
2. Ceremony: calls to the signer's attention the legal significance of the signer's act.
3. Approval: a signature expresses the signer's approval or authorisation of the writing.

It can be seen that the functions defined in the digital signature guidelines and the functions defined for manuscript signature correlate closely. There exist three broad categories of electronic signatures, all with their own level of security implications, discussed subsequently.

2.2 Simple electronic signatures

A simple electronic signature includes the copying and pasting of scanned signatures, tick box approvals or typing a name (DBEIS, 2016; Heyink, 2014). This method is currently used usually where records are electronic and where a traditional paper manuscript would have been used in the past (Heyink, 2014; Srivastava, 2009). Figure 1 (a) shows a typical example of a digital approval form which the applicant has signed with the use of a simple electronic signature in the form of an image and Figure 1 (b) shows a signature at the bottom of an email. Both instances can be considered valid electronic signatures as they meet the three functional criteria of a signature: the person who wrote/read the document inserted the signature, the person intended it to be a signature, and there is an association between the document and the signature (Pulse, 2015; Delaney and Francis, 2018).



I understand and agree to the above guidelines.

Jane Doe

Applicant

Line Manager

Administrative Signature

To... John@companyY.com

Jane Doe
Manager

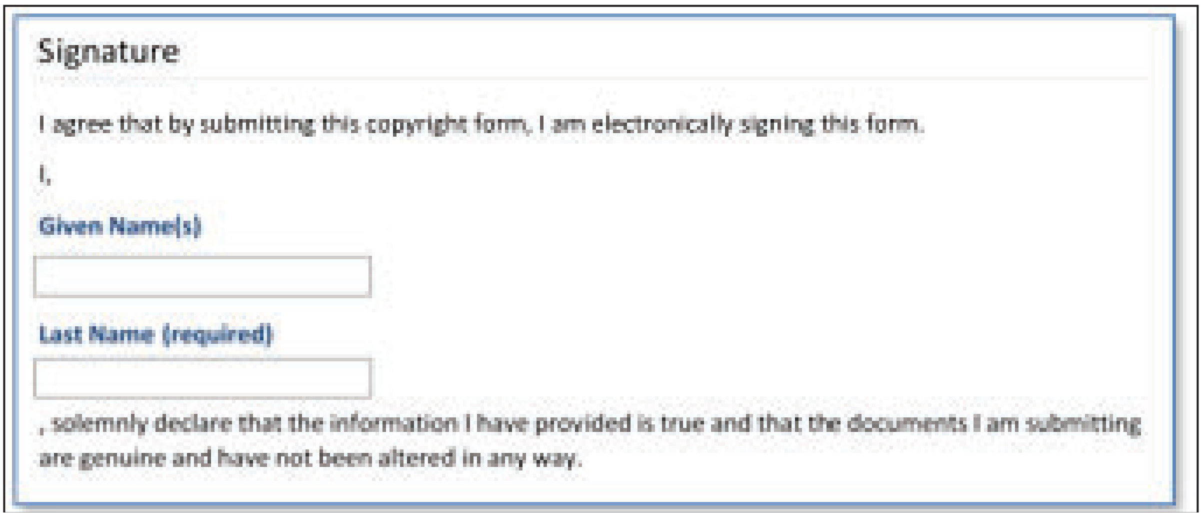
COMPANY X

Company X
Johannesburg, South Africa
Tel: +27 11 559-1234
E-mail: jane.doe@X.com

(a) (b)

Figure 1: Example of simple electronic signatures in (a) an approval form and (b) email signature.

Typing your name into an electronic form and clicking “I Agree are used in many cases including transferral of copyright material and agreeing to terms of a software license (Nolo, 2018). An example of such a signature in shown in Figure 2. By using this method, a person has legally electronically signed a document as it meets the three criteria of a signature.



Signature

I agree that by submitting this copyright form, I am electronically signing this form.

I,

Given Name(s)

Last Name (required)

, solemnly declare that the information I have provided is true and that the documents I am submitting are genuine and have not been altered in any way.

Figure 2: E-Signature for online form

It can be seen from the examples that simple electronic signatures may be considered the functional equivalence of a manuscript signature and considered legally binding, unless it can be disputed. Simple electronic signatures do not provide safeguards against content manipulation and fraud. In the paper environment people have created safeguards to protect them against unauthorised signing and amendment, which includes signing each page, ruling out blank space and use of handwriting experts. In the same way as paper safeguards are used, safeguards in the digital environment must also be utilised.

2.3 Digital signatures or advanced electronic signatures (AES)

In an international context, the phrases “digital signature” and “advanced electronic signature” can cause confusion. For example, in South Africa this second category of signatures are called digital signatures as defined by its ECT Act, but in Europe this is referred to as an advanced electronic signature (CEF, 2018; Heyink, 2014). For clarity in this work, this second category of signatures will be referred to as “digital signatures”.

Digital signatures are defined as a measure to “satisfy the legal and business requirements of authenticity, integrity, non-reputability and writing and signature” (Smedinghoff, 1996) or “a type of electronic signature which is created and verified using cryptography” (Srivastava, 2009). Digital signature utilises Public Key Infrastructure (PKI) to encrypt the resulting message with the sender’s private key (Smedinghoff, 1998; Blythe, 2005). A digital signature differs from a simple electronic signature as the digital signature is derived from and linked to the document to be signed (Smedinghoff, 1998, Heyink, 2014). Asymmetric encryption is utilised in the creation of a digital signature. Once signed, the contents of the document cannot be changed, but only viewed and verified (Heyink, 2014; Blythe, 2005). These processes are illustrated in Figure 3.

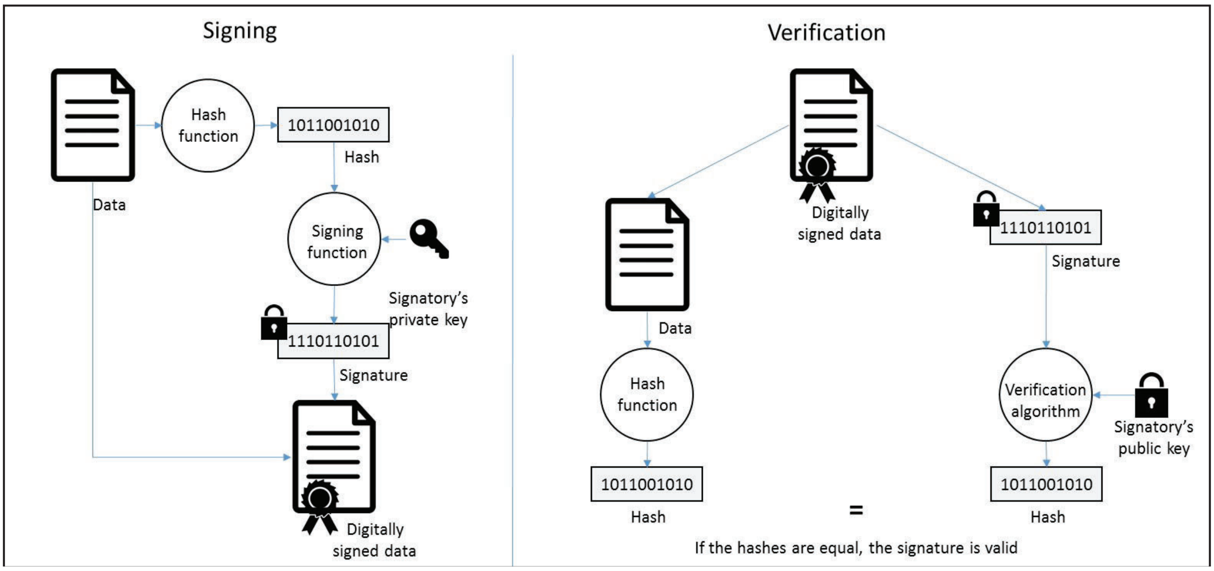


Figure 3: Creation and verification processes of a digital signature

The first step in the signing process is where a hash is created from the document. The signatory's private key is used together with the hash to create the digital signature, unique to the specific document and signatory. After the document has been signed, any person can verify the integrity of the document as well as identify the signatory. A first hash value is computed from the content of the document, while a second hash function is computed from the digital signature, using the signatory's public key. If these two hashed match, the contents of the document is proved to be valid and unchanged. Any changes to the content of the document will result in a different hash value that will not match the hash computed from the digital signature.

The digital signature process allows the signatory's identity to be linked to the signature and to "lock" the contents of the document to ensure the integrity of the contents. A digital signature adds a timestamp so that the time and date of the signature can be verified. It can be seen that digital signatures have included a set of safeguards to ensure the integrity of the document and enable the identification of the signatory, not present with simple electronic signatures.

2.4 Advanced Electronic Signatures or Qualified Electronic Signatures

As before, in an international context, the phases "advanced electronic signature" and "qualified electronic signatures" can cause confusion. South African law defined this third category of signatures as advanced electronic signatures, but it is known as qualified electronic signatures in the EU (DBEIS, 2016; Heyink, 2014). For clarity in this work, this third category will be referred to as "advanced electronic signatures". An advanced electronic signature is defined by the ECT act as an electronic signature which results from a processes accredited by "authentication service providers" or an "accreditation authority" (ECT Act, 2002). The advanced electronic signature, accredited by a Supervisory Body, makes the signature prima facie valid, meaning that it is always assumed to be valid unless the contrary is proved (ECT Act, 2002; DBEIS, 2016). An advanced electronic signature links the identity of the signatory to the signature and ensures the integrity of the document. The main difference between advanced electronic signatures and digital signatures is that advanced electronic signatures are approved by an authentication service provider.

Certificate authorities (CAs) are organisations who can issue different types of certificates for a variety of purposes and levels of trust. The process of digital signatures through the use of certificates issued by a CA is illustrated in Figure 4 (IBM, 2020; Sachimani, 2013).

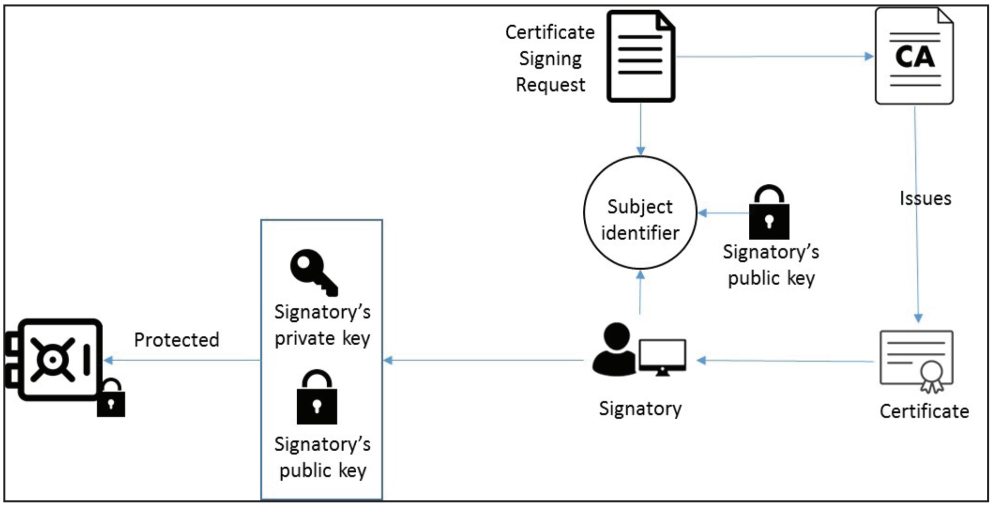


Figure 4: Process of obtaining a digital certificate from a CA.

The CA receives a request from a client, whereby the authority verifies their identity before building the certificate and returning it to the client as a personal certificate. It can be seen from the process that the utilisation of a CA to issue a certificate links the identity of the signatory to the signature and ensures the integrity of the document.

3. Potential exploits and recommendations

"Electronic signatures are only as secure as the business processes and technology used to create them" (DBEIS, 2016). In general, high valued transactions need better and more secure signatures. Large corporations utilising electronic signatures would typically utilise advanced electronic signatures as it can provide the required level of assurance and also ensure trust in the underlying system. It can link the signatory to the information, link the signature to the signatory and ensure the integrity of the document. Unfortunately, the utilisation of advanced electronic signatures is limited due to the requirement for an authentication service provider or certification authority to accredit the signature process. This makes the utilisation of these signatures impractical for the day to day use within an organisation, meaning that employees generally use simple electronic signatures and digital signatures.

3.1 Simple electronic signatures

Simple electronic signatures are currently used in the majority of instances where an electronic document is used in place of a traditional paper manuscript (Heyink, 2014; Srivastava, 2009). It has become a convenient and acceptable method of approving an electronic version of a document instead of hard copies, which must be posted or scanned and emailed. As described in Section 2.2, a simple electronic signature does not provide any level of security, although it is considered a valid signature unless proven otherwise. This paper argues that the biggest concern of an individual regarding simple electronic signatures should not be the legal acceptability thereof, but fraud.

When considering the first function of a signature, a signature must provide evidence to identify the signatory. Where manuscript signatures could be subjected to evaluation by handwriting experts when a dispute occurs, it is more difficult from a forensic point of view to prove that a simple electronic signature was not added to a document by the owner. When a person inserts a digital image of their signature to a document or at the bottom of an email, they run the risk of that signature being fraudulently stolen and misused (Mednet, 2013; Knight, 2000). In cases within an organisation where the same line manager must sign off or approve documents and does so with an inserted image of his/her signature, instances can occur where employees simply bypass the chain of approval and add the signature themselves. The action of simply inserting an image or a name in text, also creates the possibility of altering the content of the signed document. As the document is not "locked" after an image is inserted into a document, the signatory cannot be ensured that the contents will not be altered, and it may be very difficult to prove if disputed. In the abovementioned example, an employee can change details within an approved document after the line manager has included his/her signature.

These examples are extremely simple, yet the scanned image of personal signatures is widely used within and amongst organisations. Therefore, it is imperative that when simple electronic signatures are used, precautions must be taken to minimise the risk of misappropriation. Scanned image signatures inserted into a document are not physically bound to that document, therefore measures should be taken to ensure the relationship of a scanned signature and its associated document are preserved and that the signature cannot be used in other documents (Albany, 2017; Mednet, 2013).

An image of your personal signature must therefore never be inserted into the body of an email as it can be copied, saved and used by a recipient of any email. A document which can be edited, such as a Word or Excel document, must never be distributed when it contains a signature image file. When signatures are added to a document, a password-protected or encrypted PDF must be created. Once a file is password-protected or encrypted, the password must not be sent in the same email as the document. This adds an addition layer of security to deter fraudsters from cropping a signature out of the PDF. An individual must avoid sending scanned images of signatures to external organizations, if possible. When a signature is included in a document, that specific signature file must be deleted after use.

Organisations must ideally develop procedures to identify, evaluate, and record where simple electronic signatures are permitted and where more secure signatures must be used. The individual or organisation must also retain a copy, or log, of all documents signed electronically such that they can be retrieved if required.

3.2 Digital signatures

Digital signatures to a large extent address many of the security concerns of simple electronic signatures. A digital signature is uniquely linked to the signatory via a private key (password protected), links the data in the document to the signature, provides a unique timestamp and makes any subsequent change to a document detectable (DBEIS, 2016; CEF, 2018; Finjan, 2017). Literature has reported on various content attacks on digitally signed document through the manipulation of dynamic content and file format exploits (Buccafurri, 2008; Srivastava, 2009; Alsaid et al., 2005; Kain et al., 2002). However, many of these exploits are solved with security and software updates and updated signature creation processes. This paper does not consider the technical exploits which can compromise the signing process or document integrity, as the encryption technology behind a digital signature does not ensure that the signature has been applied by the owner of the cipher key. Many security threats associated with digital signatures relates to human behaviour.

Security issues relating to digital signatures are widely debated in the literature as scholars argue that a password is not an adequate method for protecting a private key (Srivastava, 2009; Myers, 1999; Mason and Bohm, 2003). Many individuals select password which are easy to guess, reuse passwords or write them down. They also neglect to change important passwords often, which makes the password behind a private key prone to attack (Mason and Bohm, 2003; Julia-Barcelo et al., 1998). Social engineering strategies continue to be effective as individuals are tricked or convinced into disclosing their passwords (Whitman, 2004; Regan, 2006). Malware, such as software which remotely copies data from an individual's computer or logs keystrokes, can be used by hackers to gain access to passwords and secured documents (Srivastava, 2009).

Securing digital signatures therefore mainly resides around the use and protection of strong passwords. Studies have found repeatedly that despite password security practices within an organisation, employees show carelessness toward password management (Srivastava, 2009; Ernst & Young, 2006; Mulligan, 2005). The protection of a private key also requires the individual to protect the computer where it resides on. These well-known password protection methods are essential to secure an individual against possible signature fraud. The security of stored digital signatures can easily be compromised if the signatories are not vigilant in keeping their passwords protected and their computers secure. The existence of a strong security infrastructure for digital signatures is not sufficient to keep it safe and must be complemented with awareness and training programs for employees.

4. Conclusion

This paper provided an overview on the various categories of electronic signatures which exit. Electronic signatures are considered valid signature unless proven otherwise and has the potential to speed up and simplify the processes within organisations. This paper argues that the biggest concern relating to electronic signatures are not the legal acceptability thereof, but possible fraud. Various insecure practices of individuals around electronic signatures and how they can be exploited are discussed and argues that although there exist sound encryption technologies and strong security infrastructure for digital signatures, this alone cannot protect individuals against signature fraud. Individuals must learn the value of their signature and use it securely. An image of their signature must be used with care and must be protected. Passwords of private keys, as any other password, must be managed securely and the computers where these signatures are kept must be kept secure.

As many companies are in the early phases of utilising electronic signatures, the insecure practices which exist opens companies and individuals to fraud relating to their signatures. The impact of the unlawful use of a signature may have massive consequences in business and personal sphere, therefore awareness and training around safe signature use must be improved in the future.

References

Adobe Systems Incorporated (2015) "Global Guide to Electronic Signature Law: Country by country summaries of law and enforceability". San Jose, USA.

Alsaid, A and Mitchell, C.J. (2005) "Dynamic Content Attacks on Digital Signatures". Information Management & Computer Security, 13(4).

American Bar Association (1996) "American Bar Association Guideline to Digital Signatures". White Paper. Information Security Committee, Section of Science & Technology, American Bar Association.

Blythe, S.E. (2005) "Digital Signature Law of the United Nations, European Union, United Kingdom and United States: Promotion of Growth in E-Commerce With Enhanced Security". Journal of Law & Technology, 11(2).

Buccafurri, F. (2008) "Digital Signature Trust Vulnerability: A new attack on digital signatures". Information Systems Security Association Journal, 10.

Christensen, S., Duncan, W. and Low, R. (2003) "The Statute of Frauds in the Digital Age - Maintaining the Integrity of Signatures". Murdoch University Electronic Journal of Law, 10 (4).

Connecting Europe Facility (CEF) (2018) "What is an electronic signature?" [WWW Document]. URL https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action? pageId=46992760 (accessed 10.06.18).

Delaney, H. and Francis, B. (2018) "Electronic signatures and their legal validity in Australia". [WWW Document]. URL http://www.findlaw.com.au/articles/ 5777/electronic-signatures-and-their-legal-validity-in-.aspx (accessed 13.06.18).

Department of Business, Energy & Industrial Strategy (DBEIS) (2016) "Electronic Signatures and Trust Service Guide. London.

Ernst & Young (2006) "Cybersecurity regained: preparing to face cyber attacks". EY Global Information Security Survey 2017-18 [WWW Document]. URL https://www.ey.com/gl/en/services/advisory/ey-global-information-security-survey-2017-18 (accessed 14.06.18).

Finjan Blog (2017) "Digital Signatures and Information Security" [WWW Document]. URL https://blog.finjan.com/digital-signatures-and-information-security/ (accessed 03.06.18).

Heyink, M (2014) "Electronic Signatures for South African Law Firms". Law Society of South Africa.

IBM (2020) "Obtaining personal certificates from a certificate authority". IBM knowledge centre [WWW Document]. URL https://www.ibm.com/support/knowledgecenter/en/SSFKSJ_8.0.0/com.ibm.mq.sec.doc/q009870_.htm (accessed 13.03.20).

Julia-Barcelo, R. and Vinje, T. (1998) "Towards a European Framework for Digital Signatures and Encryption". Computer Law and Security Report, 14(2).

Kain, K., Smith, S.W. and Asokan, R. (2002) "Digital Signatures and Electronic Documents: A Cautionary Tale". Proceedings of the Sixth IFIP Conference on Communication and Multimedia Security.

Knight, W. (2000) "Digital Signatures Pose Security Risk, says Expert". ZDNet [WWW Document]. URL https://www.zdnet.com/article/digital-signatures-pose-security-risk-says-expert/ (accessed 13.06.18).

Mason, S. and Bohm, N. (2003) "The Signature in Electronic Conveyancing: An Unresolved Issue?" The Conveyancer and Property Lawyer.

Mednet (2013) "Scanned Handwriting Signature". UBC Faculty of Medicine [WWW Document]. URL https://mednet.med.ubc.ca/ServicesAndResources/Communications/InternalCommunications/Pages/Scanned-Handwritten-Signatures.aspx (accessed 10.06.18).

Michaelsons (2017) "Electronic Signature Handbook". Michealsons.

Mulligan, J. and Elbirt, A.J. (2005) "Desktop Security and Usability Trade-offs: An Evaluation of Password Management Systems". Information Systems Security, 10(10).

Myers, S.G. (1999) "Potential Liability under the Illinois Electronic Commerce Security Act: Is it a Risk Work Taking?" The John Mitchell Journal of Computer and Security Law, 17(3).

Nolo (2018) "Electronic Signatures and Online Contracts" [WWW Document]. URL https://www.nolo.com/legal-encyclopedia/electronic-signatures-online-contracts-29495.html (accessed 13.06.18).

Pulse, H.R. (2015) "Understanding the legal side of electronic signatures in South Africia" [WWW Document]. URL http://www.itnewsafrica.com/2015/07/ understanding-the-legal-side-of-e-signatures-in-south-africa/ (accessed 13.06.18).

Regan, K. (2006) "The Fine Are of Password Protection". E-commerce Times [WWW Document]. URL https://www.ecommercetimes.com/story/21776.html (accessed 14.06.18).

Sachimani (2013) “x509 Certificate – Asymmetric encryption and Digital Signatures”. WordPress [WWW Document]. URL https://sachi73blog.wordpress.com/2013/11/21/x509-certificate-asymmetric-encryption-and-digital-signatures/

Smedinghoff, T. J. (1996) “Online Law: The SPA's Legal Guide to Doing Business on the Internet”. Addison-Wesley Developers Press.

Smedinghoff, T.J. (1998) “Electronic Contracts and Digital Signatures”. Journal of Equipment Lease Financing, 16(2).

Srivastava, A. (2009) “Electronic Signatures and Security Issues: An Empirical Study”. Computer Law and Security Review, 25, pp 432 – 446.

Whitman, M.E. and Mattord, H.J. (2004) Management and Information Security 4th Ed. Cengage Learning.

Artificial Intelligence and its' Legal Risk to Cybersecurity

MM Watney
University of Johannesburg, South Africa
mwatney@uj.ac.za

DOI: 10.34190/EWS.20.026

Abstract: The risk pertaining to data and technology are continuously growing in scale and severity as the dependence on, and advancement in technology grows. The last decade has seen hundreds of cases of identity theft, loss of money and data breaches. It is estimated that by the year 2021, cybercrime losses may cost upwards of $6 million annually. Due to the constant risk of intrusions, the tech industry, businesses and government bodies must safeguard technology and data and this is where cybersecurity plays a major and critical role. It may be considered as the first line of defense against intrusions. The tools and techniques developed and supported by AI and machine learning (ML) are now expanding to cybersecurity to protect against cybercrime. There are many advantages in using AI-ML technology for cybersecurity but it is a double-edged sword. Just as AI cybersecurity technology may be used to more accurately identify and stop intrusions, the AI systems may be exploited for the commission of cybercrime. Cybersecurity is not an issue that a government can address on its own; it requires multi-stakeholders to work together in addressing the legal risk AI-ML technology presents to cybersecurity. The discussion is divided into two parts: Part 1 explores the beneficial use of AI cybersecurity technology; and Part 2 considers the harmful use of AI such as data breaches and the commission of cybercrimes. Cybersecurity and cybercrime are issues that cannot be separated from each other. AI-driven cybersecurity technology must keep pace with the legal risk that the use of AI in the commission of cybercrime poses otherwise it cannot effectively prevent, detect, respond to and recover from an intrusion. A preventative approach rather than a detection-focused approach should be applied to cybersecurity. The discussion identifies the legal risk that the use of AI for cybersecurity presents and how it may be addressed by means of ethics, policies and laws.

Keywords: AI-driven cybersecurity; AI-driven cybercrime, data breach, the legal risk AI-ML technology presents to cybersecurity

1. Introduction

Emerging technologies such as AI, ML, 5G, quantum computing and evolving technologies such as IoT, autonomous vehicles and cloud computing have an impact on cybersecurity and the commission of cybercrime. The technologies create cybersecurity challenges and expand the threat landscape. For example, by 2020 75 million IoT devices are expected to be in use and it is predicted that a quarter of cyber-attacks on enterprises will involve IoT devices (Press, 2019). 5G, which enables faster speeds and increased bandwidth, will result in even more devices connecting to the Internet and the attack surface for the commission of hacking and DDoS attacks may increase (Press, 2019).

The role that cybersecurity plays in protecting against cybercrime evokes a great deal of discussion. Over the years businesses and financial institutions have experienced many cyber intrusions. Cybersecurity cannot be effectively managed without knowing the risk it faces. For example, in 2017 a cyberattack referred to as "WannaCry" affected more than 200 000 computers in 150 countries over the course of just a few days. It exploited a vulnerability that was first discovered by the United States (US) National Security Agency (NSA) and later stolen and disseminated online (Fruhlinger, 2018). After breaching the computer, WannaCry encrypted that computer's files and rendered them unreadable. In order to recover the encrypted material, targets of the attack were told that they had to purchase special decryption software. The so-called "ransomware" siege affected individuals as well as large organizations, including the United Kingdom's National Health Service, Russian banks, Chinese schools, Spanish telecom giant Telefonica and the US-based delivery service FedEx. By some estimates, total losses approached $4 billion (Gottsegen, 2020; Sobers, 2020).

The discussion explores the legal risk and threat that the use of AI-ML technology presents to cybersecurity. The aim of the discussion is to provide an overview of the legal risk that AI presents to cybersecurity in general. Against this background, the legal risk that AI cybersecurity presents to a particular domain may be considered.

Many law enforcement agencies are employing AI-ML tools aimed at the prevention, detection and investigation of crimes committed in a physical world, for example predictive policing, CCTV cameras or drones equipped with face recognition technology and license plate readers to name but a few (Watney, 2019). In this discussion, the focus will be at securing cyberspace by means of AI-ML technology and the legal risk facing cybersecurity such as personal information protection, surveillance and cybercrime.

Cyberspace faces many risks in securing it. Cyberspace is the communication space made of network infrastructure (such as servers and cables), devices (like computers and smart phones), software (both human-to-machine and machine-to-machine interfaces) and data carried over the network (Paris Call for Trust and Security in Cyberspace, 2018). Although AI-ML technology is very useful in securing a safer and more secure digital environment, it is not infallible. AI-driven technology has a dual-use; it can be used for beneficial and harmful purposes (Crane, 2019 Ramachandran, 2019; Brundage et al, 2018). For example, police may use drones for law enforcement purposes, but a drone can be hacked and used as an instrument to commit a crime in a physical environment. Banks are more likely to receive phishing or ransomware attacks than being conventionally robbed. In 2018 the banking industry incurred the most cybercrime costs at $18,3 million. Industries that store valuable information such as healthcare and finance are usually bigger targets for hackers (Sobers, 2020).

Part 1 Understanding AI-ML cybersecurity technology

2. Conceptualising cybersecurity, AL and ML

Although an international and harmonised definition for cybersecurity, AL and ML do not exist, it is important to understand the meaning of concepts such as cybersecurity, AI and ML as it explains why AI and ML are expanding to cybersecurity.

The South African National Cybersecurity Policy Framework of 2015 defines “ccybersecurity” as the practice of making the networks that constitute cyberspace secure against intrusions, maintaining confidentiality, availability and integrity of information, detecting intrusions and incidents that do occur, and responding to and recovering from them.

AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages. ML is an application or subset of AI that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed (Gottsegen, 2019). ML uses algorithms derived from previous datasets and statistical analysis to make assumptions about a computer's behaviour. The computer can then adjust its action and even perform functions for which it has not been explicitly programmed (Gottsegen, 2019).

3. Advantages of AI-ML cybersecurity tools and techniques

- Organisations must be able to detect a threat in advance to be able to defend itself against such a threat. With its ability to sort through millions of files and identify potentially hazardous ones, ML is increasingly being used to pro-actively detect threats and automatically eliminate them before they can wreak havoc (Gottsegen, 2019; Ramachandran, 2019).
- AI can scan vast amounts of data and then use ML algorithms to look for patterns, learn how cyberattacks begin and guide human decision-makers on how to respond (Crane, 2019; Sparapani, 2019). Human security experts cannot match the speed and scale at which AI software can accomplish the data analysis tasks.
- AI cybersecurity technology can learn how cybercrime applications operate such as recognising deepfakes. Deepfakes use machine-learning AI technology to merge two videos together, such as swapping one person’s face for another to create completely new and artificial video and audio impersonations (Sparapani, 2019).
- Additionally, AI-based cybersecurity data analysis software can complete the task with consistently higher accuracy than human analysts (Press, 2019). Large-scale data analysis and anomaly detection are some of the areas where AI might add value in cybersecurity today (Bharadwa, 2019). When it comes to reducing errors and faults in the operational tasks and when it is about finding anomalies and irregularities, AI is way ahead of the human efficiency and capability.
- AI and ML may complement the shortage in human resources and save cost in cybersecurity (Brundage et al, 2018; Ramachandran, 2019).

It appears that defensive AI is the only force capable of combatting offensive AI attacks (Crane, 2019; Press, 2019).

4. Challenges facing AI cybersecurity

- Data is at the heart of any AI implementation. To be effective, AI algorithms must be driven by the right data systems. The data must not just exist but should be current. AI seeks to mimic human intelligence and should therefore be designed to continuously improve itself based on new knowledge. Identifying the required data sets must be the first thing a business does in their quest to operationalize the AI-driven cybersecurity algorithms (Stephen, 2019).
- As indicated, AI systems typically require large volumes of so-called training data to learn their functions. If the data used is biased, then the AI is going to understand only a partial view of the world and make decisions based on that narrow understanding (Choudhury and Lim, 2019). Bias can occur in three areas — the program, the data and the people who design those AI systems. Inherently biased AI programs can pose serious problems for cybersecurity at a time when hackers are becoming more sophisticated in their attacks (Choudhury and Lim, 2019; Press, 2019).

The legal risks are as follows:

- Businesses, institutions and governmental bodies are collecting and storing large amounts of personal data. Technologies such as the IoT and cloud computing are exacerbating the amount of available data. Individuals, businesses and government bodies must protect collected and stored data. In 2018 Mariott-Starwood disclosed a data breach in which 500 million consumers' data dating back to 2014 had been compromised (Sobers, 2020).
- Legislation such as the EU General Data Protection Regulation (GDPR) of 2018 imposes a legal duty on companies to secure private information. It puts the onus on the controller and processor to implement technical and organisational measures to ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected. Many countries have followed suit and have implemented similar data protection legislation.
- Data breaches expose sensitive information that often leave exposed users at risk for identity theft, and may impact negatively on a company's reputation. A company may be held liable for compliance violations (Sobers, 2020).
- In 2018 Liberty Holdings, an insurance company in South Africa, suffered a data breach which resulted in thousands of Liberty Holdings' investors' financial details stolen. The hackers had sought to extract money from the group by threatening to release confidential information. The South African Information Regulator indicated that Liberty Holdings had adequate cybersecurity measures in place and could not be held liable for the data breach. Liberty Holdings has European investors and therefore it had a legal duty to disclose the intrusion in terms of the GDPR.
- Surveillance is closely linked to personal information protection. AI systems enable the gathering of information by means of surveillance at a more efficient scale. AI-driven surveillance illustrates the dual nature of AI; it may pose a major risk to the privacy rights of individuals but likewise it can be used for preventing terrorism and other serious crimes (Brundage et al, 2018). The 2013 disclosure by former NSA contractor, Snowden, of mass surveillance conducted by the United States serves as a warning of the dangers inherent to surveillance.
- For every advance in cybersecurity, there are new cybercrime enhancements that set cybersecurity back again (Sparapani, 2019). Cyber criminals may exploit AI technology to commit a cybercrime (Ramachandran, 2019; Brundage et al, 2018; see part 2 hereafter). For example, data breaches where the perpetrator gains unauthorised access to personal information constitutes a cybercrime.

Part 2 AI and the commission of cybercrime

5. Impact of AI on cybercrime commission

The growing use of AI systems may cause the following changes in the threat landscape (Ramachandran, 2019; Brundage et al, 2018):

5.1 Modification of the nature of existing threats.

AI can hack into a system's vulnerability much faster and better than a human can (Sparapani, 2019). It can be used to disguise attacks so effectively that a business might never know that their network or device has been affected. Brundage et al (2018) expect AI-driven attacks to be more effective, more finely targeted, more difficult to attribute and more likely to exploit vulnerabilities in AI systems.

5.2 Expansion of threats and attacks

Brundage et al (2018) indicates that for many familiar attacks, the progress in AI will result in the expansion of the set of actors who are able to carry out the attack, the rate at which these actors can carry it out and the set of plausible targets.

The following examples illustrate the expansion of the threats and attacks:

- Phishing may extend to spear-phishing. Cybercriminals may use AI to direct their attack against a specific target. For example, a fully automated spear phishing system could create tailored tweets on the social media platform, Twitter, based on a user's demonstrated interests achieving a high rate of clicks to a link that could be malicious (Brundage, 2018). Not only does AI make spear-phishing more efficient but it also allows execution to take place at a much higher rate than if the process was manually run. This, in turn, expands the potential attack surface. Traditional signature-based security tools struggle to match up against this kind of attack environment. Such adversaries can arguably only be effectively contained by AI-powered security controls (Brundage, 2018).
- Threat actors use botnets — networks of infected computers or devices — for various cybercriminal purposes, most significantly distributed denial-of-service (DDoS) attacks against predefined targets. Today, botnets with DDoS capabilities are for sale on the Dark Web. IoT devices such as webcams, digital video recorders and routers can be used as bots. The adoption of 5G will enable a massive increase in connected devices and it will empower criminals to launch much larger botnet and DDoS attacks (Press, 2019).

5.3 Further advancement in AI can also give birth to new types of cyber threats.

The use of AI systems may give rise to new threats, such as:

- Business email compromise (BEC) attacks rely on emails that appear to come from ranking executives so the hackers can trick the recipient, often the chief financial officer, into transferring funds to the hacker (Sparapani, 2019). Threat actors can use AI to learn business users' communication patterns to mimic their writing styles and automatically compose convincing content. They are also using deepfake techniques, particularly audio impersonations of the executive in BEC attacks. For example, an unsuspecting employee (the victim) may receive an email purportedly from the chief financial officer to transfer funds to the hacker. This may be followed by a phone call from the hacker impersonating the financial officer thereby adding credibility to the ploy (Sparapani, 2019).
- A self-driving autonomous car is just like any other computer-enabled device, prone to a cybercrime (Elezaj, 2019). Criminals might be motivated to hack into the vehicles' operating systems and steal important passenger data, or else disrupt its operation and jeopardize the passenger's safety. As with any other hacking scenario, hacking into an autonomous car could expose a great deal of personal data, such as the passenger's destination. With this information, someone could potentially track the user with an aim towards robbery or assault (Elezaj, 2019). Another possibility is that an attack could be launched by means of an adversarial system causing the vehicle to crash (Brundage et al, 2019). For example, an image of a stop sign with a few pixels changed in specific ways, which a human would easily recognise as still being an image of a stop sign, might nevertheless be misclassified as something else entirely by an AI system and cause the self-driving vehicle to crash (Brundage, 2019).

6. Impact of AI on threat actors

As indicated, AI technology may enable larger scale and more numerous attacks (Brundage, 2018). The threat actors are not only non-state actors, but also nation-states or state-sponsored actors. It was alleged that the 2017 WannaCry ransomware attack could be attributed to North Korea but it was not confirmed (Fruhlinger, 2018).

The geopolitical landscape may drastically shift as nation-states or state-sponsored actors gradually increase their use of AI using complex sophisticated malware to attack one another's infrastructure (Press, 2019). Some countries are pursuing global dominance and in order to achieve it, they may use cyber espionage and disruptions, sometimes using smaller countries as proxies, in the goal of gaining political advantage (Press, 2019).

Countries are taking cyber threats seriously. In 2020 the UK indicated that a National Cyber Force (NCF), operated by the Ministry of Defence and GCHQ, will come into operation (Sengupta, 2020). It will conduct both proactive and retaliatory attacks against state and non-state actors accused of perpetrating cyberattacks on the UK and promoting "fake news" on social media, and shut down digital platforms used by terrorists and criminals. The NCF will operate alongside the National Cyber Security Centre (NCSC), which primarily concentrates on defensive cyber activities to protect government departments, strategic infrastructure and industry (Sengupta, 2020).

7. Recommendations in addressing the risk to cybersecurity from a legal perspective

Cybersecurity requires not only a technical response, but a multi-disciplinary and a multi-stakeholder's response. To effectively address the risk, a government needs the co-operation and assistance of, such as, the tech industry.

It is possible to have singled out a specific domain and to have made recommendations pertaining to that specific domain, for example, cyber-physical security or IoT security or Internet security. The purpose of this discussion however, is to provide an overview of the risk AI presents to cybersecurity in general. A common denominator that the domains share is the risk that cybercrime poses to cybersecurity.

Ethics, policy and law play a big role in ensuring an effective cybersecurity framework. Furthermore, it is suggested that businesses and government bodies should move towards a preventative approach to cybersecurity over a detection-focused approach (Press, 2019). This approach enables a business or government body to prevent an intrusion or otherwise allow them to mitigate the impact of the intrusion.

The recommendations are as follows:

- The cybersecurity industry, government and law enforcement agencies should take note of the possible impact of AI technology on crime commission, such as spear-phishing, hacking of a self-driving vehicle or the theft of datasets used for facial recognition (Brundage et al, 2018).
- Policymakers should collaborate closely with technical researchers to investigate, prevent and mitigate potential malicious uses of AI-driven technology.
- Brundage et al (2018) recommend that researchers and engineers in AI should take the dual-use nature of their work seriously. Researches should consider it their responsibility to take whatever steps they can to promote the beneficial uses of technology and prevent harmful uses.
- Legislation has to provide for the criminalisation of unlawful conduct as this ensures compliance with the legality principle in terms of which conduct, that is not recognised as a crime, cannot be investigated nor can the accused be prosecuted.
- The new ways that criminals may use AI technology such as deepfakes may necessitate legislation if existing legislation does not regulate it. As indicated, deepfake technology is false yet highly realistic AI-created media, such as a video showing people saying things they never said and doing things they never did.
- Deepfakes can have a negative impact on elections. During the 2016 US elections fake news and disinformation received a great deal of attention. In 2019 the US President Trump signed the first federal law related to deepfakes which requires the government to notify Congress of foreign deepfake-disinformation activities targeting US elections (Ferraro, Chipman and Preston, 2019). In 2019 US state, Virginia became the first state in the US to impose criminal penalties on the distribution of non-consensual, deepfake pornography. California enacted two laws that, collectively, allow victims of non-consensual, deepfake pornography to sue for damages and give candidates for public office the ability to sue individuals or organizations that distribute "with actual malice" election-related deepfakes without warning labels near election day (Ferraro, Chipman and Preston, 2019).
- In January 2020 China implemented new rules governing video and audio content online, including a ban on the publishing and distribution of fake news created with technologies such as AI and virtual reality. Any use of AI or virtual reality also needs to be clearly marked in a prominent manner and failure to follow the rules could be considered a criminal offense (Yang and Goh, 2019).
- Data protection is very important in a digital interconnected environment. As indicated, many countries have implemented data protection legislation. It is important that companies adhere to the legislation and that the government enforce compliance. A company must have a legal duty to disclose a data

breach. It cannot merely be a moral duty. A company must be held liable for inadequate or no cybersecurity which results in cybercrime.

- Companies leveraging IoT devices should focus on privacy and security (Burgess, 2016). As indicated, advancements in technology has an impact on crime commission, for example the adoption of 5G will result in a massive increase in connected devices and with the faster speeds and ultra-high bandwidth, DDoS attacks may increase. In future cybercriminals may use ransomware attacks targeting consumers through these devices especially where they lack built-in security controls (Press, 2019).
- IoT will remain insecure unless government steps in and fixes the problem (Brundage et al, 2018). A government could impose security regulations on IoT manufacturers, forcing them to make their devices secure and allowing people to sue them in instances where insecure devices resulted in a cybercrime. Any of these would raise the cost of insecurity and give companies incentives to spend money making their devices secure.
- There have been calls for the responsible disclosure of AI vulnerabilities to vendors (Brundage et al, 2018). For example, the NSA found a vulnerability in Microsoft software and wrote a code, called EternalBlue, to target it. However, in 2017 it was stolen and ultimately used in the 2017 WannaCry attack and other attacks such as the 2019 Baltimore ransomware attack (Perlroth and Shane, 2019).
- Brundage et al (2018) proposes that some types of AI research results should be subjected to some kind of risk assessment. Results should be shared selectively with a predetermined set of people and organisations that meet certain criteria such as the adherence to appropriate ethical norms. The reason being is that cybercriminals have access to the same publicly available AI technology that cybersecurity companies do and they may use it to commit cybercrimes (Sparapani, 2019). Selective disclosure of AI results may serve as a defensive measure against the risk of cybercrime.
- The use of AI systems to identify threats should be explainable and transparent which will ensure trust in the use of AI cybersecurity (Press, 2019). This may be achieved by means of ethical standards.
- In 2019 the Organisation for Economic Cooperation and Development (OECD), adopted the first intergovernmental standard for AI. It refers to five complementary values-based principles and five recommendations to governments. One of the recommendations is that organisations and individuals developing, deploying or operating AI systems should be held accountable for the proper functioning in line with the above principles.
- In 2019 the European Commission's high-level group on AI and ethics put forward a set of guidelines on how companies and governments should develop ethical applications of AI (European Commission, 2019). The guidelines list seven key requirements that AI systems should meet in order to be trustworthy. These requirements are applicable to different stakeholders partaking in the AI systems' life cycle, such as developers, deployers and end-users as well as the broader society. The 7 requirements are human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity non-discrimination and fairness, societal and environmental wellbeing and accountability.
- Countries need to work together pertaining to the protection of the safety of Internet users against state-nation attacks.
- Microsoft has proposed that governments adopt a Digital Geneva Convention to protect civilians on the Internet against nation-state attacks. Microsoft, which has been leading this initiative, has also called for an independent attribution body that would be tasked with determining who is responsible for an attack (Krahulcova and Mitnick, 2018).
- In 2018 there was the Paris Call for Trust and Security in Cyberspace which is based around nine common principles to secure cyberspace. It calls all cyberspace actors to work together and encourage states to cooperate with private sector partners, researchers and civil society to ensure responsible behaviour within cyberspace (Paris Call for Trust and Security in Cyberspace, 2018).
- It is commendable that a number of leading technology companies have begun to work together to address cybersecurity issues, launching the Cybersecurity Tech Accord (Tech Accord) and the Digital Geneva Convention (DGC).
- More than 30 companies have signed the Tech Accord. The companies that have signed the Tech Accord have agreed to resist nation-state and criminal attacks to protect their customers and have taken a step toward promoting corporate respect for the human rights of their users. The Tech Accord asks global tech companies not to assist governments in offensive operations, to protect technology against tampering, and to offer products that prioritize privacy and security. It does not identify the harms users face from these attacks nor does it address the proactive steps companies should take to protect users,

such as implementing data security and data privacy measures that would improve accountability across sectors. The Tech Accord could serve to apply more pressure on companies to respect human rights when developing and deploying their products (Krahulcova and Mitnick, 2018).

- As indicated, a company or IoT manufacturer should be held liable for an intrusion where inadequate or no cybersecurity technology was in place.
- Cyber-physical security is not only relevant in respect of autonomous vehicles but also in respect of robots and drones. Where a robot is used as an instrument to commit a crime, it is possible that the programmer or manufacturer or the person behind the robot could be held liable (Simmler and Markwalder, 2019). The legal issue of whether a robot itself that commits a crime can be held criminally responsible is a topic that is still undecided (Simmler and Markwalder, 2019).
- The prosecution of perpetrator(s) should be prioritised as it will serve as a deterrent for crime commission. Prosecution faces many legal challenges, such as attribution and the cross-border nature of the crime commission which may complicate the gathering of evidence.

8. Conclusion

AI is inherently data-driven and therefore, a host of issues and concerns arise, including cybersecurity, privacy, accuracy and bias associated with its use.

A business, a financial institution or a government body must implement cybersecurity as the consequence of inadequate or no cybersecurity will have dire repercussions for it and the affected Internet users. AI-ML cybersecurity technology and tools cannot be implemented without taking into consideration the impact of emerging and evolving technologies on the commission of cybercrime. As AI-ML technology will be used for the commission of cybercrime, the cybersecurity defence has to be AI-ML driven.

AI cybersecurity technology cannot be implemented as a technical solution in isolation, but it should be seen within a multi-disciplinary context. In this regard, the discussion focused on providing an overview of the legal risk AI presents to cybersecurity and the manner in which the legal risk may be counter-balanced by means of the law, ethics and policies.

For the unforeseeable future, AI-driven cybercrime will pose a huge legal risk to AI-driven cybersecurity.

References

Bharadwa, R. (2019) “Artificial Intelligence in Cybersecurity – Current Use-Cases and Capabilities”, [online], https://emerj.com/ai-sector-overviews/artificial-intelligence-cybersecurity/.

Burgess, C (2016) “How Will the Internet of Things Be Leveraged to Ruin Your Company’s Day? Understanding IoT Security”, [online], https://securityintelligence.com/will-internet-things-leveraged-ruin-companys-day-understanding-iot-security/.

Brundage, M. et al. (2018) “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation”, [online], https://arxiv.org/pdf/1802.07228.pdf.

Choudhury, S.R. and Lim, B. (2019) “A.I. has a bias problem that can be a big challenge in cybersecurity”, [online], https://www.cnbc.com/2019/07/17/ai-has-a-bias-problem-that-can-be-a-big-challenge-in-cybersecurity.html. Crane, C. (2019) “Artificial intelligence in cyber security: The saviour or enemy of your business?” [online], https://www.thesslstore.com/blog/artificial-intelligence-in-cyber-security-the-savior-or-enemy-of-your-business/.

Elezaj, R. (2019) “Cybersecurity or cyberthreats”, [online], https://www.machinedesign.com/mechanical-motion-systems/article/21837958/autonomous-cars-safety-opportunity-or-cybersecurity-threat.

European Commission (2019) “Ethical Guidelines for Trustworthy AI”, [online], https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1#Human%20agency.

Ferraro, M.F, Chipman, J.C. and Preston, S.W. (2019) “United States: First Federal Legislation On Deepfakes Signed Into Law”, [online], http://www.mondaq.com/unitedstates/x/878878/Terrorism+Homeland+Security+Defence/First+Federal+Legislation+On+Deepfakes+Signed+Into+Law.

Fruhlinger, J. (2018), “What is WannaCry ransomware, how does it infect and who was responsible”, [online], https://www.csoonline.com/article/3227906/what-is-wannacry-ransomware-how-does-it-infect-and-who-was-responsible.html.

Gottsegen, G. (2019) “Machine learning cybersecurity: how it works and companies to know”, [online], https://builtin.com/artificial-intelligence/machine-learning-cybersecurity.

Krahulcova, L. and Mitnick, D. (2018) “A digital Rights Approach to the Tech Accord and the Digital Geneva Convention”, [online], https://www.accessnow.org/why-the-digital-geneva-convention-needs-more-work-to-protect-human-rights/.

Paris Call for Trust and Security in Cyberspace. (2018), [online], https://pariscall.international/en/.

Ramachandran, R. (2019) "How Artificial Intelligence is changing Cyber security landscape and preventing cyber-attacks", [online], https://www.entrepreneur.com/article/339509.

Perlroth, N. and Shane, S. (2019) "In Baltimore and beyond; a stolen N.S.A tool wreaks havoc", [online], https://www.nytimes.com/2019/05/25/us/nsa-hacking-tool-baltimore.html.

Ramachandran, R. (2019) "How Artificial Intelligence is changing Cyber security landscape and preventing cyber-attacks", [online], https://www.entrepreneur.com/article/339509.

Press, G. (2019) "141 Cybersecurity predictions for 2020", [online], https://www.forbes.com/sites/gilpress/2019/12/03/141-cybersecurity-predictions-for-2020/#5d74ff4b1bc5.

Sengupta, K. (2020) "UK is nearly ready to launch force to hit hostile countries with cyberattacks", [online], https://www.independent.co.uk/news/uk/home-news/cyber-warfare-security-force-iran-crisis-ministry-of-defence-a9278591.html**.**

Simmler, M. and Markwalder, N. (2019) "Guilty robots?-Rethinking the nature of culpability and legal personhood in an age of artificial intelligence", *Criminal Law Forum*, pp 1 – 31.

Sobers, R. (2020) "110 Must-know cybersecurity statistics for 2020", [online], https://www.varonis.com/blog/cybersecurity-statistics/.

South African National Cybersecurity Policy Framework. (2015) [online], https://www.gov.za/sites/default/files/gcis_document/201512/39475gon609.pdf.

Sparapani, J. (2019) "An innovation war: Cybersecurity vs cybercrime" MIT Technology Review Insights, [online], https://mittrinsights.s3.amazonaws.com/hpe-cyber.pdf.

Stephen, M.W. (2019) "Cybersecurity AI: Integrating artificial intelligence into your security policy" [online], http://techgenix.com/cybersecurity-ai/.

Watney, M. (2019) "Law Enforcement Use of Artificial Intelligence for Domestic Security: Challenges and Risks", Paper read at the *Proceedings of the European Conference of Artificial Intelligence and Robotics (ECIAR)*, Oxford, UK, November, pp 341 – 348.

Yang, Y. and Goh, B. (2019) "China seeks to root out fake news and deepfakes with new online content rules**",** [online], https://www.reuters.com/article/us-china-technology/china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU.

Smart City IoT: Co-Designing Trustworthy Public Safety Drones for Indoor Flight Missions

Harry Wingo
National Defense University, Washington, D.C., United States of America
harry.m.wingo.civ@ndu.edu

DOI: 10.34190/EWS.20.509

Abstract: This paper proposes a collaborative design ("co-design") approach to accelerating public and political acceptance of the cyberspace and information risks, inherent to the development and deployment of indoor smart building unmanned aircraft systems (UAS) to provide immediate "visual access" to law enforcement and other armed first responders in the case of an active shooter incident inside of a building, including risks concerning (1) system safety and reliability; (2) supply chain security; (3) cybersecurity; and (4) privacy. The paper explores the potential to use a virtual world platform as part of a phased pilot at the U.S. Service Academies to build trust in the privacy, security and efficacy of a counter active shooter UAS (CAS-UAS), trust necessary to win public and political support for procuring, developing, and deploying such a system. By using a live-virtual-constructive approach similar to the U.S. military's synthetic training environments (STE), to intentionally engage a cross-functional team in a pilot to field a prototype CAS-UAS, a co-created CAS-UAS could anticipate and address security and privacy concerns of the sort that slowed the roll-out of police body-worn cameras (BWC) and wide area motion imagery (WAMI), particularly in the context of privacy concerns over the use of facial recognition. The paper is intended to provide communities with a new approach to identify and address policy obstacles to deploying smart city IoT technology to save lives, here in the case of a CAS-UAS that can enhance situational awareness for first responders during active shootings inside of public buildings like shopping malls, corporate headquarters, houses of worship or schools. The paper draws on the author's personal experience testifying before the U.S. Congress concerning recent questions about the supply chain security of small commercial drones.

Keywords: Data Science; Public Safety; Cloud and Edge Computing; Artificial Intelligence; Transparent, Usable, and Comprehensive Privacy Policy Systems; 3D Visualization Cyberspace: The New Battlefield Cyber Manoeuvre Warfare; Cyber-Security in Smart Cities; Privacy

1. Introduction: Using Cyber-Physical Ranges to Co-Design Privacy-Aware "Life Drones"

Smart city Internet of Things (IoT) technologies include sensor networks able to provide unprecedented situational awareness to public safety teams, like police body-worn cameras (BWC) and wide-area motion imagery (WAMI) that enhance situational awareness for first responders but also raise privacy concerns. (Purser 2019; Michel 2019). These concerns might be addressed through collaborative design or "co-design" (also "participatory design") that enables communities to observe and contribute to urban challenges (Choque et al. 2019; Castelnovo et al. 2015). Smart city co-design through "virtual world" platforms will allow citizens to actually contribute to the privacy-aware development of public safety technology like BWC and WAMI.

Privacy concerns are also raised by law enforcement use of unmanned aircraft systems (UAS), more widely known as drones. The surveillance capabilities of small UAS (sUAS) is less powerful than WAMI, but collectively, the data collected is a matter of privacy concern. A specialized use case for sUAS -- providing situational awareness during an active shooting *inside* of buildings -- would likely rely on an installed network of security cameras, and perhaps even smart gunshot detectors. Such a counter active shooter UAS (CAS-UAS) could protect high-occupancy buildings like shopping malls, corporate offices, sports venues, places of worship or schools by allowing "life drones" to instantly provide first responders access to an active shooting (Wingo 2018). While a CAS-UAS could greatly enhance situational awareness, it will predictably face privacy concerns similar to those facing BWC and WAMI. These concerns might be addressed through a serious game approach to engaging CAS-UAS stakeholders: Law enforcement, building owners and operators, daily occupants and the community in which the smart building protected by a life drone system would be installed.

A CAS-UAS co-design platform could demonstrate the value of life drones, particularly those using "stream of life" data like facial features, gait and other personal details about every person within a building. Using differential privacy methods to train deep learning (DL) systems can allow life drones to more rapidly locate an active shooter, and to also provide more relevant and timely information to first responders about what is happening (or may happen) during the chaos of an active shooting. This can save lives. (Nunavath 2017).

2. Co-Designing Privacy-Aware Life Drones: A Live, Virtual, Constructive (LVC) Approach

Co-design platforms call to mind crowdsourcing tools like the University of Washington's Foldit (Curtis 2015), the more-scientifically accurate Foldit Standalone (Kleffner 2017), and Google's CAPTCHA security tool that enlists users to review computer images before being granted access to a website (Dzieza 2018). In the U.S. Department of Defense (DoD), collaborative training and experiment can be conducted through live, virtual, constructive (LVC) (DoD n.d., 119) platforms like the Atterbury-Muscatatuck Urban Training Center, "CyberTropolis" (DoD-Muscatatuck n.d.). CyberTropolis includes a multi-story hospital that might allow life drones to be tested in a live-fire and electronic warfare environment. A "digital twin" of CyberTropolis might allow law enforcement anywhere in the U.S. to fly virtual missions in order to test whether a CAS-UAS being considered by a local hospital would be up to standards adopted in other jurisdictions.

The virtual aspects of a co-design platform may be the most compelling for demonstrating the balance between efficacy and privacy. Consider for example that a 3D simulated version of the State Capitol for each of the United States might be hosted on a serious game platform. Providing the functionality of a multiplayer first-person-point of view (FPV) drone simulator like LiftOff, Velocidrone, DRL (Drone Racing League) or FPV Air 2 (OSPrey 2019), tailored to the specifics of each State Capitol complex would allow first responders, drone operators and policy makers to experience through serious gameplay the anticipated functionality of a CAS-UAS. The CAS-UAS simulator, as part of a CAS-UAS co-design platform, could also gather feedback in several ways from those who engaged with the platform (Lee et al. 2020).

If designed with imminent emerging technology like 5G and artificial intelligence (AI) in mind, a co-design platform could grow along with near- and long-term developments in the CAS-UAS space. It might assist the development of ethical AI, as intended DARPA's Urban Reconnaissance through Supervised Autonomy (URSA) program (Root n.d.), or even contribute to the trek towards "Third Wave AI" (DARPA n.d.). Harvesting the human workflow within a co-designed life drone platform might provide insight for taking AI beyond current deep learning (DL) approaches and towards a machine form of "common sense" (Marcus 2019; Mitchell 2019). Also, differential privacy in the context of AI is a matter of growing concern, and is increasingly researched. (Rohringer 2019), along with the trustworthiness of autonomous machines (Michael 2019; Scharre 2018). A forward-looking co-design platform, if properly designed, could eventually help with these future-oriented challenges.

3. A National Capital Region Testbed at the U.S. Naval Academy & University of Maryland

The State of Maryland is home to two academic institutions with a nexus of national security excellence and proximity to the cyber and drone talent to launch a compelling pilot of the co-design platform discussed in this paper. The U.S. Naval Academy in Annapolis, Maryland, has a cybersecurity degree program that has demonstrated interest in allowing students (called Midshipmen, a category of officer candidate) to explore the cybersecurity and engineering challenges related to small drones, including their potential use for force protection or in the context of counter-UAS (C-UAS) operations. The University of Maryland (College Park), has the nation's largest number of computer science undergraduate students and, in addition to an on-campus drone fabrication lab, has demonstrated world-class innovation in the UAS field through its worlds-first delivery of a human organ by drone (U.Maryland 2018). The University of Maryland's UAS Test Site is connected to other talent and research clusters in the DC region and beyond (UAS Test Site n.d.).

The Naval Academy and U. Maryland would find many stakeholder partners in the National Capital Region, like the U.S. Department of Commerce's National Institute of Standards and Technology (NIST), which has spearheaded an initiative to spur co-design of security and privacy for smart cities (NIST-Global Cities 2019; Global Cities Challenge n.d.), as well as the Department of Homeland Security, Army Research Lab, DARPA, and the Naval Research Labs. As for the private sector, Maryland-based Lockheed Martin is spearheading AI-enabled drone innovation through its AlphaPilot competition (Lockheed n.d.)

Between the Naval Academy and U. Maryland, the former would likely have fewer privacy-related concerns raised by faculty and the student body. As a working military facility for officer candidate training, Annapolis would be able to proceed with an LVC co-design pilot that engaged Midshipmen without the potential objections that might arise within the U. Maryland campus community.

4. The China-USA Privacy Gap: A Strategic Case for Co-Designing Smart City IoT Privacy

In June 2018, the author testified before Congress concerning the strategic advantage that Chinese companies have in the market for commercial drones (Wingo-Drones 2018). The over 80 percent market share that Chinese drone companies currently enjoy within the U.S. is akin to China's challenge of the U.S. in other smart city technologies that implicate privacy, like the AI-enabled computer vision algorithms and micro-electronics behind high-tech cameras offered by Hikvision, or the next-generation radio technology fueling China's growing push into 5G.

Congress and other policymakers are sounding the alarm and taking legislative action about the risks to U.S. security presented in the UAS context. For example, the recently passed National Defense Authorization Act (NDAA 2020) bans DoD from purchasing or operating Chinese drones (Host 2019). Another bill, by Florida Senator Rick Scott, the "American Security Drone Act of 2019" seeks to ban any Federal procurement of Chinese UAS (Vavra 2019). A co-design platform might allow immediate feedback to collaborators about the supply chain details and risks attending particular drones or other CAS-UAS elements. The co-design platform might also underscore the privacy shortcomings of non-U.S. supply chains, a function that could bolster the case for building the overall U.S. commercial sUAS industrial base.

5. Conclusion

Choque et al. propose developing tools to allow smart city service providers to keep pace with the need for novel solutions that can be validated by citizens (2019 p.1). By developing a web-based, co-design platform to allow the co-creation of privacy-aware life drones, communities may more rapidly support the use of such public safety technology even if it involves surveillance. Unlike the "constant stare" persistent surveillance technology of WAMI (Bishop 2018), as currently proposed for deployment in the city of Baltimore, a life drone co-design pilot at the U.S. Naval Academy in Annapolis would present an easier case in the context of privacy. Also, the platform used to co-create differential privacy for CAS-UAS may also set the stage for the co-creation of "common sense" AI, the goal of DARPA's "third wave AI" efforts. Finally, co-designing smart city IoT may help to address the strategic gap in training data that China is able to gather from its cities as compared to the United States, and our Western allies and partners with similar privacy traditions.

Disclaimer. The opinions, conclusions, and recommendations expressed or implied are the authors' and do not necessarily reflect the views of the Department of Defense or any other agency of the U.S. Federal Government, or any other organization.

References.

Bishop, T. (2018) *Secret surveillance a hard sell in Baltimore*. Baltimore Sun. Available from: https://www.baltimoresun.com/opinion/op-ed/bs-ed-op-0427-bishop-surveillance-20180426-story.html.

Castelnovo, W., Misuraca, G. and Savoldelli, A. (2015) *Citizen's engagement and value co-production in smart and sustainable cities*. In International conference on public policy (pp. 1-16). Milan. Available from: https://www.ippapublicpolicy.org/file/paper/1433973333.pdf.

Choque, J., Diez, L., Medela, A. and Muñoz, L. (2019) *Experimentation Management in the Co-Created Smart-City: Incentivization and Citizen Engagement*. *Sensors*, *19*(2), p.411. Available from: https://www.mdpi.com/1424-8220/19/2/411/pdf.

Curtis, V. (2015) *Motivation to participate in an online citizen science game: A study of Foldit. Science Communication*, 37(6), pp.723-746.

Defense Advanced Research Projects Agency (DARPA) (n.d.) *AI Next Campaign.* Available from: https://www.darpa.mil/work-with-us/ai-next-campaign.

Dzieza, J. (2019) *Why CAPTCHAs have gotten so difficult: Demonstrating you're not a robot is getting harder and harder.* Verge. Available from: https://www.theverge.com/2019/2/1/18205610/google-captcha-ai-robot-human-difficult-artificial-intelligence.

Global City Teams Challenge (n.d.) *Smart Regions Collaborative.* Available from: https://pages.nist.gov/GCTC/smart-regions/.

Host, P. (December 30, 2019) *US bans Pentagon use, procurement of Chinese UAVs.* Jane's International Defence Review. https://www.janes.com/article/93423/us-bans-pentagon-use-procurement-of-chinese-uavs.

Kleffner, R. et al. (2017) *Foldit Standalone: a video game-derived protein structure manipulation interface using Rosetta*, Bioinformatics, Volume 33, Issue 17, 01 September 2017, Pages 2765–2767. Available from: https://doi.org/10.1093/bioinformatics/btx283.

Lee, K., Gibson, J., and Evangelos A. Theodorou, E.A. (2020) *Aggressive Perception-Aware Navigation using Deep Optical Flow Dynamics and PixelMPC*. Available from: https://arxiv.org/pdf/2001.02307.pdf.

Lockheed Martin (n.d.) AlphaPilot -- *Lockheed Martin AI Drone Racing Innovation Challenge*. Available from: https://www.lockheedmartin.com/en-us/news/events/ai-innovation-challenge.html.

Marcus, G. and Davis, E. (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Michael, J.B. (Nov/Dec 2019) *Trustworthiness of Autonomous Machines in Armed Conflict.* IEEE, Security & Privacy, p. 4-5. Available from: https://ieeexplore.ieee.org/document/8886896.

Mitchell, M. (2019) *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.

Murison, M. (November 5, 2019) *Interview: Skydio CEO Adam Bry on Autonomy, Scaling Up & Taking on DJI*. Available from: https://dronelife.com/2019/11/05/interview-skydio-ceo-adam-bry/.

OSPrey (March 4, 2019) *FPV Drone Simulators: Everything You Need to Know.* getfpv.com. Available from: https://www.getfpv.com/learn/fpv-essentials/fpv-drone-simulators/ (Accessed on 17 January 2020).

Johns Hopkins University (JHU) (2016) *Johns Hopkins scientists brings down hobby drones to expose security flaws*. Available from: https://isi.jhu.edu/2016/06/08/johns-hopkins-scientists-brings-down-hobby-drones-to-expose-security-flaws/#.XiDIR8hKiM8.

National Institute of Standards and Technology (NIST) (2019) *Global City Teams Challenge Smart and Secure City Expo Convenes July 10-12 in Washington, D.C. Cybersecurity and privacy included in 2019 problem-solving goals*. Available from: https://www.nist.gov/news-events/news/2019/06/global-city-teams-challenge-smart-and-secure-city-expo-convenes-july-10-12.

Nunavath, V., Radianti, J., Comes, M. and Prinz, A. (2015) *Visualization of Information Flows and Exchanged Information: Evidence from an indoor fire game*. Available from: https://uia.brage.unit.no/uia-xmlui/bitstream/handle/11250/2374548/Nunavathetal_ISCRAM2015.pdf?sequence=3.

Parslow, G.R. (2013) *Commentary: Crowdsourcing, foldit, and scientific discovery games. Biochemistry and Molecular Biology Education*, 41(2), pp.116-117. Available from: http://www.academia.edu/download/49738438/pdf.pdf.

Purser, B. and Adams, K. (2019) *Who is really being watched by police body cameras?* Marketplace. Available from: https://www.marketplace.org/2019/11/12/who-is-really-being-watched-by-police-body-cameras/.

Riazi, Rouhani and Kouahnfar, (2019) *Deep Learning on private data,* IEEE Security & Privacy.

Rohringer, T.J., Budhkar, A. and Rudzicz, F (2019) *Privacy versus artificial intelligence in medicine*. University of Toronto Medical Journal, 96(1). Available from: https://pdfs.semanticscholar.org/6203/0e2258fd15e1861a346ecbfa0baaf1ed40d6.pdf.

Root, P. (n.d.) *Urban Reconnaissance through Supervised Autonomy (URSA)*. Available from: https://www.darpa.mil/program/urban-reconnaissance-through-supervised-autonomy.

Scharre, P. (2018) *Army of none: autonomous weapons and the future of war.* WW Norton & Company.

Togelius, J. (2018) *Playing Smart: On Games, Intelligence, and Artificial Intelligence.* MIT Press. Cambridge, MA.

University of Maryland School of Medicine. (April 26, 2019) *The University of Maryland is the first to use a drone to complete a successful kidney transplant.* YouTube. https://www.youtube.com/watch?v=7QbQsBt5kHg.

U.S. Congress (n.d.) S.1790 - National Defense Authorization Act for Fiscal Year 2020. 12/20/2019, Became Public Law No: P.L. 116-92. Congress.gov. Available from: https://www.congress.gov/bill/116th-congress/senate-bill/1790.

U.S. Department of Defense (DoD) (n.d.) *Modeling and Simulation (M&S) Glossary.* Available from: www.acqnotes.com/Attachments/DoD%20M&S%20Glossary%201%20Oct%2011.pdf.

U.S. Department of Defense (DoD-Muscatatuck) (n.d.) *Muscatatuck Urban Training Center - As Real as it Gets.* https://www.atterburymuscatatuck.in.ng.mil/Muscatatuck/About-Muscatatuck-Urban-Training-Center/.

Vavra, S. (2019) *Republican senators ask DOT, FAA to cease using Chinese drones*. CyberScoop. Available from: https://www.cyberscoop.com/dji-letter-department-of-transportation-federal-aviation-administration/.

Verhoef, P.C., van Doorn, J., and Beckers, S.F.M. (2013) *Understand the Perils of Co-Creation. Harvard Business Review*. Available from: https://hbr.org/2013/09/understand-the-perils-of-co-creation.

Wingo, H. (2018) *Beyond the Loop: Can Cyber-Secure, Autonomous Micro-UAVs Stop Active Shooters?*. In *International Conference on Cyber Warfare and Security* (pp. 497-502). Academic Conferences International Limited.

Wingo, H. (2019) *Statement of Mr. Harry Wingo. Drone Security: Enhancing Innovation and Mitigating Supply Chain Risks*. Available from: https://www.commerce.senate.gov/2019/6/drone-security-enhancing-innovation-and-mitigating-supply-chain-risks.

Winter, M. et al. (2020) *Learning to Read by Learning to Write: Evaluation of a Serious Game to Foster Business Process Model Comprehension*. Available from: https://games.jmir.org/2020/1/e15374/pdf.

WPTV News (November 7, 2018) *Fort Pierce Central High School students use drone technology to keep fellow students safe.* Available from: https://www.youtube.com/watch?v=7s_UXPwFOH8. (Accessed on 15 January 2020).

The Evolution of Ransomware Variants

Ashley Charles Wood and Thaddeus Eze
University of Chester, UK
ashley.wood@chester.ac.uk
t.eze@chester.ac.uk

DOI: 10.34190/EWS.20.027

Abstract: This paper investigates how ransomware is continuing to evolve and adapt as time progresses to become more damaging, resilient and sophisticated from one ransomware variant to another. This involves investigating how each ransomware sample including; Petya, WannaCry and CrySiS/Dharma interacts with the underlying system to implicate on both the systems functionality and its underlying data, by utilising several static and dynamic analysis tools. Our analysis shows, whilst ransomware is undoubtedly becoming more sophisticated, fundamental problems exist with its underlying encryption processes which has shown data recovery to be possible across all three samples studied whilst varying aspects of system functionality can be preserved or restored in their entirety.

Keywords: Ransomware, WannaCry, Petya, Dharma, CrySiS, Malware analysis

1. Introduction:

Ransomware is an ever-evolving threat, which in recent times is becoming more persistent and resilient as time progresses. The WannaCry ransomware in 2017 was responsible for widespread disruption within the NHS, leading to 80 NHS trusts, 595 GP Practices and 8 other NHS organisations being adversely affected, which resulted in implications upon patient care in one third of NHS trusts in England, with 34 trusts locked out of devices and a further 46 not believed to be infected but reporting disruption (Smart, 2018). The Petya ransomware also caused substantial disruption in 2017, with organisations such as Merck Pharmaceuticals facing disruption to research, manufacturing and sales of products costing around $670 million in damages (Davis, 2017). Petya impacted upon shipping companies such as; Maersk and TNT express, who subsequently suffered substantial disruption and incurred nine figure costs as a result (Greenberg, 2018). CrySiS/Dharma continued to be the most prevalent form of ransomware in Q1 2019 (Coveware 2019) with attacks increasing by a margin of 148% from February to April 2019, potentially due to CrySiS/Dharma's use of multiple attack vectors (Arntz, 2019).

This study carried out an in-depth behavioural analysis of ransomwares such as; WannaCry, Petya and CrySiS using a variety of static and dynamic analysis tools to reveal the underlying functionality of each to ascertain how each interacts with the system and its underlying data. To achieve this, a virtual machine environment running Windows 7 SP1 was created within VMWare Workstation, each of the samples and a selection of sample data was then populated onto the virtual machine and each of the ransomware samples executed to allow the full effects and implications of each variant to be ascertained. Windows 7 was selected for this study as this was the operating system in widespread use across the NHS, when the devastating NHS WannaCry ransomware took place. Windows 7 was selected as 79% of organisations still have at least one operational Windows 7 machine (Sanders, 2019) whilst an additional 50% of organisations continue to use it as their primary operating system after security updates and support have ceased (Parmar, 2019).

This paper is organised as follows; the second section provides a general background and overview of previous authors work to assist the reader in understanding the problem of ransomware and the work done previously to examine ransomware, the third section provides an overview and analysis of the ransomwares behaviour and how this differs between variants to discuss the implications on both system functionality and underlying data. The fourth section discusses the propagation techniques of each of the ransomware variants to detail how each ransomware sample propagates from one system to another, the final section discusses the main key findings and conclusions of this study with recommendations for future work.

2. Previous Work

There are several works which have delved into the underlying implications on both systems and data of recent ransomware samples, all of which have come to different findings regarding the WannaCry, Petya and CrySiS ransomwares.

Suiche (2017) conducted analysis of the WannaCry ransomware, focusing on WannaCry's network propagation techniques and discovered it to utilise an exploit known as DoublePulsar in combination with the EternalBlue exploit as a means of propagating itself to additional systems. Should this fail, the alternative MS17-010 Microsoft vulnerability would be utilised to exploit the target host (Suiche, 2017). Akbanov, Logothetis & Vassilakis (2019) in their dynamic analysis of WannaCry, discovered the presence of a kill-switch domain which when activated would stop the propagation behaviour of WannaCry and if not active will allow WannaCry to continue execution. Concerning data, Ivanov, Mamedov and Sinitsyn (2017) detailed flawed encryption logic within WannaCry which will read contents of original files before encrypting their contents and saving it to a file with the .WNCRYT before deleting the original file.

Analysis by Secureworks (2017) found WannaCry additionally performs modifications to the Windows registry to maintain its persistence on the system, namely it will set its payload to execute upon every subsequent restart or log-on event and it will define its working directory in the registry. Concerning the encryption process the analysis by Secureworks (2017) found WannaCry will utilize a combination of RSA and AES algorithms to encrypt files. Upon encryption, the taskdl.exe process will periodically delete the original files without modification, the authors argue this process renders data recoverable forensically. Further, analysis of WannaCry's network functionality finds WannaCry installs the Tor software in the TaskData folder within the working directory of WannaCry, Tor is executed as the taskhsvc.exe process, this process is then utilised to connect to WannaCry's command and control servers (Secureworks, 2017).

A technical analysis by Hung & Joven (2017) shows Petya will implicate upon system functionality by overwriting the Master Boot Record (MBR) which renders the system unable to boot correctly following initial infection. The master file table of the system is then damaged following the first system restart, rendering it unable to boot and inoperable (Kroustrek, 2017). It is also reported that select variants of Petya can propagate to systems which have been patched against both the EternalBlue and EternalRomance vulnerabilities (Křoustek, 2017).

Microsoft (2017) conducted analysis into a Petya variant which contained worm-like capabilities and detailed some of Petya's network propagation behaviour. This work found this Petya variant will attempt to exploit the CVE-2017-0144 or EternalBlue (MS17-010) vulnerability and additionally the CVE-2017-0145 vulnerability to spread to out-of-date vulnerable machines. The variant will then attempt to use these exploits by generating SMBv1 packets, Microsoft (2017) finds if the exploit is successful the MBR will be overwritten and the machine will restart to display a false system message.

Analysis of the network functionality and behaviour of CrySiS/Dharma by Panda Security (2017), revealed that unlike Petya and WannaCry, CrySiS/Dharma utilised a remote desktop protocol (RDP) vulnerability to propagate to additional systems before proceeding to create registry keys to maintain system persistence. Further analysis by Abrams (2018) found CrySiS/Dharma exploits vulnerable systems on the internet remotely via port 3389 and then attempts to brute-force a connection in order to leverage the payload (Abrams, 2018). Panda Security (2017) however found no evidence of a propagation system within CrySiS/Dharma and suggest human interaction is explicitly required to execute CrySiS/Dharma.

A study by Maclejack and Yang (2018) into the CrySiS/Dharma ransomware, found CrySiS/Dharma on execution will firstly store its contents on the stack which are then loaded. It will then configure itself to automatically execute when the user logs in to Windows, allowing it to continually encrypt files upon every subsequent execution. Commands are then processed by the ransomware to enforce the deletion of shadow copies of files. Upon completing these steps, the ransomware will then proceed to start encrypting the contents of mapped network drives before starting to encrypt files on the root of the OS drive using AES Encryption.

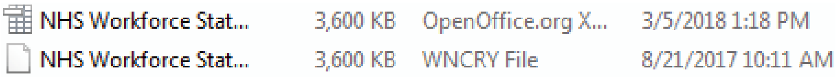
3. Behavioural Analysis

This section provides an overview and analysis of the Petya, WannaCry and CrySiS/Dharma ransomware variants and the implications on both system functionality and underlying data. The next section details the propagation techniques of each of the ransomware variants and examines how each ransomware sample propagates from one system to another.

3.1 WannaCry

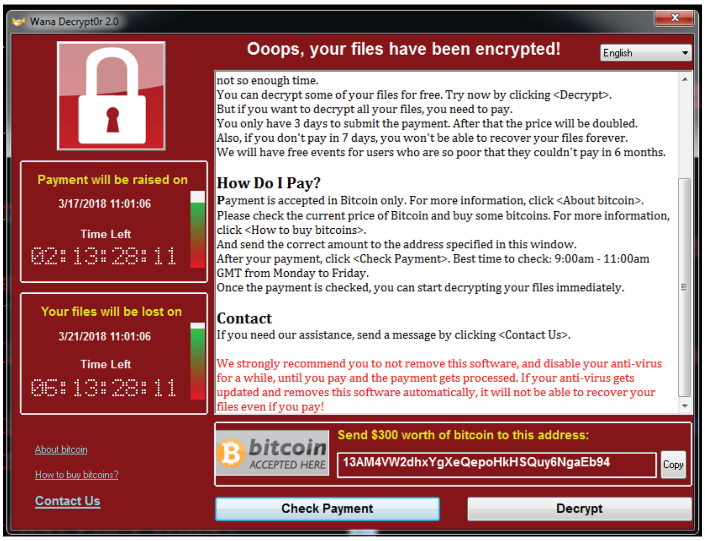
Firstly, two samples of the WannaCry ransomware were tested, the first is a sample containing only the WannaCry payload itself and the second a sample containing network propagation and the WannaCry software.

Upon execution of the first WannaCry sample, several files within the working directory of the WannaCry ransomware are created, which include; b.wnry, c.wnry, r.wnry, s.wnry, t.wnry and two executables taskdl.exe and taskse.exe. The populated files are then seen to be appended with the .wncryt extension, before an encrypted duplicate is created with the .wncry extension (Figure **1**) and the original file deleted. The WannaCry ransomware then displays its user interface (Figure 2) and demands a ransom consisting of $300 of bitcoin from the victim to decrypt the files. Later analysis of the drives data with data recovery tools, such as *Restoration* and *Recuva*, revealed many of the original files remained intact on the drive and as such proved recoverable, as the study showed that no modification of data occurred, prior to deletion.



NHS Workforce Stat...	3,600 KB	OpenOffice.org X...	3/5/2018 1:18 PM
NHS Workforce Stat...	3,600 KB	WNCRY File	8/21/2017 10:11 AM

Figure 1: Original file and encrypted duplicate



Ooops, your files have been encrypted!

not so enough time.
You can decrypt some of your files for free. Try now by clicking <Decrypt>.
But if you want to decrypt all your files, you need to pay.
You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.
We will have free events for users who are so poor that they couldn't pay in 6 months.

How Do I Pay?
Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.
Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.
And send the correct amount to the address specified in this window.
After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am GMT from Monday to Friday.
Once the payment is checked, you can start decrypting your files immediately.

Contact
If you need our assistance, send a message by clicking <Contact Us>.

We strongly recommend you to not remove this software, and disable your anti-virus for a while, until you pay and the payment gets processed. If your anti-virus gets updated and removes this software automatically, it will not be able to recover your files even if you pay!

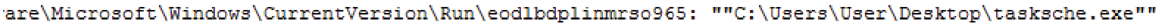
Payment will be raised on
3/17/2018 11:01:06
Time Left
02:13:28:11

Your files will be lost on
3/21/2018 11:01:06
Time Left
06:13:28:11

Send $300 worth of bitcoin to this address:
13AM4VW2dhxYgXeQepoHkHSQuy6NgaEb94

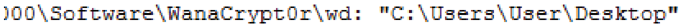
Figure 2: WannaCry user interface and ransom demand displayed following encryption process

Examining the implications upon the system, WannaCry was monitored with *RegShot*, which indicated WannaCry configures the system to automatically run the executable tasksche.exe from the location where WannaCry was first executed, in our test, the desktop (Figure 3). Further, WannaCry is observed to have defined the initial execution location as its working directory (Figure 4). Analysis of tasksche.exe reveals it to be a duplicate of the first sample and indicates how WannaCry maintains its persistence on the system.



```
are\Microsoft\Windows\CurrentVersion\Run\eodlbdplinmrso965: ""C:\Users\User\Desktop\tasksche.exe""
```

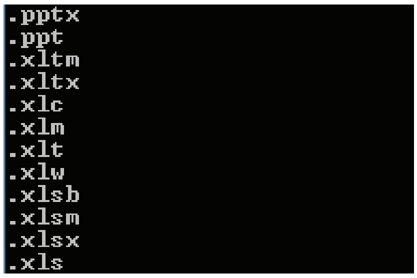
Figure 3: tasksche.exe set to run from the initial execution location



```
)00\Software\WanaCrypt0r\wd: "C:\Users\User\Desktop"
```

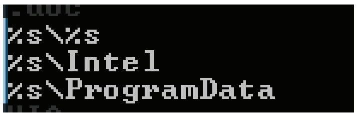
Figure 4: WannaCry defining its working directory

Our static analysis of WannaCry with the *Strings* tool, revealed references to 177 filetypes (Figure 5) and multiple key directories (Figure 6). Testing indicated WannaCry only encrypts files of these types and appears to exclude files contained within the referenced directories. All 177 filetypes found to be encrypted by WannaCry are shown in (Table 1).



```
.pptx
.ppt
.xltm
.xltx
.xlc
.xlm
.xlt
.xlw
.xlsb
.xlsm
.xlsx
.xls
```

Figure 5: References to multiple filetypes



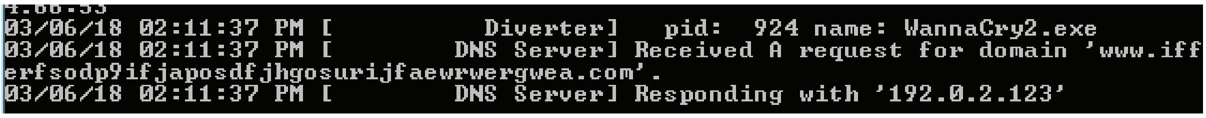
```
%s\%s
%s\Intel
%s\ProgramData
```

Figure 6: References to select directories

Table 1: 177 Filetypes encrypted by WannaCry

.der	.pfx	.key	.crt	.csr	.p12	.pem	.odt	.ott	.sxw
.stw	.uot	.3ds	.max	.3dm	.ods	.ots	.sxc	.stc	.dif
.slk	.wb2	.odp	.otp	.sxd	.std	.uop	.odg	.otg	.sxm
.mml	.lay	.lay6	.asc	.sqlite3	.sqlitedb	.sql	.accdb	.mdb	.db
.dbf	.odb	.frm	.myd	.myi	.ibd	.mdf	.ldf	.sln	.suo
.cs	.cpp	.pas	.asm	.js	.cmd	.bat	.ps1	.vbs	.vb
.pl	.dip	.dch	.sch	.brd	.jsp	.php	.asp	.rb	.java
.jar	.class	.sh	.mp3	.wav	.swf	.fla	.wmv	.mpg	.vob
.mpeg	.asf	.avi	.mov	.mp4	.3gp	.mkv	.3g2	.flv	.wma
.mid	.m3u	.m4u	.djv	.svg	.ai	.psd	.nef	.tiff	.tif
.cgm	.raw	.gif	.png	.bmp	.jpg	.jpeg	.vcd	.iso	.backup
.backup	.zip	.rar	.7z	.gz	.tgz	.tar	.bak	.tbk	.bz2
.PAQ	.ARC	.aes	.gpg	.vmx	.vmdk	.vdi	.sldm	.sldx	.sti
.sxi	.602	.hwp	.snt	.onetoc2	.dwg	.pdf	.wk1	.wks	.123
.rtf	.csv	.txt	.vsdx	.vsd	.edb	.eml	.msg	.ost	.pst
.potm	.potx	.ppam	.ppsx	.ppsm	.pps	.pot	.pptm	.pptx	.ppt
.xltm	.xltx	.xlc	.xlm	.xlt	.xlw	.xlsb	.xlsm	.xlsx	.xls
.dotx	.dotm	.dot	.docm	.docb	.docx	.doc			

Analysing the secondary samples behaviour with *FakeNet-NG*, reveals that upon execution, the WannaCry process will make attempts to connect to the www.ifferfsodp9ifjaposdfjhgosurijfaewrwergwea.com address before awaiting a response. This could be a potential kill-switch check which would correlate with the analysis performed by Akbanov, Vassilakis and Logothetis (2019).

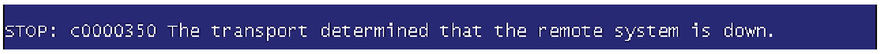


```
03/06/18 02:11:37 PM [           Diverter]   pid:  924 name: WannaCry2.exe
03/06/18 02:11:37 PM [         DNS Server] Received A request for domain 'www.iff
erfsodp9ifjaposdfjhgosurijfaewrwergwea.com'.
03/06/18 02:11:37 PM [         DNS Server] Responding with '192.0.2.123'
```

Figure 7: Potential kill-switch activity

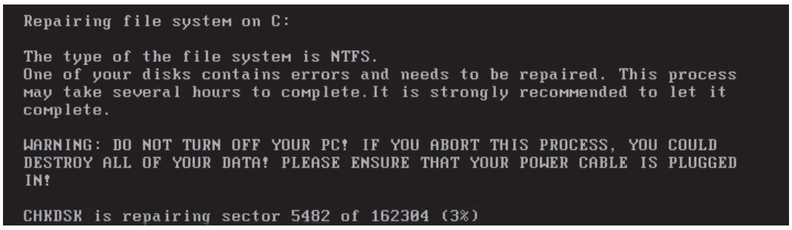
3.2 Petya

The analysis of Petya ransomware showed that upon execution, the Petya ransomware will crash the system (Figure 8) before proceeding to restart it and to begin the alleged CHKDSK process (Figure 9). Once this process concludes, a message is displayed upon every subsequent boot displaying a user interface and ransom demand (Figure 10), notably restricting access to the operating system and underlying data, rendering the system inoperable.



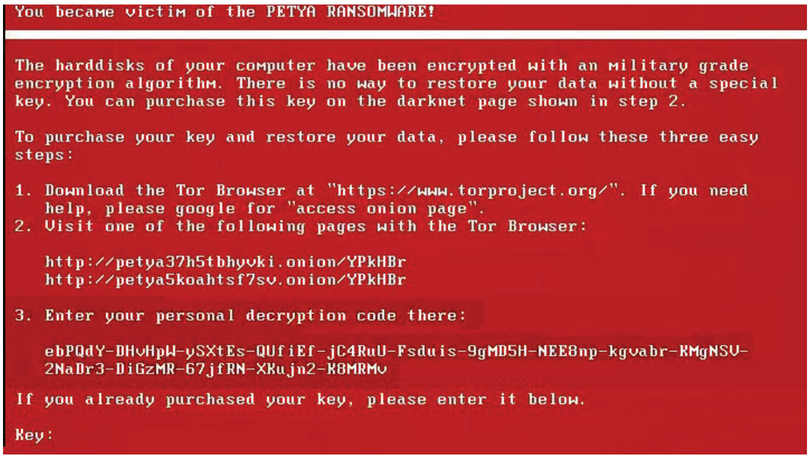
```
STOP: c0000350 The transport determined that the remote system is down.
```

Figure 8: System crashes shortly after execution



```
Repairing file system on C:

The type of the file system is NTFS.
One of your disks contains errors and needs to be repaired. This process
may take several hours to complete.It is strongly recommended to let it
complete.

WARNING: DO NOT TURN OFF YOUR PC! IF YOU ABORT THIS PROCESS, YOU COULD
DESTROY ALL OF YOUR DATA! PLEASE ENSURE THAT YOUR POWER CABLE IS PLUGGED
IN!

CHKDSK is repairing sector 5482 of 162304 (3%)
```

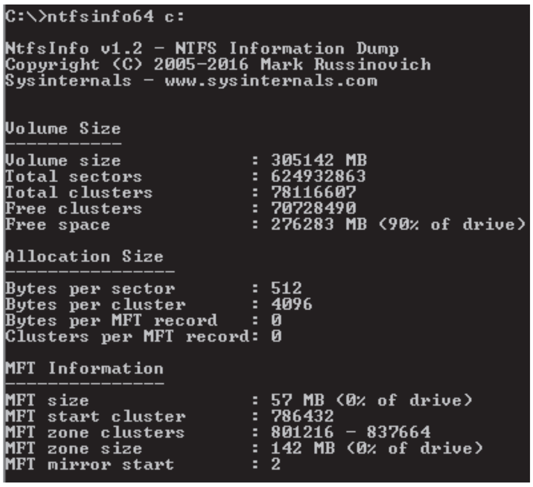
Figure 9: Falsified CHKDSK process



```
You became victim of the PETYA RANSOMWARE!

The harddisks of your computer have been encrypted with an military grade
encryption algorithm. There is no way to restore your data without a special
key. You can purchase this key on the darknet page shown in step 2.

To purchase your key and restore your data, please follow these three easy
steps:

1. Download the Tor Browser at "https://www.torproject.org/". If you need
   help, please google for "access onion page".
2. Visit one of the following pages with the Tor Browser:

   http://petya37h5tbhyvki.onion/YPkHBr
   http://petya5koahtsf7sv.onion/YPkHBr

3. Enter your personal decryption code there:

   ebPQdY-DHvHpW-ySXtEs-QUfiEf-jC4RuU-Fsduis-9gMD5H-NEE8np-kgvabr-KMgNSV-
   2NaDr3-DiGzMR-67jfRN-XKujn2-K8MRMv

If you already purchased your key, please enter it below.

Key:
```

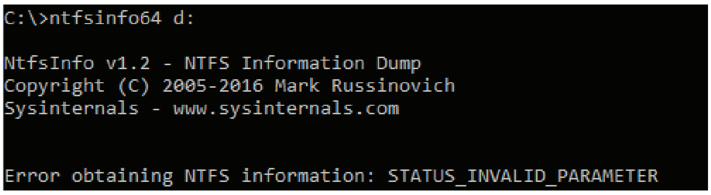
Figure 10: Resulting ransom demand displayed upon subsequent boots

This behaviour suggests that upon first examination Petya is inherently the most damaging of the ransomware samples due to Petya's claim of file encryption and loss of system functionality. To further investigate Petya's implications, execution was repeated with a physical machine environment and disk forensics applied at each stage of the execution process. This showed that if Petya was interrupted prior to the false CHKDSK process, the volume on the drive remained intact along with its data as indicated by ntfsinfo64 (**Figure 11**) which revealed that manual repairs to the MBR/BCD can be achieved at this stage to restore both system functionality and data. However, if the system proceeds into the CHKDSK process and allowed to complete, analysis showed that no information regarding the drive volume exists with ntfsinfo64 with examination resulting in an error (**Figure 12**). This indicates both MFT and volume information is lost, with any manual MBR/BCD restoration likely to prove problematic.



```
C:\>ntfsinfo64 c:

NtfsInfo v1.2 - NTFS Information Dump
Copyright (C) 2005-2016 Mark Russinovich
Sysinternals - www.sysinternals.com


Volume Size
-----------
Volume size              : 305142 MB
Total sectors            : 624932863
Total clusters           : 78116607
Free clusters            : 70728490
Free space               : 276283 MB (90% of drive)

Allocation Size
----------------
Bytes per sector         : 512
Bytes per cluster        : 4096
Bytes per MFT record     : 0
Clusters per MFT record: 0

MFT Information
---------------
MFT size                 : 57 MB (0% of drive)
MFT start cluster        : 786432
MFT zone clusters        : 801216 - 837664
MFT zone size            : 142 MB (0% of drive)
MFT mirror start         : 2
```

Figure 11: ntfsinfo64 shows record of original volume following initial execution

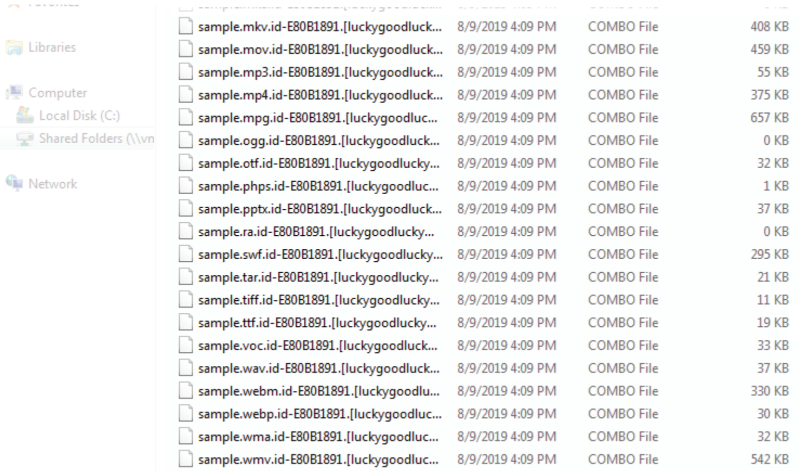


```
C:\>ntfsinfo64 d:

NtfsInfo v1.2 - NTFS Information Dump
Copyright (C) 2005-2016 Mark Russinovich
Sysinternals - www.sysinternals.com


Error obtaining NTFS information: STATUS_INVALID_PARAMETER
```

Figure 12: ntfsinfo64 reports an error immediately following complete Petya execution

3.3 CrySiS/Dharma

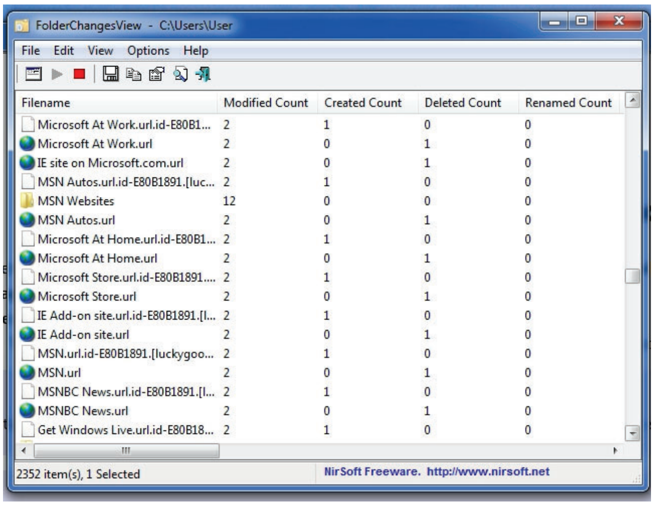
The final ransomware of this study is CrySiS, which ostensibly appears to be the least sophisticated of the studied ransomwares, however the analysis results indicate this ransomware sample offers lowest prospects to the victim regarding restoration of system functionality and data. CrySiS upon initial execution will firstly prioritise the encryption of resources upon mapped network locations (Figure 13) before encrypting the entirety of the local machines files, including both user files and program files, before presenting its user interface which results in impaired system functionality and data.



sample.mkv.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	408 KB
sample.mov.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	459 KB
sample.mp3.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	55 KB
sample.mp4.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	375 KB
sample.mpg.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	657 KB
sample.ogg.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	0 KB
sample.otf.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	32 KB
sample.phps.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	1 KB
sample.pptx.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	37 KB
sample.ra.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	0 KB
sample.swf.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	295 KB
sample.tar.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	21 KB
sample.tiff.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	11 KB
sample.ttf.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	19 KB
sample.voc.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	33 KB
sample.wav.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	37 KB
sample.webm.id-E80B1891.[luckygoodlu...	8/9/2019 4:09 PM	COMBO File	330 KB
sample.webp.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	30 KB
sample.wma.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	32 KB
sample.wmv.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	542 KB

Figure 13: Files encrypted firstly within network locations

During analysis, files were monitored with *FolderChangesView*, this confirmed that encrypted duplicates of files were created which were named using the original filenames and the extension .id-E80B1891.[luckygoodlucky@tuta.io].combo, before the original files were shown to be firstly modified and deleted (**Figure 11**). Analysis showed that immediately prior to deletion the original files had been stripped of their data, meaning each were 0kb compared to their original size as shown with program.py (**Figure 15**). Such behaviour mirrors that of WannaCry but takes additional steps to damage the integrity of original files prior to deletion, evidently to thwart data recovery attempts.

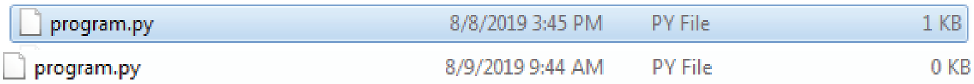


Filename	Modified Count	Created Count	Deleted Count	Renamed Count
Microsoft At Work.url.id-E80B1...	2	1	0	0
Microsoft At Work.url	2	0	1	0
IE site on Microsoft.com.url	2	0	1	0
MSN Autos.url.id-E80B1891.[luc...	2	1	0	0
MSN Websites	12	0	0	0
MSN Autos.url	2	0	1	0
Microsoft At Home.url.id-E80B1...	2	1	0	0
Microsoft At Home.url	2	0	1	0
Microsoft Store.url.id-E80B1891....	2	1	0	0
Microsoft Store.url	2	0	1	0
IE Add-on site.url.id-E80B1891.[l...	2	1	0	0
IE Add-on site.url	2	0	1	0
MSN.url.id-E80B1891.[luckygoo...	2	1	0	0
MSN.url	2	0	1	0
MSNBC News.url.id-E80B1891.[l...	2	1	0	0
MSNBC News.url	2	0	1	0
Get Windows Live.url.id-E80B18...	2	1	0	0

2352 item(s), 1 Selected

NirSoft Freeware. http://www.nirsoft.net

Figure 14: Encrypted duplicates of files created, before the original is modified and deleted



program.py	8/8/2019 3:45 PM	PY File	1 KB
program.py	8/9/2019 9:44 AM	PY File	0 KB

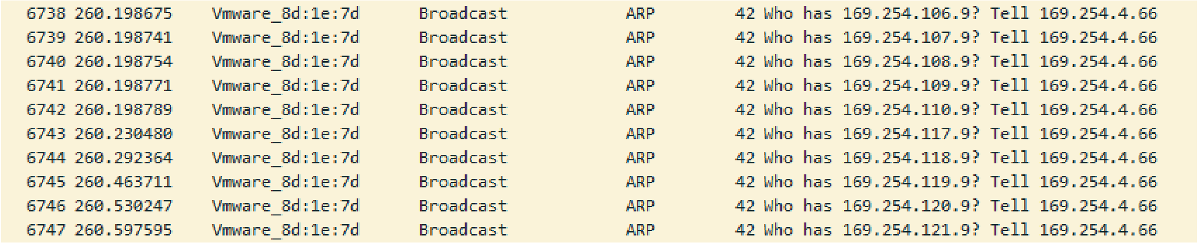
Figure 15: The file program.py before (above) and during (below) CrySiS execution

4. Propagation Techniques

This study investigated the network functionality and network propagation techniques of each ransomware variant, to establish how each can successfully propagate across networks from one system to another and further to ascertain if system functionality and data is adversely affected on other systems. In order to contain each ransomware variant, a virtual machine environment and virtualised network with multiple hosts were created to test and analyse any propagation behaviour present within each variant.

4.1 WannaCry Network Functionality

To test the network functionality of the WannaCry ransomware, two samples were obtained, the first being a sample containing the raw payload of WannaCry whilst the second was an adaption of the first but additionally incorporating network functionality and propagation behaviour. The second sample of WannaCry was permitted to execute within the prepared virtual machine and given access to the virtual network within VMWare, incoming and outgoing network activity was then closely monitored using Wireshark. The results show that upon execution the WannaCry sample attempts to sweep the entire address range of the local network to find other hosts in which the sample can propate itself on the network (Figure 16). Evidence suggests once a remote host is found on the network, the WannaCry sample will attempt to infect the remote host by leveraging a vulnerability known as EternalBlue within the windows SMB2 protocol, allowing it to effectively propagate itself to other hosts on the network.

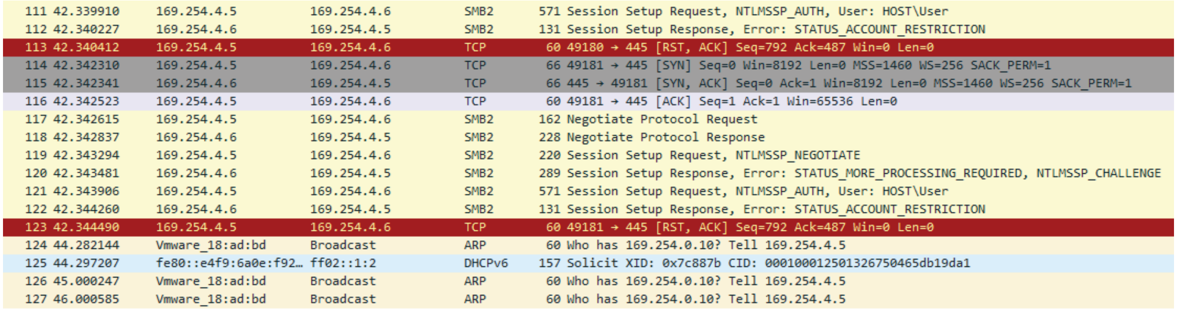


```
6738 260.198675   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.106.9? Tell 169.254.4.66
6739 260.198741   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.107.9? Tell 169.254.4.66
6740 260.198754   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.108.9? Tell 169.254.4.66
6741 260.198771   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.109.9? Tell 169.254.4.66
6742 260.198789   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.110.9? Tell 169.254.4.66
6743 260.230480   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.117.9? Tell 169.254.4.66
6744 260.292364   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.118.9? Tell 169.254.4.66
6745 260.463711   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.119.9? Tell 169.254.4.66
6746 260.530247   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.120.9? Tell 169.254.4.66
6747 260.597595   Vmware_8d:1e:7d    Broadcast      ARP     42 Who has 169.254.121.9? Tell 169.254.4.66
```

Figure 16: WannaCry ARP Protocol Traffic

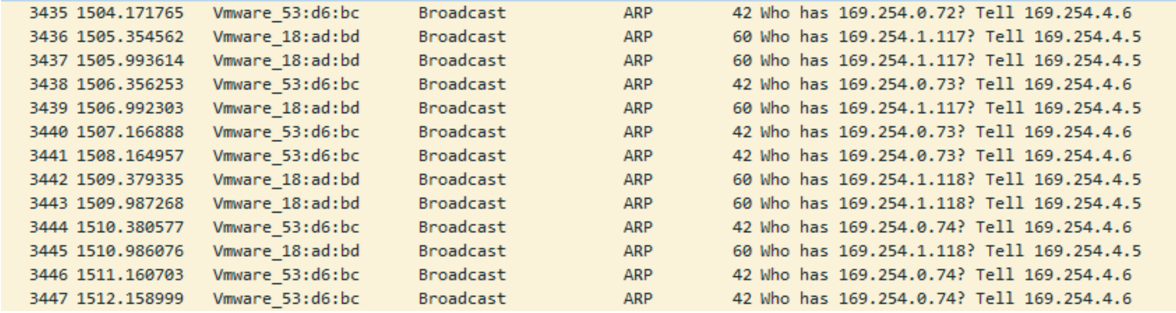
4.2 Petya Network Functionality

Petya's network activity was also studied for comparison, the results indicate Petya exhibits similar activity but upon finding a host to infect will begin to exchange substantial TCP and SMB2 traffic with the identified vulnerable host (Figure 17) to propagate Petya's payload from one system to another. The secondary infected host will proceed to sweep the network for additional hosts to infect together with the originally infected host (Figure 18) indefinitely until the system is powered off or restarted, at which point the falsified CHKDSK process will present itself and overwrite the master-file-table (MFT).



```
111 42.339910   169.254.4.5           169.254.4.6    SMB2     571 Session Setup Request, NTLMSSP_AUTH, User: HOST\User
112 42.340227   169.254.4.6           169.254.4.5    SMB2     131 Session Setup Response, Error: STATUS_ACCOUNT_RESTRICTION
113 42.340412   169.254.4.5           169.254.4.6    TCP       60 49180 → 445 [RST, ACK] Seq=792 Ack=487 Win=0 Len=0
114 42.342310   169.254.4.5           169.254.4.6    TCP       66 49181 → 445 [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=256 SACK_PERM=1
115 42.342341   169.254.4.6           169.254.4.5    TCP       66 445 → 49181 [SYN, ACK] Seq=0 Ack=1 Win=8192 Len=0 MSS=1460 WS=256 SACK_PERM=1
116 42.342523   169.254.4.5           169.254.4.6    TCP       60 49181 → 445 [ACK] Seq=1 Ack=1 Win=65536 Len=0
117 42.342615   169.254.4.5           169.254.4.6    SMB2     162 Negotiate Protocol Request
118 42.342837   169.254.4.6           169.254.4.5    SMB2     228 Negotiate Protocol Response
119 42.343294   169.254.4.5           169.254.4.6    SMB2     220 Session Setup Request, NTLMSSP_NEGOTIATE
120 42.343481   169.254.4.6           169.254.4.5    SMB2     289 Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED, NTLMSSP_CHALLENGE
121 42.343906   169.254.4.5           169.254.4.6    SMB2     571 Session Setup Request, NTLMSSP_AUTH, User: HOST\User
122 42.344260   169.254.4.6           169.254.4.5    SMB2     131 Session Setup Response, Error: STATUS_ACCOUNT_RESTRICTION
123 42.344490   169.254.4.5           169.254.4.6    TCP       60 49181 → 445 [RST, ACK] Seq=792 Ack=487 Win=0 Len=0
124 44.282144   Vmware_18:ad:bd       Broadcast      ARP       60 Who has 169.254.0.10? Tell 169.254.4.5
125 44.297207   fe80::e4f9:6a0e:f92…  ff02::1:2      DHCPv6   157 Solicit XID: 0x7c887b CID: 000100012501326750465db19da1
126 45.000247   Vmware_18:ad:bd       Broadcast      ARP       60 Who has 169.254.0.10? Tell 169.254.4.5
127 46.000585   Vmware_18:ad:bd       Broadcast      ARP       60 Who has 169.254.0.10? Tell 169.254.4.5
```

Figure 17: SMB2/TCP traffic between originally infected host and identified vulnerable host on the network

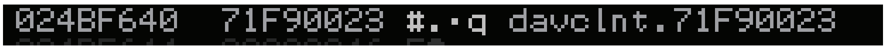


```
3435 1504.171765   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.72? Tell 169.254.4.6
3436 1505.354562   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.117? Tell 169.254.4.5
3437 1505.993614   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.117? Tell 169.254.4.5
3438 1506.356253   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.73? Tell 169.254.4.6
3439 1506.992303   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.117? Tell 169.254.4.5
3440 1507.166888   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.73? Tell 169.254.4.6
3441 1508.164957   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.73? Tell 169.254.4.6
3442 1509.379335   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.118? Tell 169.254.4.5
3443 1509.987268   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.118? Tell 169.254.4.5
3444 1510.380577   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.74? Tell 169.254.4.6
3445 1510.986076   Vmware_18:ad:bd    Broadcast      ARP      60 Who has 169.254.1.118? Tell 169.254.4.5
3446 1511.160703   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.74? Tell 169.254.4.6
3447 1512.158999   Vmware_53:d6:bc    Broadcast      ARP      42 Who has 169.254.0.74? Tell 169.254.4.6
```

Figure 18: Originally Infected host and secondary infect host sweep the network for additional vulnerable hosts

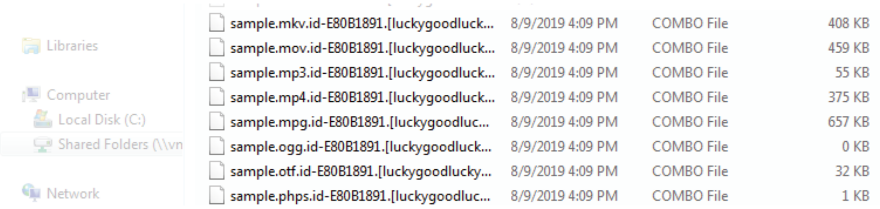
4.3 CrySis/Dharma Functionality

Our work to investigate incoming and outgoing network traffic with Wireshark for CrySiS/Dharma were inconclusive, results of static analysis with *Immunity Debugger* however revealed several references to WebDav in the form of davcint amongst others (Figure 19). This suggests the sample may be intent on spreading itself to servers on both the network and potentially the internet. This may also indicate intent to download additional files from remote locations. Network functionality however was identified, as CrySiS successfully encrypted the contents of a mapped network location, which indicates it can implicate upon additional systems on the network without necessarily infecting other systems.



```
024BF640  71F90023 #.·q davclnt.71F90023
```

Figure 19: References to davcint i.e. WebDav



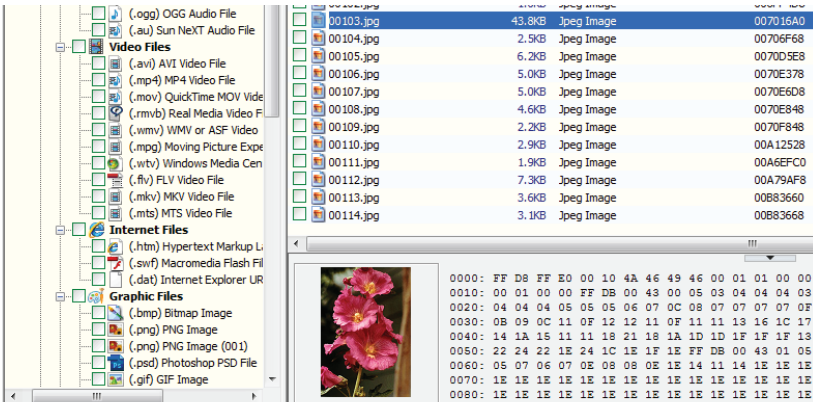
sample.mkv.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	408 KB
sample.mov.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	459 KB
sample.mp3.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	55 KB
sample.mp4.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	375 KB
sample.mpg.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	657 KB
sample.ogg.id-E80B1891.[luckygoodluck...	8/9/2019 4:09 PM	COMBO File	0 KB
sample.otf.id-E80B1891.[luckygoodlucky...	8/9/2019 4:09 PM	COMBO File	32 KB
sample.phps.id-E80B1891.[luckygoodluc...	8/9/2019 4:09 PM	COMBO File	1 KB

Figure 20: Files on the mapped network drive encrypted

5. Conclusions and Future Work

Our work, throughout its analysis of ransomwares such as; WannaCry, Petya and CrySiS led to many key findings and determined that despite the increasing obstinance of ransomware, numerous issues exist within ransomware which drastically reduce the overall effectivity of each variant.

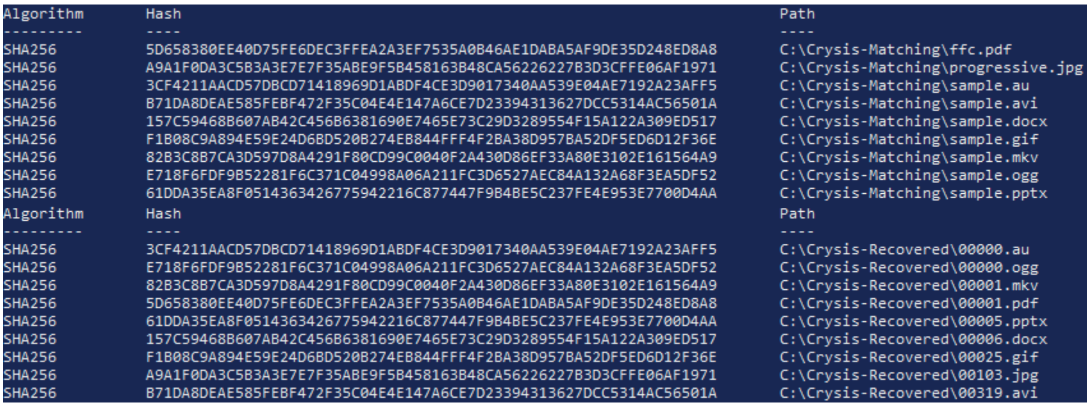
Our analysis of CrySiS/Dharma showed that whilst CrySiS/Dharma appeared to be the least sophisticated superficially, it behaves similarly to WannaCry and will during the encryption process, create encrypted duplicates of files before deleting the originals. Attempting to avoid WannaCry's mistake of leaving data intact prior to deletion, CrySiS/Dharma will nullify the data within files before deleting it, meaning all files will display as 0kb in size. Our findings show this technique effectively thwarts conventional data recovery tools directly, making recovery unfeasible within the context of the virtual environment. Data recovery however proved viable with physical hardware and disk forensic tools such as *DiskGenius*, by removing the physical hard drive from the system and connecting it to an alternative system and utilising disk forensics tools to search the drive sector by sector. This resulted in large volumes of the populated data proving to be intact and recoverable (Figure 21), SHA256 verification showed the files integrity remained unmodified between the recovered files and the originally populated files (Figure 22).



00103.jpg	43.8KB	Jpeg Image	007016A0
00104.jpg	2.5KB	Jpeg Image	00706F68
00105.jpg	6.2KB	Jpeg Image	0070D5E8
00106.jpg	5.0KB	Jpeg Image	0070E378
00107.jpg	5.0KB	Jpeg Image	0070E6D8
00108.jpg	4.6KB	Jpeg Image	0070E848
00109.jpg	2.2KB	Jpeg Image	0070F848
00110.jpg	2.9KB	Jpeg Image	00A12528
00111.jpg	1.9KB	Jpeg Image	00A6EFC0
00112.jpg	7.3KB	Jpeg Image	00A79AF8
00113.jpg	3.6KB	Jpeg Image	00B83660
00114.jpg	3.1KB	Jpeg Image	00B83668

```
0000: FF D8 FF E0 00 10 4A 46 49 46 00 01 01 00 00
0010: 00 01 00 00 FF DB 00 43 00 05 03 04 04 04 03
0020: 04 04 04 05 05 05 06 07 0C 08 07 07 07 07 0F
0030: 0B 09 0C 11 0F 12 12 11 0F 11 11 13 16 1C 17
0040: 14 1A 15 11 11 18 21 18 1A 1D 1D 1F 1F 1F 13
0050: 22 24 22 1E 24 1C 1E 1F 1E FF DB 00 43 01 05
0060: 05 07 06 07 0E 08 08 0E 1E 14 11 14 1E 1E 1E
0070: 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E
0080: 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E 1E
```

Figure 21: One of the populated files shown to be intact during disk forensics



```
Algorithm      Hash                                                                 Path
---------      ----                                                                 ----
SHA256         5D658380EE40D75FE6DEC3FFEA2A3EF7535A0B46AE1DABA5AF9DE35D248ED8A8     C:\Crysis-Matching\ffc.pdf
SHA256         A9A1F0DA3C5B3A3E7E7F35ABE9F5B458163B48CA56226227B3D3CFFE06AF1971     C:\Crysis-Matching\progressive.jpg
SHA256         3CF4211AACD57DBCD71418969D1ABDF4CE3D9017340AA539E04AE7192A23AFF5     C:\Crysis-Matching\sample.au
SHA256         B71DA8DEAE585FEBF472F35C04E4E147A6CE7D23394313627DCC5314AC56501A     C:\Crysis-Matching\sample.avi
SHA256         157C59468B607AB42C456B6381690E7465E73C29D3289554F15A122A309ED517     C:\Crysis-Matching\sample.docx
SHA256         F1B08C9A894E59E24D6BD520B274EB844FFF4F2BA38D957BA52DF5ED6D12F36E     C:\Crysis-Matching\sample.gif
SHA256         82B3C8B7CA3D597D8A4291F80CD99C0040F2A430D86EF33A80E3102E161564A9     C:\Crysis-Matching\sample.mkv
SHA256         E718F6FDF9B52281F6C371C04998A06A211FC3D6527AEC84A132A68F3EA5DF52     C:\Crysis-Matching\sample.ogg
SHA256         61DDA35EA8F0514363426775942216C877447F9B4BE5C237FE4E953E7700D4AA     C:\Crysis-Matching\sample.pptx
Algorithm      Hash                                                                 Path
---------      ----                                                                 ----
SHA256         3CF4211AACD57DBCD71418969D1ABDF4CE3D9017340AA539E04AE7192A23AFF5     C:\Crysis-Recovered\00000.au
SHA256         E718F6FDF9B52281F6C371C04998A06A211FC3D6527AEC84A132A68F3EA5DF52     C:\Crysis-Recovered\00000.ogg
SHA256         82B3C8B7CA3D597D8A4291F80CD99C0040F2A430D86EF33A80E3102E161564A9     C:\Crysis-Recovered\00001.mkv
SHA256         5D658380EE40D75FE6DEC3FFEA2A3EF7535A0B46AE1DABA5AF9DE35D248ED8A8     C:\Crysis-Recovered\00001.pdf
SHA256         61DDA35EA8F0514363426775942216C877447F9B4BE5C237FE4E953E7700D4AA     C:\Crysis-Recovered\00005.pptx
SHA256         157C59468B607AB42C456B6381690E7465E73C29D3289554F15A122A309ED517     C:\Crysis-Recovered\00006.docx
SHA256         F1B08C9A894E59E24D6BD520B274EB844FFF4F2BA38D957BA52DF5ED6D12F36E     C:\Crysis-Recovered\00025.gif
SHA256         A9A1F0DA3C5B3A3E7E7F35ABE9F5B458163B48CA56226227B3D3CFFE06AF1971     C:\Crysis-Recovered\00103.jpg
SHA256         B71DA8DEAE585FEBF472F35C04E4E147A6CE7D23394313627DCC5314AC56501A     C:\Crysis-Recovered\00319.avi
```

Figure 22: SHA256 hashes of the recovered files and matching files before/after CrySiS execution

Analysis showed that whilst Petya appeared to be the most damaging, no evidence of file encryption was found, with Petya employing the alternative technique of firstly corrupting the MBR/BCD in the first instance before proceeding to corrupt the MFT table. Our work indicated that the MBR may be restored immediately prior to the MFT being corrupted i.e. the falsified CHKDSK process, however the timescale is evidently limited. Data recovery remained viable following Petya's full execution, despite the loss of the MFT and MBR. Future work could aim to investigate whether repairs may be performed to restore the MFT, which would permit the restoration of the MBR and allow full restoration of data and system functionality to a state comparable to to Petya's execution.

The first ransomware sample WannaCry, was shown outwardly to cause catastrophic consequences for the system and its data due to the nature of the propagation techniques used. Our study however found it to be the least sophisticated, containing errors which only allowed it to encrypt certain filetypes and files from limited directories. Data also proved to be intact and recoverable following reported file encryption as files were copied into encrypted duplicates and the originals then deleted which permitted full data recovery. Our analysis found WannaCry does not implicate on system functionality and only implicates upon data, thereby allowing retainment of full system functionality following WannaCry's removal.

Our study has shown that whilst ransomware collectively remains a prevalent threat, there are several shortcomings within the studied variants which render each less effective than their potential. Our studies key finding was the discovery that it was possible to recover substantial quantities of data following the encryption process of all three studied ransomware findings, due to the varying faults within the encryption process of each ransomware variant. Notably however, our method of recovering data only proved successful in recovering raw files, meaning identifying information about the files such as; filename, filetype and file/directory structure were lost. Future work should therefore focus on attempting to restore both data accompanied with such identifying information to reduce the most severe risks and devastation caused by such recent strains of ransomware. A full taxonomy of our data recovery attempts is detailed in (Table 2), whilst a summary of our key findings is provided in (Table 3).

Table 2: Taxonomy/Summary of the data recovery attempt.

Sample	Execution Stage	Data Recovery Technique	Result
Petya	Initial execution; following modification of the MBR/BCD and prior to MFT corruption.	It was ascertained data remained intact on the disk following the modification of the MBR/BCD. Data proved recoverable by manually removing the hard disk from the infected system and connecting it to an alternative host to recover data. Our study found the MBR/BCD can be repaired at this stage to restore system functionality and data.	Data and system functionality were restored without modification.
	Complete execution; following the modification of the MBR/BCD and MFT corruption	Our study found, data appeared to remain on the systems hard drive following MFT corruption, albeit with loss of system functionality. Disk forensics were carried out with *DiskGenius*, which confirmed the MFT is overwritten but data remained intact and unmodified. Large quantities of populated data proved to be recoverable, albeit with the loss of original filenames and structure.	Data was recoverable with the loss of the original filenames and directory structure.

Sample	Execution Stage	Data Recovery Technique	Result
WannaCry	Following the execution and encryption process.	This study found that during encryption, data is copied into encrypted duplicates before the originals are deleted. Attempts were made to recover these deleted files with *Restoration* and *Recuva*, which permitted the recovery of large volumes of populated data albeit without original filenames, filetypes and file structure. Filetypes were determined manually using *TrIDNet* which allowed the recovery and restoration of original files intact.	Data was recoverable with the loss of original filenames, directory structure and filetypes.
CrySiS/ Dharma	Following the execution and encryption process.	The study of CrySiS/Dharma revealed that original files data is copied into encrypted duplicates, before the original files data is nullified and the original file is deleted. Manual recovery attempts are attempted with *Restoration* and *Recuva* albeit unsuccessfully. Data was later found to be recoverable through disk forensics tools such as *DiskGenius*, the use of *DiskGenius* revealed original data remained intact without original filenames and file structure.	Data recovery was successful with large quantities of data recoverable without original filenames and directory structure.

Table 3: Taxonomy/Summary of the main findings.

Sample	System Implications	Data Implications
WannaCry	WannaCry on execution will create several files before encrypting files on the system and presenting its user interface. The system will still function, and all programs will still execute, the WannaCry ransomware can be deleted at this stage to regain full system functionality.	Analysis has shown WannaCry encrypts a total of 177 filetypes and excludes those not defined in its list of filetypes and in blacklisted directories. WannaCry is observed to firstly create encrypted duplicates of files before deleting the original files on the drive, data recovery software such as; *Restoration* and *Recuva* can be used at this stage to recover files albeit without their original filenames of filetypes.
Petya	Petya on execution overwrites the MBR/BCD before crashing the system and restarting the system. A falsified CHKDSK process will then be displayed which corrupts the master-file-tree.	Analysis indicated restoration of data and system functionality would prove possible in the initial infection stages by manually repairing the MBR of the system. Should full execution occur and the falsified CHKDSK process complete, data may then be recovered intact and unmodified using disk forensics tools.
CrySiS/Dharma	CrySiS/Dharma will on execution target remote networked systems and encrypt their contents. CrySiS/Dharma will then encrypt files and program data on the local system which renders much of the system inoperable with multiple programs unable to execute.	CrySiS/Dharma upon execution prioritizes encryption of mapped network drives before proceeding to encrypt files in the root directory of the system. Files are copied into encrypted duplicates before the original files data is nullified and the original file is deleted. Data proved recoverable using disk forensics tools such as *DiskGenius*.

References

Abrams, L. (2018). *New Brrr Dharma Ransomware Variant Released.* Retrieved from https://www.bleepingcomputer.com/news/security/new-brrr-dharma-ransomware-variant-released/

Akbanov, M., Logothetis, M., & Vassilakis, V. *WannaCry Ransomware: Analysis of Infection, Persistence, Recovery Prevention and Propagation Mechanisms.* Retrieved from https://www.il-pib.pl/czasopisma/JTIT/2019/1/113.pdf

Arntz, P. (2019). *Threat spotlight: CrySIS, aka Dharma ransomware, causing a crisis for businesses.* Retrieved from https://blog.malwarebytes.com/threat-analysis/2019/05/threat-spotlight-crysis-aka-dharma-ransomware-causing-a-crisis-for-businesses/

Coveware. (2019). *Ransom amounts rise 90% in Q1 as Ryuk increases*. Retrieved from https://www.coveware.com/blog/2019/4/15/ransom-amounts-rise-90-in-q1-as-ryuk-ransomware-increases

Davis, J. (2017). *Petya attacks now appear to be causing permanent damage.* Retrieved from https://www.healthcareitnews.com/news/petya-attacks-now-appear-be-causing-permanent-damage

Greenberg, A. (2018). *The Untold Story of NotPetya, the Most Devastating Cyberattack in History.* Retrieved from https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/

Hung, J.,& Joven, M. (2017). *THREAT RESEARCH Petya's Master Boot Record Infection.* Retrieved from https://www.fortinet.com/blog/threat-research/petya-s-master-boot-record-infection.html

Ivanov, A., Mamedov, O., & Sinitsyn, F. (2017). WannaCry mistakes that can help you restore files after infection. Retrieved from https://securelist.com/wannacry-mistakes-that-can-help-you-restore-files-after-infection/78609/

Křoustek, J. (2017). *Things we have learned about Petna, the Petya-based malware.* Retrieved from https://blog.avast.com/things-we-have-learned-about-petna-the-petya-based-malware

Maclejak, D., & Yang, K, Y. (2018). Dharma Ransomware: What It's Teaching Us. Retrieved from https://www.fortinet.com/blog/threat-research/dharma-ransomware--what-it-s-teaching-us.html

Microsoft. (2017). New ransomware, old techniques: Petya adds worm capabilities. Retrieved from https://www.microsoft.com/security/blog/2017/06/27/new-ransomware-old-techniques-petya-adds-worm-capabilities/

Parmar, M. (2019). *Windows 7 Still Used in Almost 50% of Surveyed Businesses.* Retrieved from https://www.bleepingcomputer.com/news/microsoft/windows-7-still-used-in-almost-50-percent-of-surveyed-businesses/

Sanders, J. (2019). *It's 2019, and one third of businesses still have active Windows XP deployments.* Retrieved from https://www.techrepublic.com/article/its-2019-and-one-third-of-businesses-still-have-active-windows-xp-deployments/

Smart, W. (2018). *Lessons learned review of the WannaCry Ransomware Cyber Attack*. Retrieved from https://www.england.nhs.uk/wp-content/uploads/2018/02/lessons-learned-review-wannacry-ransomware-cyber-attack-cio-review.pdf

Suiche, M. (2017). WannaCry — The largest ransom-ware infection in History More than 70 countries are reported to be infected. Retrieved from https://blog.comae.io/wannacry-the-largest-ransom-wareinfection-in-history-f37da8e30a58

Secureworks. (2017). WCry Ransomware Analysis. Retrieved from https://www.secureworks.com/research/wcry-ransomware-analysis

Vulnerabilities Analysis of Software Solutions for Critical Activity Sectors

Alin Zamfiroiu[1,2], Victor Emmanuell Badea[2] and Paul Pocatilu[1]
[1]Bucharest University of Economic Studies, Romania
[2]National Institute for Research and Development in Informatics – ICI Bucharest, Romania
alin.zamfiroiu@csie.ase.ro,
emmanuell.badea@ici.ro,
ppaul@ase.ro

DOI: 10.34190/EWS.20.048

Abstract: The number of security incidents is constantly increasing and represents a serious threat to all critical infrastructure sectors. Software systems, used in the fields of activity of the sectors and subsectors from the annex of Directive 1148/2016 (energy, transport, banking, financial market infrastructures etc.), are the main targets of attacks and harmful actions deliberately to affect their normal functioning. In the last year, there have been numerous cyber-attacks targeting the technologies and solutions of the essential service operators. We have grouped these incidents for each sector separately.

Sometimes security incidents can be prevented and it is ideal to take measures so that these incidents do not occur. Prevention measures must be taken based on existing vulnerabilities, identified and reported. The Common Vulnerabilities and Exposures (CVE) platform presents the vulnerabilities identified so far for different software solutions. In this paper we analyse the recent incidents that have compromised critical solutions or infrastructures that have been reported and have had a great impact on these sectors. A special attention is given to the energy sector, in order to determine the vulnerabilities of the applications used here. We identified the impact score for each vulnerability found on the CVE platform, thus determining the impact level for all identified vulnerabilities as well as the average impact level for each solution.

Keywords: vulnerabilities, software, critical sectors, critical infrastructure.

1. Introduction

All modern countries have a model of society characterized by a high "quality of life". This expression represents the possibility of access to a set of "basic" services and opportunities that are made available to each citizen in order to best express their skills within the society.

In the European Union, the different member states have historically had very different approaches to regulating the protection of their critical infrastructures, as well as very uneven levels of cyber-defense preparedness. This fragmentation in itself was recognized as a vulnerability (Stubbings et al, 2019), (Srinivas, Das, Kumar, 2019).

The Directive (EU) 1148/2016 aims to improve national capabilities, strengthen cooperation at EU level and promote a culture of risk management and incident reporting among key economic players providing essential services and digital services. This directive starts from the risks involved in security incidents on the economic activities, the potential financial losses that companies, citizens or administrations of the Member States can register, as well as the intentional or unintended disruptions to the electronic systems that support essential services for which can affect several Member States at the same time.

The EU strengthens its cyber security, so that all citizens can enjoy greater security in all areas of our digital lives. Operators of Essential Services (OES) and Digital Service Providers (DSPs) must adopt risk management practices and notify national authorities of significant incidents.

Let's take a simple example: A ransomware can easily infect the computers of an electricity supplier, blocking access to important data, thus shutting down all industrial systems used for electricity distribution.

According to the NIS Directive, as the main electricity supplier, the company would be identified by the national authorities as an Essential Services Operator (OES). This implies that the Essential Services Operator (OES) should have measures in place to prevent risks, ensure the security of system networks, and manage incidents.

If an accident does occur, however, the electricity company must immediately notify the relevant national authority.

The exchange of information on both sides is essential to support and prevent future incidents. Also, the company would receive operational assistance from the National Cyber Incident Response Center (CSIRT), so that it could shorten the time to remedy the incident.

Currently, our societies are really dependent on critical infrastructures, which define "quality of life", for example: energy supply services, banking system, transport system, health system, etc.

The proper functioning and protection of these critical infrastructures is synonymous with the protection of the "quality of life" itself (Freedom from Fear Magazine, 2020).

Critical infrastructures are complex elements that constantly require the adoption of protection measures at both national and international level

Regarding the protection of critical infrastructures, two premises are unanimously accepted:
It is practically impossible to provide 100% protection in case of critical infrastructure;
The absence of a unique solution to solve this problem.

The increasingly aggressive terrorist threat makes the scenarios much more complex and even more dramatic. Industrial cyber-attacks are the most dangerous, especially if they target strategic sectors such as energy. Industrial IT incidents can cause production disruptions, leading to very large, difficult and complex economic losses to be addressed.

Many companies around the world take advantage of the benefits of digitalization in order to expand their activity and optimize their industrial processes. This digital evolution greatly simplifies the processes, on the one hand, while on the other hand it determines a potential risk.

Hackers take advantage of exploiting the vulnerability of systems that are not updated or poorly protected, launching cyber-attacks on the energy sector, in our case, this having a huge impact on production, on the continuity of services, further affecting the image of companies.

Computer vulnerabilities are sought for commercialization on the black market and actively exploited in cyber-attacks, causing significant losses and endangering citizens' data and the functioning of today's digital society systems and services.

The expressions "coordinated disclosure" or "responsible disclosure" of computer vulnerabilities refer to the responsibility towards the owner or producer organizations, but especially to the users of the vulnerable systems that may be affected by the publication of information on the identified vulnerabilities. In this context, it is desirable to strike a balance between wanting to publicly signal a vulnerability and taking measures such as notifying the organization and coordinating with it to remedy the problem before publication and carefully weighing the consequences of publication (Turcu, 2018), (Markopoulou, Papakonstantinou, and Hert, 2019).

An additional source of concern is also the fact that some countries have begun to open the energy market for smaller private operators. This means that almost anyone can use mini-power plants (wind or photovoltaic). These operators often do not have a capable and complete IT staff who can technically support the available facilities, which can lead to more vulnerable installations (Kumar, Pandey & Punia, 2014).

In recent years there have been numerous cyber-attacks against the energy sector. Not all were based on well-trained attackers, some incidents were just collateral damage caused by malware infections or configuration failure (Wueest, 2014). Although widely tested, all operating systems and applications are released with bugs (software errors) that affect their security, performance and stability. Inability/failure to update the software of different applications in the energy sector is a common mistake made by IT professionals.

The paper is structured as follows. The next section presents the cyber security incidents identified for the sectors of activity of the essential service operators. The third section is dealing with vulnerabilities analysis for the energy sector. The paper ends with conclusions and future work.

2. Cyber security incidents

In the last year, there have been numerous cyber-attacks targeting the technologies of the sectors of activity of the essential service operators. These are presented in (Significant Cyber Incidents, 2019). We selected from these, sixteen representative incidents:

- **IN16 – October 2019**. Moroccan authorities arrested terrorists from the Daesh group who were planning an attack on the water supply system of Casablanca;
- **IN15 - August 2019**. It has been found that state-sponsored hackers in China targeted several cancer institutes in the US to obtain information on the latest cancer research;
- **IN14 - August 2019**. A suspected Indian cyber espionage group has launched a phishing campaign targeting Chinese government agencies and state-owned enterprises for information on economic trade, defense issues and foreign relations;
- **IN13 - August 2019**. Networks of several Bahrain government agencies and critical infrastructure providers have been infiltrated by Iran-related hackers;
- **IN12 - August 2019**. It has been found that a previously unidentified Chinese spy group has been working since 2012 to gather data from foreign companies in industries identified as strategic priorities by the Chinese government, including telecommunications, healthcare, semiconductor manufacturing and machine learning. The group was also active in the theft of virtual currencies and in monitoring Hong Kong dissidents;
- **IN11 - June 2019**. suspected Iranian group found telecom services in Iraq, Pakistan and Tajikistan;
- **IN10 - April 2019**. Iranian hackers are alleged to have launched a hacking campaign against banks, local government networks and other UK public agencies;
- **IN09 - April 2019**. Bayer pharmaceutical company announced it has prevented an attack by Chinese hackers targeting sensitive intellectual property;
- **IN08 - March 2019**. Iranian cyber spy group targeted governmental and industrial digital infrastructure in Saudi Arabia and the U.S.;
- **IN07 - March 2019**. Russian hackers targeted a number of European government agencies ahead of the EU elections;
- **IN06 - March 2019**. Indonesia's National Election Commission reported that Chinese and Russian hackers analyzed Indonesia's voter database ahead of the country's presidential and legislative elections;
- **IN05 - March 2019**. US officials report that at least 27 US universities were targeted by Chinese hackers as part of a theft campaign from naval technology research;
- **IN04 - February 2019**. European aerospace company Airbus reveals that it was targeted by Chinese hackers who stole the personal and identifying information of some of its European employees;
- **IN03 - January 2019**. Hackers associated with Russian intelligence services were found to target the Centre for Strategic and International Studies;
- **IN02 - January 2019**. US Department of Justice announced operation to disrupt a North Korean botnet that was used to target media, aerospace, financial and critical infrastructure companies;
- **IN01 - January 2019**. Iran has been shown to be involved in a multi-year global DNS hijacking campaign targeting telecommunications and internet infrastructure providers, as well as government entities in the Middle East, Europe and North America.

Considering the sectors of activity in the annex of the Directive 1148/2016 of the critical infrastructures, we grouped these sixteen incidents for each sector separately, Table 1.

Table 1: Incidents by activity sectors (2019)

Sector	Subsector	Incidents
Energy	Electricity Oil Gas	
Transport	Air transport Rail transport Road transport	**IN02**, **IN04**, **IN05**, **IN12**
Banking		**IN10**
Financial market infrastructures		**IN03**, **IN06**, **IN07**, **IN14**
Health sector		**IN09**, **IN12**, **IN15**
Drinking water supply and distribution		**IN16**

Sector	Subsector	Incidents
Digital Infrastructure	IXPs DNS TLD	**IN01**, **IN08**, **IN11**, **IN12**, **IN13**

Figure 1 shows that in March and August 2019 were identified the highest number of incidents (4 incidents).

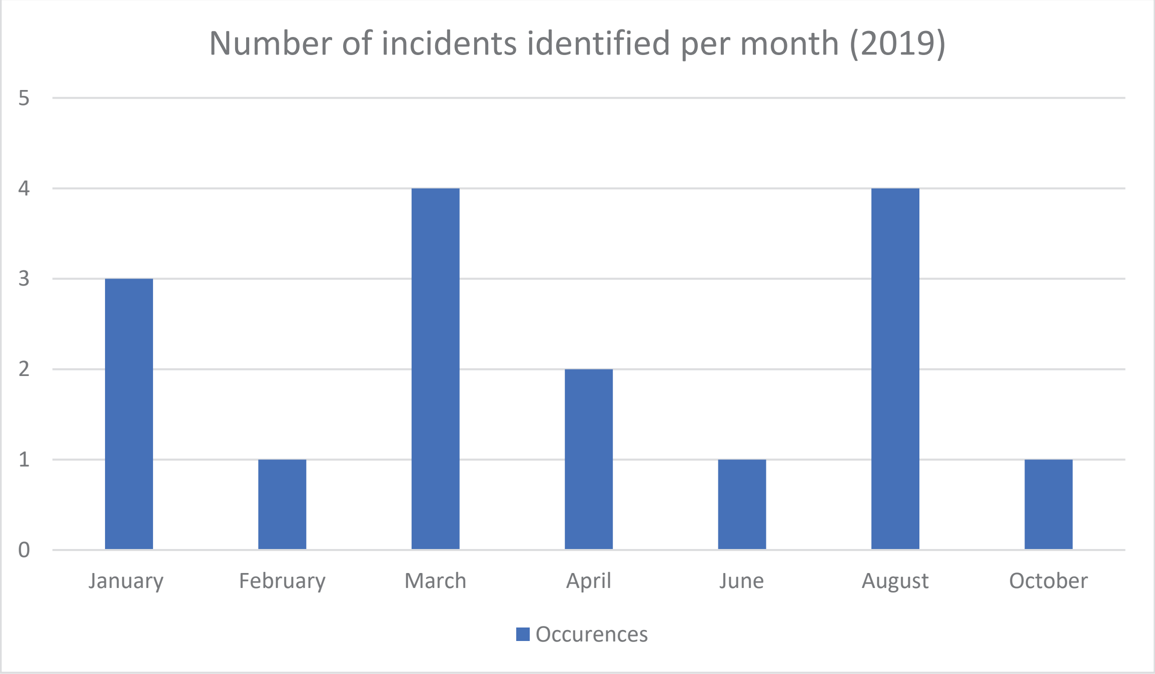


Figure 1: Frequency of incident identification in 2019

The IN12 incident is classified into three sectors (Transport, Healthcare, Digital Infrastructure), being an incident for a longer period (2012-2019). This incident was identified in 2019, but since the analysis began and until the malicious persons were identified, they had time to intervene in several areas.

According to this incident reporting report, no incidents were identified for the energy sector, even there are many vulnerabilities identified for this critical sector, as presented in the next chapter.

3. Vulnerabilities of the critical sectors of activity: Energy

Sometimes security incidents can be prevented and it is ideal to take measures so that these incidents do not occur. Prevention measures must be taken based on existing vulnerabilities, identified and reported (Aleem, 2019).

The Common Vulnerabilities and Exposures platform (CVE, 2019) presents the vulnerabilities identified so far for different software solutions.

We have used this platform and found the vulnerabilities identified so far for the several applications used in the energy sector of activity. These are presented in Table 2.

For each vulnerability identified on the CVE platform, we calculate the impact score of the respective vulnerability on the National Vulnerability Database (NVD, 2019).

On the NVD platform are provided two score calculators for the identified vulnerabilities (CVSS). Both use vulnerability features and allow new scores to be added for new vulnerabilities.

Table 2: Vulnerabilities of software solutions in the energy sector

Software solution	Identified vulnerabilities	Average
Cisco EnergyWise	CVE-2017-3863 CVE-2017-3862 CVE-2017-3861 CVE-2017-3860 CVE-2014-3327	7
Cisco Energy Management Suite	CVE-2018-15445 CVE-2018-15444 CVE-2018-0468	4.66
IBM Insights Foundation for Energy	CVE-2017-1345 CVE-2017-1342 CVE-2017-1311 CVE-2017-1141	4.25
Power Monitoring Expert	CVE-2018-7797	0
Locus Energy LGate	CVE-2016-5782	7
Rockwell Automation FactoryTalk	CVE-2018-18981 CVE-2017-6015 CVE-2017-14022 CVE-2016-4531 CVE-2016-4522 CVE-2014-9209 CVE-2012-4714 CVE-2012-4713 CVE-2012-0222 CVE-2012-0221 CVE-2011-2957 CVE-2010-5305	6.33
EnergyMetrix	CVE-2016-4531 CVE-2016-4522	7
DTE Energy Android Application	CVE-2016-1562 CVE-2014-6002	4.5
ePortal	CVE-2018-8802	6

The impact score is calculated by the formula (NVD, 2019):

$$ISC = 6.42 * (1 - (1 - I_C) * (1 - I_I) * (1 - I_A))$$

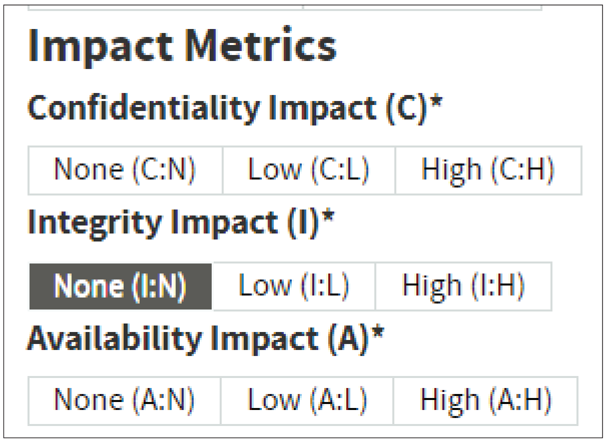
where:

I_C – Confidentiality impact;

I_I – Integrity Impact;

I_A – Availability impact.

These values are specified by the user on the platform, Figure 2.



Impact Metrics

Confidentiality Impact (C)*

None (C:N) Low (C:L) High (C:H)

Integrity Impact (I)*

None (I:N) Low (I:L) High (I:H)

Availability Impact (A)*

None (A:N) Low (A:L) High (A:H)

Figure 2: Impact metrics on the NVD platform

With these scores we can calculate the average of the impact for each solution. These calculations are presented in Table 2, column Average.

For the Power Monitoring Expert solution, the average obtained is 0, because a single vulnerability has been identified for this solution, and for this vulnerability the impact score on the NVD platform has not been calculated so far. Figure 3 presents the average vulnerability score for each analysed solution, as provided by NVD.

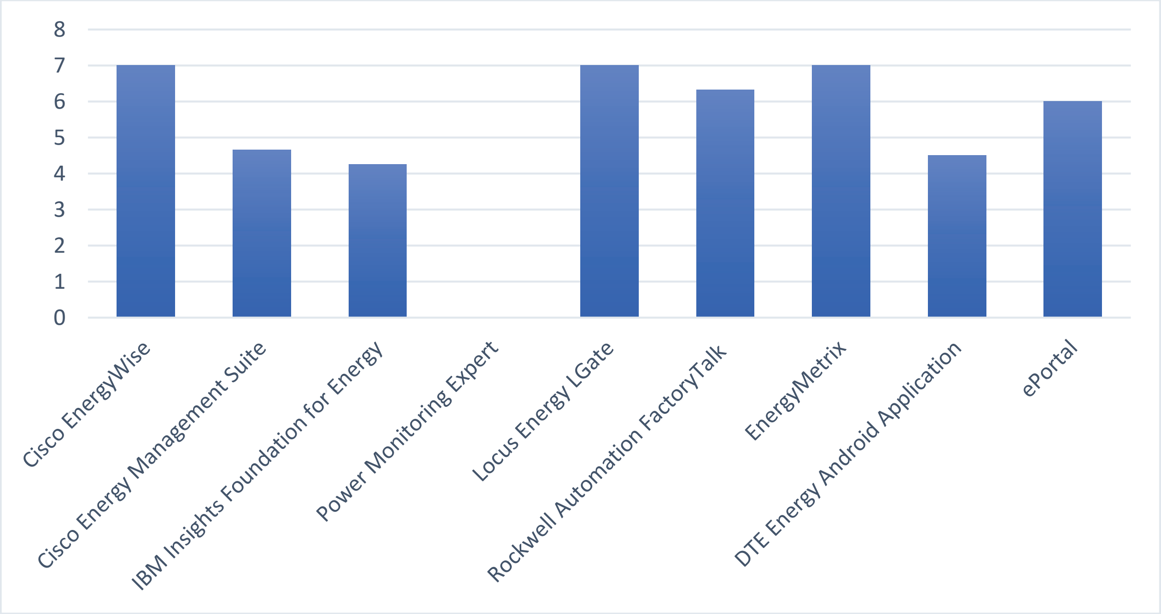


Figure 3: The average vulnerability score for each solution

The existence of these vulnerabilities with a high impact score suggests that in the energy sector, even if there were no incidents in the previous year, there are vulnerabilities that can contribute to future incidents. And CERT teams need to be prepared to handle such incidents at the time of emergence and in the energy sector. The use of vulnerable applications, which have a high score, can contribute to the occurrence of such security incidents in the future.

These values can be obtained by analysing the existing vulnerabilities. In case of new vulnerabilities these values will be different. However, the calculation method remains the same. For future analysis, new solutions may be added or other sectors of activity can be chosen.

4. Conclusions

Both security and privacy are challenges in these critical sectors of activity defined by the EU Directive 2016/1148. There are still open issues where efforts need to be made to improve the security of the dedicated solutions used in these sectors. We saw that even there are no recorded incidents on the energy sector, there are several vulnerabilities that could affect software systems used in this sector.

The analysed applications have various vulnerabilities with a high score. These applications being vulnerable may be the reason for future security incidents in the energy sector.

We plan to extend the current analysis to new software solutions and other sectors of activity and to find patterns (region, period of time, abnormal usage etc.) that could be used to prevent potential attacks.

Acknowledgment

This paper was co-financed from the Human Capital Operational Program 2014-2020, project number POCU / 380/6/13/125245 no. 36482 / 23.05.2019 "Excellence in interdisciplinary PhD and post-PhD research, career alternatives through entrepreneurial initiative (EXCIA)", coordinator The Bucharest University of Economic Studies”.

References

Aleem, A. (2019). Treading water: why organisations are making no progress on cyber security. Network Security, 2019(11), 15-18.

Cisco EnergyWise IOS Configuration Guide, EnergyWise Version 2.7. (2020), Online: https://www.cisco.com/c/en/us/td/docs/switches/lan/energywise/phase2_5/ios/configuration/guide/2_5ewise/mult_ent.html

Cisco Energy Management Suite. (2020), Online: https://www.cisco.com/c/en_ca/products/switches/energy-management-technology/index.html

Common Vulnerabilities and Exposures. (2020), Online: https://cve.mitre.org/

DTE Energy Android Application. (2020), Online: https://detroit.cbslocal.com/2011/11/28/dte-energy-introduces-a-new-android-app-to-get-outage-information/

EnergyMetrix. (2020), Online:https://www.rockwellautomation.com/rockwellsoftware/products/factorytalk-energymetrix.page?

ePortal. (2020), Online: https://eportal.eu/en/

For Power Monitoring Expert. (2020), Online: https://www.schneider-electric.com/en/work/solutions/power-management/demos/power-monitoring-expert.jsp

Freedom From Fear Magazine, Defending Quality of Life Through Critical Infrastructure Protection, Online: http://f3magazine.unicri.it/?p=335

IBM Insights Foundation for Energy product overview, Online: https://www.ibm.com/support/knowledgecenter/en/SSZMQW_2.0.0/com.ibm.ifeop.doc/overview/kc_welcome.html

Kumar, V. A., Pandey, K. K., & Punia, D. K. (2014). Cyber security threats in the power sector: Need for a domain specific regulatory framework in India. Energy policy, 65, 126-133.

Locus Energy LGate, Online: https://www.locusenergy.com/lgate_info

Markopoulou, D., Papakonstantinou, V., & de Hert, P. (2019). The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation. Computer Law & Security Review, 35(6), 105336.

NVD – National Vulnerability Database, Online: https://nvd.nist.gov/vuln-metrics/cvss

Rockwell Automation FactoryTalk, Online:
https://www.rockwellautomation.com/rockwellsoftware/products/factorytalk-view-se.page?

Significant Cyber Incidents, Online: https://www.csis.org/programs/technology-policy-program/significant-cyber-incidents

Srinivas, J., Das, A. K., & Kumar, N. (2019). Government regulations in cyber security: Framework, standards and recommendations. Future Generation Computer Systems, 92, 178-188.

Stubbings T, Watteyne B., Kristmar T., Corbe M., Dimara D., Amokrane A., Cornu T., Sevinga A., Galimberti M., Sanchez Vaquero I. (2019). Complying with the European NIS Directive, April, 2019. Online at: https://assets.kpmg/content/dam/kpmg/be/pdf/2019/05/ADV-brochure-Complying-with-the-EU-NIS-directive-en-LR.pdf

Turcu P. (2018), Directiva NIS în contextul Pieței Unice Digitale CERT-RO, November 2018 Online: http://ier.gov.ro/wp-content/uploads/2018/11/Paul-Turcu_CERT.pdf

Wueest C. (2014), Targeted Attacks Against the Energy Sector, January 2014, Symantec, Online: https://bluekarmasecurity.net/wp-content/uploads/2014/09/Symantec_Targeted-Attacks-Against-the-Energy-Sector_whitepaper.pdf

Small Network Telescope for Advanced Security Monitoring

Linda Zeghache and Hakimi Yacine
Research Centre for Scientific and Technical Information, Algiers, Algeria
Saad Dahleb University, Blida, Algeria
lzeghache@cerist.dz
hakimi.yacine189@gmail.com

DOI: 10.34190/EWS.20.094

Abstract: Network telescope or commonly called darknet is a block of public IP address space monitored to infer malicious activities originated from cyberspace. Connected research works showed the effectiveness of leveraging darknet data to increase awareness of network security events. They consent that the bigger the darknet space the more effective is the inference of malicious activities. However, with the growth of the connected devices to the internet, the IPv4 space is limited. This paper discusses a dataset collected by a network telescope utilising a /27 netblock physically located in Algeria. The purpose of this work is to show that even on small network telescopes, extracting threat intelligence is still possible.

Keywords: Network telescope, darknet, passive monitoring, threat intelligence

1. Introduction

The ease, efficiency and convenience of using the Internet implies in some extent exposing users to cyber threats. This technology has become an inexpensive tool to generate malicious activities such as viruses, worms, denial of service (DoS), scans and Botnets. Although their concept is not new, the techniques used by these threats have evolved in recent years. Recently, emerging cyber attacks have shown expanding attack bandwidth and increased degree of difficulty to mitigate.

To address these concerns there is a pressing need for the development of early warning systems which could provide detailed forensic information on new threats in a timely manner and allow organizations to follow a proactive cybersecurity approach. Monitoring the cyber space is an effective way for generating cyber threat intelligence. It collects threat knowledge before a cyber attacker targets a victim system to help organizations understand and mitigate the risks related to zero-day exploits, Advanced Persistent Threats (APTs), internal and external threat actors (Maglaras et al, 2018).

Network telescopes, also known as Darknets (Cymru et al, 2004), Internet Background Radiation(IBR), sinkholes or blackhole monitors (Baily et al, 2006), are trap based monitoring systems capturing network traffic destined to an unallocated IP address space. These unused IP addresses are not supposed to receive any packets from the Internet so all packets targeting the darknet are either malicious or misconfigured (Ban et al, 2012).

Darknet can provide valuable insights and knowledge about threats to help security policy to be proactive. Evidently many global entities rely on Darknet data to generate cyber threat intelligence. Ban et al (2016) state that Darknet, usually provides a good trade-off between the monitoring cost and global knowledge acquisition. Network telescope is easier to deploy, non-intrusive, contains less traffic compared to the real network and covers larger malicious activity traces. The passive nature of Darknet makes it capable of capturing cyber-attacks without being affected.

Many authors state that a large network telescope is important for monitoring global events so current Darknet designs require large, contiguous blocks of unused IP addresses. However, this is not always feasible for enterprise network operators especially in African countries because of their dependence on IPV4. Small network telescope can be an interesting option since it allows to preserve IP addresses for production purposes and save efforts and troubles for data analysts by providing smaller datasets as stated by (Chindipha, Irwin and Herbert, 2018).

In this study we aim to show the effectiveness of using a small number of contiguous IP addresses in a network telescope for cyber security monitoring. To do so, we deployed a small network telescope that monitors a block of 32 routed IP addresses then we analysed the incoming traffic.

The remainder of this paper is organized as follows. Section 2 provides the related work of this study. Section 3 presents the dataset. In section 4 we analyse the collected data to uncover the global properties of our network telescope. Section 5 describes the threat analysis and the paper concludes in Section 6.

2. Related works

Network telescope was first introduced by (Moore et al, 2004). A while later, many works have been devoted to leverage network telescopes for advanced security monitoring and threats inference.

Irwin (2013) used the network telescope for tracking and monitoring the Conficker malware outbreak. Fachkha et al (2014) analysed the DNS amplification DDoS attacks to understand the nature of this threat and uncover its mechanism. Ban et al (2016) proposed a clustering approach to discover new emerging attacks at their early stage. Bou-Harb et al (2013) used a number of statistical techniques and probabilistic distribution methods to understand and analyse probing activities. Dainotti et al (2011) analysed in detail the characteristics of country-wide Internet outages and censorship. Antonakakis et al (2017) analysed Mirai botnet to demonstrate the damage that an IoT botnet can inflict upon the public Internet.

The aforementioned works used real captured packets provided by darknets operated all over the world. Among the most popular large scale network telescopes operating on /8 netblocks we name Nicter operated by the National Institute of Information and Communications Technology (NICT) in Japan (Masashi et al, 2011). The UCSD (University of California, San Diego) Network managed by the Cooperative Association for Internet Data Analysis (Caida, 2019). The University of Michigan deployed Internet Motion Sensor (IMS), a distributed globally scoped telescope using 28 blocks, ranging in size from /25 to /8 network address blocks (Bailey et al, 2005).

Small scale network telescopes were also deployed by many organizations to monitor global security events. Harder et al (2006) set up a small network telescope with /24 netblock to show that even on such a relatively small telescope, the classification of malware-generated traffic is practical. Irwin (2011) deployed a small telescope (/24) at Rhodes university In South Africa.

Relating to the sizing of network telescopes Wustrow et al (2010) stated that a large address space is needed in order to obtain meaningful data as they observe a wide variety of unique IP sources.

Benson et al (2015) studied the effect of using a smaller telescope by varying the size from /16 to /8. They observed a power law relationship between the number of /24 blocks observed and the number of darknet IP addresses monitored.

Chindipha, Irwin and Herbert (2018) used a sampling approach to measure the influence of the subnets size (/25, /26 and /27) on the data gathered by the original telescope (/24). They analyzed traffic from two different network sensors within the same period. They found that large subnets are still preferable as they can observe more unique IP sources. Their study revealed that the lower subnets received more unique IP hosts than the rest of the subnets. However the influence of the subnets size on the nature of traffic was not discussed.

3. DataSet

The data used in this study is obtained from a small network telescope which consists of a /27 netblock with 32 IPV4 addresses physically located in Algeria. The network monitoring is totally passive as the telescope doesn't interact with the incoming traffic. Only the first packet of the TCP three-way handshake is captured and no data payload can be captured. For UDP and ICMP packets all the information is available from the initial packet.

The network telescope host itself ensured that no return packets could be sent, and is located outside of the organisational firewall.

Knowing that "The lower subnets received more unique IP source than to the rest of the subnets" (Chindipha, Irwin and Herbet, 2018), the telescope is placed in the lower subnet of the network in order to have a better representation of unique IP source.

The data collection started by 19th March 2019. This study began one month later so the dataset analysed in this paper covers the period from 19th March to 21th April 2019.

The total number of packets observed in the telescope was 3611370 packets resulting in an average of 112855 packets per IP and a mean arrival rate of between 1 and 2 packets per second.

The collected traffic contained 232,753 unique source IP addresses. The results are in line with what Chindipha, Irwin and Herbert (2018) observed in the lower subnet /27 of /24 IPV4 address space.

A breakdown of the unique IP source classes is presented in Table 1. It is shown that the major part of the IP sources belongs to the class 'A' and that confirms the results found by (Fachkha et al, 2012).

Table 1: Unique IP Sources Distribution

Class A	Class B	Class C
54.2%	31.1%	14.7%

It was expected that with small telescopes fewer unique IP sources are observed than large telescopes and thus the network visibility is decreased. In the following we are breaking down the collected data to show if small telescopes can assure the same results with the same accuracy as larger telescopes.

4. Data Analysis

Most of the observed hosts sent only a couple of connection initializing packets to the Darknet. Therefore our analysis focuses on the control information in packet headers, including source and destination IP addresses, source and destination ports, and used IP protocols.

Starting the analysis from the transport layer, The incoming traffic is composed of the three common types of protocols seen in all former works : TCP, UDP and ICMP. Table 2 presents the protocol distribution of the collected traffic. It is mainly composed of TCP traffic 90% , 7% is UDP and 1,34% is ICMP. The rest of packets are composed of a lot of other protocols with negligible rates such as SCTP (Stream Control Transmission Protocol) with 1634 packets , IPv6 encapsulation protocol with only 26 packets.

Table 2: Protocols distribution

	TCP	UDP	ICMP
Packets	3283591	278916	48410
Percentage	90.89%	7.72%	1.34%

Figure 1 depicts the arrival rate of the top three protocols packets on a daily basis. It is noteworthy that the majority of traffic is TCP. This latter increases significantly in the period from the 18th to the 20th April. The breakdown of this phenomenon is presented in section 5.

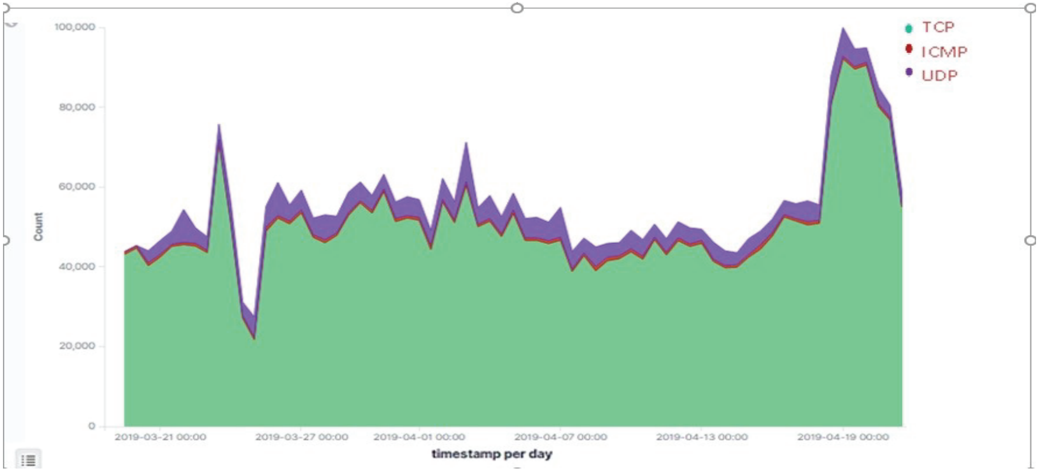


Figure 1: Daily packets count per protocol

4.1 TCP Protocol

As indicated previously, TCP packets still dominate the observed telescope traffic. Fachkha et al (2012) stated that this is related to the fact that the majority of scanning attacks use TCP and the scanning is the preliminary step before launching any attack. The top ten ranked destination TCP ports are presented in figure 2. They accounted for 37,69% of TCP traffic. Except for port 81 and port 5038 the rest of destined ports are allocated by AINA and correspond to well known services. Telnet (23) and ssh (22) are vulnerable to brute force attack. Microsoft Server Message Block (445) and Microsoft terminal server (3389) also have vulnerabilities.

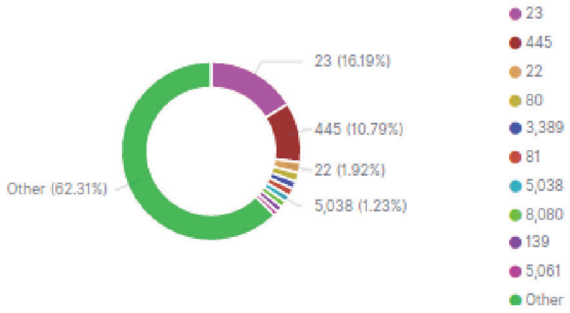


Figure2: Top 10 TCP destination ports

Considering TCP flags, we used Wustro et al (2010) profiling method to infer the nature of Darknet traffic as scanning, backscatters or misconfiguration packets. The comparison between the results provided by our telescope and the previous works with different telescope sizes are presented in table 3. It reveals that the major traffic is scanning activities. Note that the results are a bit different, this is generally related to the placement of the sensor and the collection period (security events that occurred in that period) as stated by Benson et al (2015).

Table3: TCP Flags based traffic distribution

Telescopes	Telescope Size	TCP Flags		
		SYN : Scanning Traffic	SYN+ACK,RST,RST+ACK,ACK: Backscatters	Others : Misconfiguration
Our telescope	/27	99.96%	0.03%	0.01%
(Irwin,2011)	/24	88.63%	11.52%	0.12%
(Fachkha et al, 2012)	/16	68.02%	2.00%	29.98%
(Wustro et al, 2010)	/8	93.9%	5.9%	0.2%

4.2 UDP Protocol

The top ten UDP destination ports are depicted in figure3. They accounted for 64,27% of UDP traffic with the port 5060 accounting for 40.89% of UDP packets. This latter is the Session Initiation Protocol (SIP) used in DoS attacks against Voice Over IP (Fachkha et al ,2012). The ports 123 and 1900 are Network Time Protocol (NTP) and the Simple Service Discovery Protocol (SSDP) both used in DRDos attacks (Ryba et al, 2015). The port 53 is the DNS service known to be widely used in amplification attacks. The ports 137 is the NetBios service which has suffered from vulnerabilities. The port 53413 is a backdoor vulnerability detected in Netis Router (Dulaunoy et al, 2017).

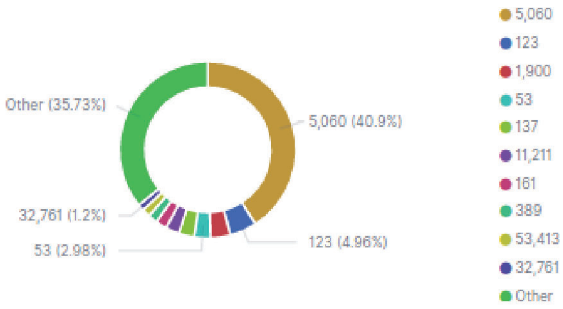


Figure3: Top 10 UDP destination ports

4.3 ICMP Protocol

ICMP packet contains two main elements : the type and the code. The top observed types are namely : type 8 (echo request) accounting for 96.76% of ICMP packets, type 3 (destination unreachable) constitutes 2.58% and type 11 (TTL Exceeded) constitutes 0.47%. The echo request is usually used for reconnaissance activities before launching the scan. Destination unreachable (type 3) is a backscatter traffic, the top observed code for this type is code 3 (port unreachable) could be generated by firewalls.

4.4 Traffic geo-localization

Analysing the incoming traffic geo-localization and how this may change according to the telescope size and placement provides an interesting insight.

The source countries reached 200 where the top countries by packets count are Russia (38.%), USA (24.7%) and China (13.25%). Considering the top geopolitical IP sources we find China, Brazil and Egypt as illustrated in figure4. These results are a bit different from previous works which confirm the statement of (Benson et al, 2015) about the telescope traffic being affected by the sensor placement.

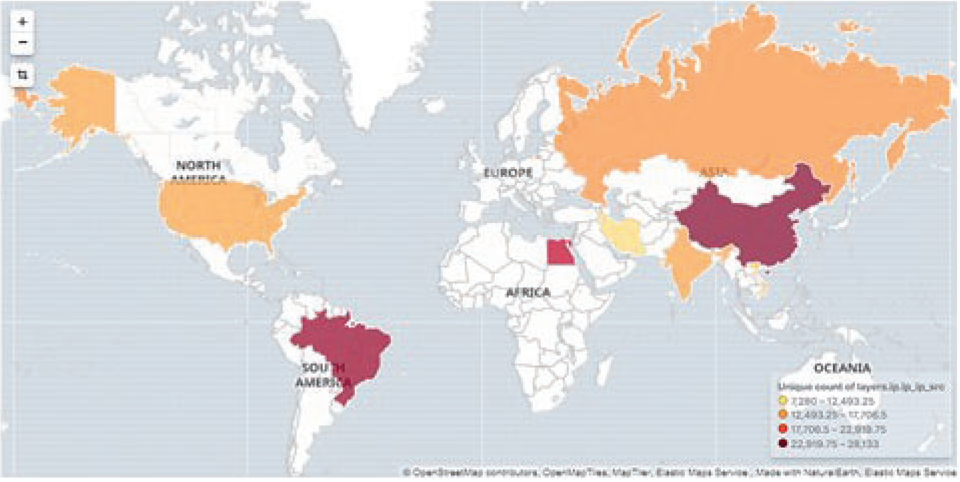

NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
OCEANIA
Unique count of layers.ip.ip_ip_src
7,280 - 12,493.25
12,493.25 - 17,706.5
17,706.5 - 22,919.75
22,919.75 - 28,133
© OpenStreetMap contributors, OpenMapTiles, MapTiler, Elastic Maps Service, Made with NaturalEarth, Elastic Maps Service

Figure 4: IP Source geo-localization

5. Threat analysis

As mentioned in section 4, the telescope observed a significant increase of TCP packets in the period from the 18th to the 20th April. To study this phenomenon, we proceeded a temporal analysis of the top ranked TCP targeted ports, the results are depicted in figure 4. It reveals that the data volume increase is caused by TCP packets targeting 5061, 5160 and 5038 ports.

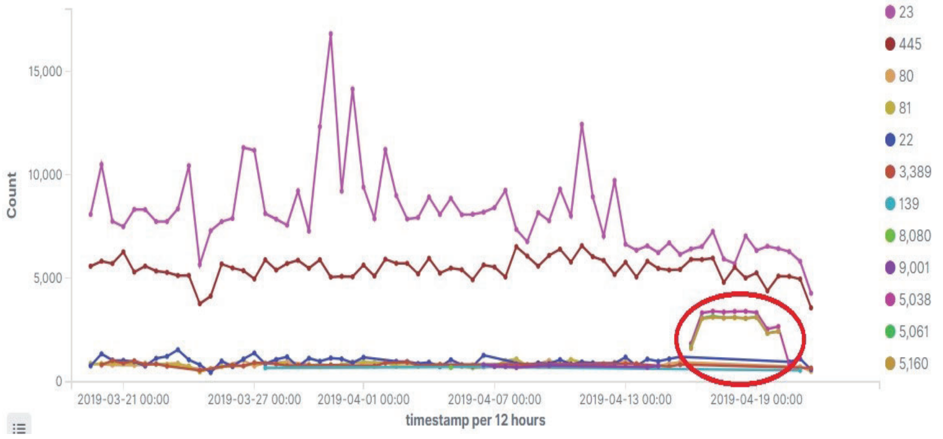


Figure 4: Tcp targeted ports activity trends

Figure 5 presents the daily IP source activity for the ports 5061, 5160 and 5038. It uncovers that all packets targeting those ports are sent from the same IP source located in Russia.

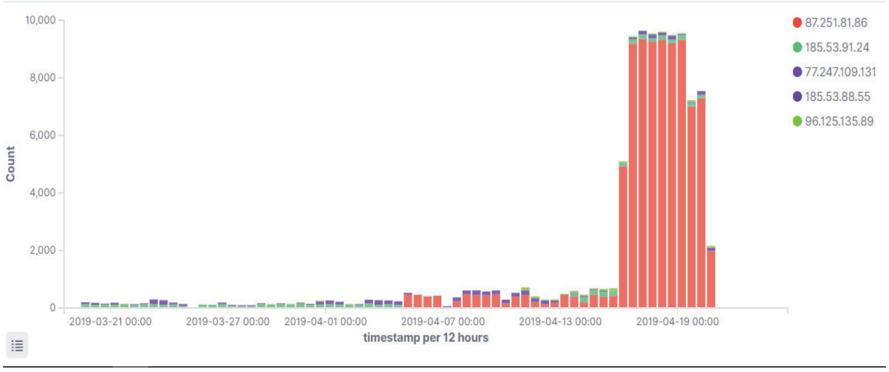


Figure 5: Daily IP source activity (port 5061, 5160 and 5038)

The ports 5060 and 5160 associated with the Session Initiation Protocol (SIP), have vulnerabilities. Unspecified vulnerability in the NAT implementation in Cisco IOS 12.1 through 12.4 and 15.0 through 15.1, and IOS XE 3.1.xSG allowing a denial of service attack by sending crafted SIP packets to TCP port 5060. An issue was discovered on DLink where an authenticated attacker may have full control over the device internals by injecting the shell command into the chkisg.htm page Sip parameter.

Port 5038 is not assigned by AINA, but Asterisk servers use this port for Managing Service. There is no vulnerability in this port but our analysis suspects a security flaw related to SIP protocol.

6. Conclusion and future work

This work was motivated by the fact that African countries are still dependent on IPv4 which makes it difficult to operate large network telescopes within IPv4 address space.

In this paper we investigated the traffic observed by /27 network telescope. A comparison with results from research done on larger sensors has been achieved to show how effective and accurate our results are.

We analysed the characteristics of the incoming traffic based on the transport and network layer protocols and the application layer. The profiling results are in line with previous works that used larger telescopes.

We categorized the collected packets by nature of traffic and geo-located the traffic origins. We concluded that the size of the telescope does not affect the nature of traffic which is generally dependent on the security events occurring during the collection period. Concerning the traffic geo-localization we confirm that it is affected by the placement of the telescope.

Regardless the number of the IP sources observed, this study shows that even a very small telescope can provide insights about security events.

For future work, as we are still collecting data, analysis on long range periods of time using statistical approaches are in progress.

References

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Arbor, A., Bursztein, E., … Zhou, Y. (2017) Understanding the Mirai Botnet . Proceedings 26th USENIX Security Symposium , Vancouver, BC, Canada.

Bailey, M., Cooke, E., Jahanian, F., Nazario, J., Watson, D. (2005) The Internet Motion Sensor: A Distributed Blackhole Monitoring System. In proceedings ofNetwork and Distributed System Security Symposium (NDSS, pages 1–13, San Diego, CA.

Bailey, M., Cooke, E., Jahanian, F., Myrick, A., Sinha, S. (2006) Practical darknet measurement. 2006 IEEE Conference on Information Sciences and Systems, CISS 2006-Proceedings, 1496–1501. https://doi.org/10.1109/CISS.2006.286376.

Ban, T., Zhu, L., Shimamura, J., Pang, S. (2012) Behavior Analysis of Long-term Cyber Attacks in the Darknet, 620–628.

Ban, T., Pang, S., Eto, M., Inoue, D., Nakao, K., Huang, R. (2016) Towards Early Detection of Novel Attack Patterns through the Lens of a Large-Scale Darknet. Proceedings - 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 341–349.

Benson, K., Dainotti, A., Claffy, K., Snoeren, A., Kallitsis, M. (2015). Leveraging Internet Background Radiation for Opportunistic Network Analysis. 423-436. 10.1145/2815675.2815702.

Bou-Harb, E., Debbabi, M., Assi, C. (2013) A Statistical Approach for Fingerprinting Probing Activities. Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES. 21-30. 10.1109/ARES.2013.9.

Caida, The Cooperative Association for Internet Data Analysis. The UCSD Network Telescope, https://www.caida.org/projects/network_telescope/. Last accessed : June 2019.

Chindipha, S., Irwin, B., Herbert, A. (2018) Effectiveness of Sampling a Small Sized Network Telescope in Internet Background Radiation Data Collection. Southern Africa Telecommunication Networks and Applications Conference (SATNAC).

Dainotti, A., Squarcella, C., Aben, E., Claffy, K., Chiesa, M. (2011) Analysis of Country-wide Internet outages Caused by Censorship. IMC'11, pp. 1–18.

Dulaunoy, A., Wagener, G., Mokaddem, S., Wagner, C. (2017) An Extended Analysis of an IOT Malware From a Blackhole Network Paper type Keywords.

Fachkha, C., Bou-Harb, E., Boukhtouta, A., Dinh, S., Iqbal, F., Debbabi, M. (2012). Investigating the dark cyberspace: Profiling, threat-based analysis and correlation. 7th International Conference on Risks and Security of Internet and Systems, CRiSIS 2012. 1-8.10.1109/CRISIS.2012.6378947.

Fachkha, C., Bou-Harb, E., Debbabi, M. (2014). Fingerprinting internet DNS amplification DDoS Activities. 6th International Conference on New Technologies, Mobility and Security - Proceedings of NTMS 2014 Conference and Workshops, 1–5. https://doi.org/10.1109/NTMS.2014.6814019.

Harder, U., Johnson, M., Bradley, J., Knottenbelt, W. (2006) Observing Internet Worm and Virus Attacks with a Small Network Telescope. Electronic Notes in Theoretical Computer Science. 151. 47-59. 10.1016/j.entcs.2006.03.011.

Irwin, B. (2011) A framework for the application of network telescope sensors in a global IP network. PhD thesis, Rhodes University, Grahamstown, South Africa.

Irwin, B. (2013) A Source Analysis of the Conficker Outbreak from a Network Telescope. SAIEE Africa Research Journal. 104. 38. 10.23919/SAIEE.2013.8531865.

Maglaras, L., Ferrag, M., A., Derhab, A., Mukherjee, M., Janicke, H., Rallis, S. (2018) Threats, Countermeasures and Attribution of Cyber Attacks on Critical Infrastructures. Security and Safety. 5. 1-9. 10.4108/eai.15-10-2018.155856.

Masashi, E., Daisuke, I., Jungsuk, S., Junji, N., Kazuhiro, O., Koji, N. (2011) Nicter: A large-scale network incident analysis system: Case studies for understanding threat landscape. 10.1145/1978672.1978677.

Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., and Weaver, N. (2003) Inside the Slammer worm. IEEE Security & Privacy, 1(4):33–39.

Moore, D., Shannon, C., Voelker, G., Savage, S. (2004) Network Telescopes. CAIDA Technical Report.

Reynolds, J.(1992) ASSIGNED NUMBERS. RFC 1340. URL https://tools.ietf.org/html/rfc1340.

Ryba, F., Orlinski, M., Wählisch, M., Rossow, C. Schmidt, T. (2015) Amplification and DRDoS Attack Defense - A Survey and New Perspectives.

Team Cymru. The darknet project. (2004) http://www.cymru.com/Darknet/index.html.

Wustrow, E., Karir, M., Bailey, M., Ja-hanian, F., Huston, G. (2010) Internet background radiation revisited. Proceedings of the 10th annual conference on Internet measurement, pages 62–74. ACM.

A Viewpoint on Management Practices for Cybersecurity in Industry 4.0 Environment

Najam Ul Zia and Victor Kwarteng Owusu
Tomas Bata University in Zlin, Czech Republic
zia@utb.cz
owusu@utb.cz

DOI: 10.34190/EWS.20.031

Abstract: Industry 4.0 is characterized by the smart business developments, application of cyber-physical systems, internet of things (IOTs), excessive use of big data, etc. In other words, it is about computerization and digitization of processes through the above-mentioned technologies. Industry 4.0 offers more effective and efficient resources to regulate the production system as compared to the out-dated centralized system. Today's economy is a data-driven economy and data is the key input to industry 4.0 operations. In the industry 4.0 era companies are bombarded with huge volume and variety of structured and unstructured data with vast velocity. Data has become one of the most significant strategic assets for data-driven companies and subsequent resource-based view (RBV) for sustainable competitive advantage companies require to maintain the rarity, inimitability, value, and organization of data as a strategic resource. This condition makes it very crucial to the data protection, steadfast technology, and secure access has become the main part of selling strategies of companies sell; trust is becoming a very serious factor of commercial dealings. It is an old thought that cybersecurity manages only cyber risks but it also manages the trust. The firms are exposed to cyber risks beyond their direct control and this has resulted from complex digital value chains. There is a possibility of hefty damage due to cyber-attacks in terms of continuous business operations, reputational harm and stealing of confidential information. This paper aims to present a viewpoint on the best appropriate cybersecurity practices to facilitate the businesses to match the stride of industry 4.0. The recent studies discuss only the technological aspect of cybersecurity. This paper is one of the early attempts to draw attention to the important role of cybersecurity management practices and identifying the most relevant challenges of managing cybersecurity in the context of industry 4.0.

Keywords: Cyber-security practices, Industry 4.0

1. Introduction

The industry is a very important pillar of any growing economy. Many industries faced different types of changes in technology and innovations (Shamim, Cang, Yu, & Li, 2017; Shamim, Zeng, Shariq, & Khan, 2019), and these are paradigm moves which are called "Industrial revolutions". As shown in figure 1, the first industrial revolution is characterized as "mechanization" and it came with using of steam power, the second industrial revolution is referred to high usage of electrical energy and the third one is electronics and automation computers. The fourth industrial revolution is a key requirement for today's growing economies and it comes with the internet of things, cyber-physical systems , and networks (Lasi, Fettke, Kemper, Feld, & Hoffmann, 2014). This is triggered by social, technological, economic and political changes. This 4th industrial revolution is also called as Ind 4.0 (Burmeister, Lüttgens, & Piller, 2016), which represents as digital transformation in current business and existing processes, and changing the manual working methodologies with digital computer structures (Iansiti & Lakhani, 2014).

Industry 4.0 has become a global term and its foundation is based on data storage in machines and companies which is utilized to enhance the business processes (Mehrfeld, 2019). In this way, both machines and IT systems are perfectly connected. This leads to better capacity usage and improves planned strategies. Previously machines were not connected with the outside world and these were isolated but now the risks from these developments need to handle, particularly concerning Cyber-security (Mehrfeld, 2019).

Currently, in this interconnected world, Cyber-security is one of the most challenging thing for everyday life, especially for companies having critical infrastructure (Agafonov & Damkus, 2017). Along with the technical aspects of cyber-security, there are concerns with the management practices as well. Cyber-security is defined as a set of measures that are required to be taken to detect, prevent and analyse and answer to all cyber incidents (Agafonov & Damkus, 2017). In companies, where structures are closely linked with each other, the cyber-attacks can spread widely and destroy other linked systems as well (Limba, Plėta, Agafonov, & Damkus, 2019). This must be noted that only technical solutions are not enough to solve the cyber-security issues and model on cyber-security management should be developed and improved along with speedily growing technology and legal features (Limba et al., 2019).

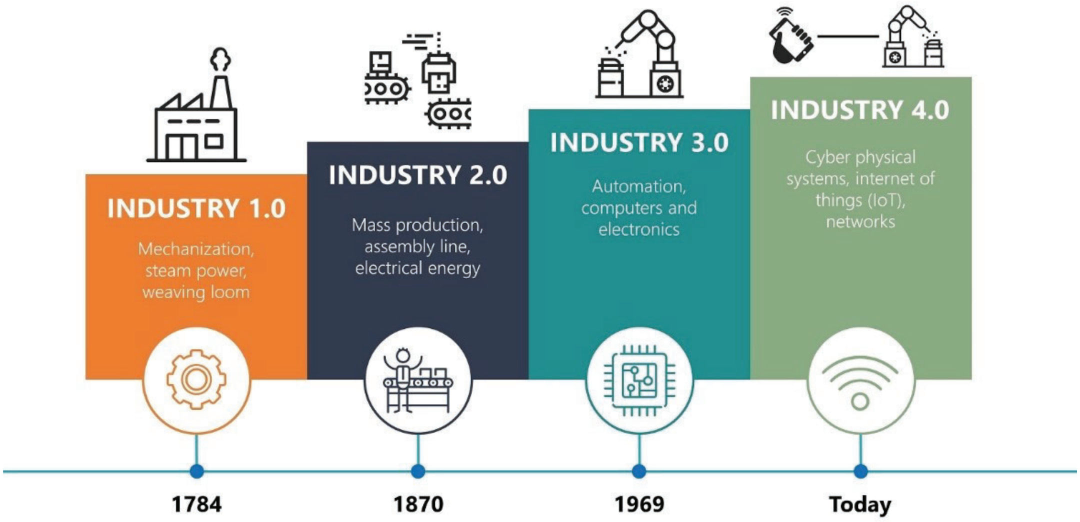

INDUSTRY 1.0
Mechanization, steam power, weaving loom
INDUSTRY 2.0
Mass production, assembly line, electrical energy
INDUSTRY 3.0
Automation, computers and electronics
INDUSTRY 4.0
Cyber physical systems, internet of things (IoT), networks
1784
1870
1969
Today

Figure 1

The cyber-security effectiveness depends on the application of different procedures that must be applied in a balanced portion (Baskerville, Spagnoletti, & Kim, 2014; Elkhannoubi & Belaissaoui, 2016). Furthermore, managers are the main influencers in Cyber-security implications in any company (Fenz, Heurix, Neubauer, & Pechstein, 2014; Soomro, Shah, & Ahmed, 2016). Similarly. Questions like how to get decision making for security, integrate security and organising leadership is being aggressively examined (Fenz et al., 2014; Zammani & Razali, 2016). The current overview of organizational practices shows that companies are more concerned about cyber-security as compare to previous practices, however, the competencies have been worst (Lee & Chang, 2014). Many reports show that companies are not active in cyber-security threats management and very less number of companies obey the standard security guidelines (Gafsi, Ajili, & Hajjaji, 2020)

Due to low funds and budgets, companies do not pay more attention towards cyber-security that's why this causes the main hurdle in managing their cyber-security aspects (Herath & Herath, 2014; Stewart, 2012), and this is a believable argument for underdeveloped practices, but performance is not solely dependent on budgets. Experts believe that overall cyber-security practices are decreasing over time (Prislan, Lobnikar, & Bernik, 2017), while the fact is also recognised that the increasing cyber threats environment is difficult to trail. Resultantly, many companies or public services face data loss or high cyber threat exposures on a daily basis. Researches also explain that overall cyber-threats are basic and well-known patterned in nature (Such, Vidler, Seabrook, & Rashid, 2015). In spite of having vast knowledge of cyber security, we still unable to control the basic cyber threats. For example, the chance of a yearly cyber incidence in any business ranges between 70% to 90% (Klahr et al., 2017). May companies struggle to recover the losses which take many months in their rebound and their share also drops (Bromiley, 2016). Even the latest reports from CIOs (Chief Information Security Officers) demonstrate that security practitioners are not confident in their security systems and they believe that current security tools fail to secure their businesses (Prislan et al., 2017). Several reports also show that many small companies do not even keep the basic security systems in their infrastructure like vulnerability assessment, monitoring processes and tools, threat assessment (Rohn, Sabari, & Leshem, 2016) . These companies believe that they are not open to cyber-crimes and threats (Emm, 2013).

Companies do not have a skilled workforce that can perform proactively on cyber-security issues and these organizations keep on struggling with such lack in cyber-security operations and skills (Prislan et al., 2017). Another reason for cyber-security incompetency is that there are many untailored instructions and guidelines which are common to the needs of management (Siponen & Willison, 2009)

1.1 Cyber-security Slipups

There are common mistakes that companies in industry 4.0 environment can make when analysing the cyber-security of their assets:

- Self-suppose that every infrastructure can be secure from any cyber-attack. Each organization that operates with Industry 4.0 technology must understand that complete security is just a daydream rather they should detect the most vulnerable areas and all the activities they must perform to clear the path and restore the normal activity of every infrastructure (WISAC, 2015). The most important aspect of the Industry 4.0 environment is to detect and respond to critical situations regarding cyber-security (Techrepublic, 2004).

- Untrue thoughts that hiring the best professionals can save the organization from the cyber-attacks. Every organization needs to understand that they can achieve the desired results in an organizational approach and not in a departmental approach (WISAC, 2015). Cyber-security should be an individual objective because human influence is the most susceptible part of cyber-security(Wang & Ho, 2017).
- Another false thinking that all the security tools and technologies that are used, ensure to stop vulnerabilities. Tools and technologies which are used in cyber-security do not grant a complete cyber-security (Wei, Lu, Jafari, Skare, & Rohde, 2010). Only these tools and products don't make the IT department but the responsibility comes to everyone to stand for cyber-security because human aspect is the feeblest about cyber-security (WISAC, 2015).
- False thinking that cyber-security is only efficient monitoring. Monitoring itself is not a narrow term that links organizational internal structure but it is a broader term that is responsible for outer environmental conditions and cybercrime trend tracking. Monitoring is of no use if no one can learn from it (WISAC, 2015). A mature level of monitoring output is to share information about any vulnerability on time because sharing such information will tell the actual picture security situation in the organization (Sun, Hahn, & Liu, 2018).
- False approach that all the security measures which are taken by the organization for cyber-attacks protection are of superior level. Organization must prioritize in cyber-security policies to invest in critical infrastructures and resources so these can detect any vulnerable attempt (WISAC, 2015)

2. Management Practices for Cyber-security in Industry 4.0

The role of senior management in leading the cyber-security posture of any organization is multifaceted. This is important that senior managers should know the relationship among organizational environment, critical infrastructure, and technological base because they organise different processes like risk assessment, asset identification, and establishing proper control environment.

Senior management's role in leading the information security posture of the organization is complex. It requires that senior managers understand the relationships among critical infrastructures, the organizational environment, and the technological base as they coordinate the process of asset identification risk assessment, overseeing the establishment of a proper control environment, and attaining equilibrium between the costs and benefits of control. Industry 4.0 performance is dependent on the revolution capabilities of the management (Lasi et al., 2014) and which may be related with cyber-physical systems, cyber security, product reengineering or some issues regarding supply chain (Shamim, Cang, Yu, & Li, 2016a)

If organizations need to secure from cyber-attacks, they require intelligent and skilled employees and a climate full of innovation and learning, which needs appropriate management practices. This study proposes the suitable management practices to ensure cyber-secure behaviour in Industry 4.0 environment. Management practices should promote a cyber-secure behaviour at individual level. These practices are described below.

2.1 Organizational Culture:

In all organizations, the members follow unwritten norms and values and such culture is a powerful dimension to compromise cyber-security in any organization(Dutta, 2002). For example, an organization that operates with high customer service activities may expose to cyber threats due to the sharing of secret information on call to customers without sufficient verification. The main role of senior managers here is to provide an organizational culture that supports cyber-security standard operating procedures (Dutta, 2002). Cyber-security is not only a technical issue, and organizations also need to depend on employees' behaviours for better protection from cyber-attacks (Huang & Pearlson, 2019). Ultimately these behaviours may reduce or increase the vulnerability to cyber threats. In an organizational culture, the role of employees is to recognise their individual responsibility for developing a cyber-secured behaviour and research shows that it pays off ultimately (Huang & Pearlson, 2019). Managers who suggest higher investments in cyber-security technical operations often resist in investing in organization mechanisms. All the investments in technology can not only secure the organizations from cyber-threats until the behaviours of employees is compatible with cyber-security practices. The managers should develop a cyber-security organizational culture and evaluate its outcomes.

2.2 Cyber-security Leadership:

This has been observed that management sometimes decides not to put extra efforts into cyber-security. In additional explanation, the person responsible to conduct security practices doesn't have relevant skills, abilities, and even intentions. Other decisions regarding investments are taken by company owners, here again, this is the management's role to convince them for investing in cyber-security as well. All these things refer to responsible leadership control (Prislan et al., 2017).

Leadership is an ability to convince, motivate, direct and influence the activities to attain desired organizational goals. Many leaders adopt an appropriate leadership style to achieve desired tasks, this is also mentioned in path-goal theory of leadership (House, 1971). For example, Apple Inc is one of the leading worldwide companies, and its CEO Steve Jobs extracts the output from his employees due to leadership style rather than to rely on only technical skills (Isaacson, 2012; Shah & Mulla, 2013). The same is the case with Microsoft Corporation where the leadership style of Bill Gates takes the huge output from employees (Shah & Mulla, 2013). So, there must be a specialised leadership style for Industry 4.0 with emphasis on cyber security.

The leadership in cyber-security culture denotes to the employment of a person or group with official charge for developing a cybersecurity culture (Huang & Pearlson, 2019). The responsibility of this leader is to build cyber-secure culture and has the authority and direct power to influence the cultivation process of cybersecurity culture. This is suggested to hire a Chief Information Security Officer (CISO), who acts as a leader to ensure all the actions connected to cyber-security effectiveness and who has a big schedule casing all features of cybersecurity culture (Huang & Pearlson, 2019). The organizations can only build a cyber-security culture by appointing an individual with a specific responsibility to look and monitor all the activities. These activities can go haphazardly executed and occasionally bounced without a cyber-security leadership.

2.3 Human Resource Practices

There can be many reasons for skewed management practices, but unsystematically employment in information technology is one of the possible reasons because it does not consider the surrounding vulnerabilities and security threats (Prislan et al., 2017). The absence of pre-working, obstinacy, and insignificance regarding such highly intensive cyber threats development are reported as a major cyber-security problem for any industry (Irwin, 2018). HR practices are supposed to be one of the key sources which are adopted by organizations to sharpen the skills, behaviours, capabilities, and attitudes of employees to achieve desired organizational goals (Collins & Clark, 2003). Managers enhance knowledge management capability and innovativeness among different employees by designing suitable HR practices (Chen & Huang, 2009). The common HR practices which need to design for innovation are staffing, training, performance appraisal and job design (Ma Prieto & Pilar Perez-Santana, 2014). Managers should design HR practices with the intention of a better cyber-security environment in Industry 4.0 organizations.

2.3.1 Cyber-security Training

Cybersecurity training denotes to the exercises and courses designed for the workers to build cyber-security knowledge and skills. Industry 4.0 organizations should plan training programs to enhance the capability of employees for better learning and innovation (Shamim, Cang, Yu, & Li, 2016b). There must be a specific training program for all workers to understand the basic skills need for better Cyber-security environment. This is not necessary that the training programs should directly be linked with their jobs but to enable them to multitask, multifunctional and increase the diversity of skills (Chang, Gong, & Shum, 2011). Training should also an emphasis on teamwork skills and team building, mentoring can be the monotonous activity for employees especially the newly hired ones (28). Moreover, there should be training to improve the problem resolving skills of employees (Chang et al., 2011). These trainings mature the employees in cyber-security awareness, train the internal workers to play a role in information security and instructs the users on the position of secured information(Huang & Pearlson, 2019). The newly hired employees should also be trained in cybersecurity as a part of their orientation process. This is also suggested to upgrade and update the cyber security knowledge in existing employees by offering online training programs annually. Not only this annual course update is sufficient for cyber-secure behaviour, but the trainings should be done on regular basis to sustain in a longer run. The practices of just-in time learnings and pop-up windows can also become efficient in training the individual employees on cyber-security basic practices(Huang & Pearlson, 2019).

2.3.2 Rewards & Punishments

The compensation arrangement in industry 4.0 should reflect the contribution of an employee in cyber-security management. Employees should get compensation based on individual, organizational and group performance (Ma Prieto & Pilar Perez-Santana, 2014). There should be a link between (cybersecurity) performance and rewards (Collins & Clark, 2003). Such a system of compensation always pays off in the way of innovation and learning in the industry 4.0 organization (Ma Prieto & Pilar Perez-Santana, 2014). Employees with a good and cyber-secure behaviour should be rewarded and punished accordingly. The rational choice theory, protection motivation theory and deterrence theory, all these theories explain that rewards and punishments have direct impact on individual decisions. So there should be the rewards including announcements and appreciation certificates for employees having model cyber security behaviours (Huang & Pearlson, 2019), whereas punishments like, remedial trainings and admonishments for offending employees.

2.3.3 Performance Appraisal

The performance appraisal refers to presence of measures of cyber-security actions and behaviours in the formal evaluation process of an employee (Huang & Pearlson, 2019). The assessment of cyber-security performance has no practical existence in almost many organisations (KESSEL & Allan, 2014). Measurement of performance and quality are the key factors in building efficient cyber-security tools in any organization (Prislan et al., 2017). This is explained in expectancy theory that managers use the evaluation process to clarify that what behaviours are acceptable and what behaviours are not required (Huang & Pearlson, 2019). For example, this is unacceptable for employees to give system passwords to suppliers or vendors without prior approval from top management. A better performance appraisal system for industry 4.0 is to evaluate the employees based on development, behaviour base approach and result base approach because such approaches enhance the learning and innovation of the organization (Chen & Huang, 2009). The cyber-security practices must be part of employee job objectives and should be evaluated in performance appraisal. Among many performance appraisal approaches, one specific approach of management by objective (MBO) is popular. MBO is described as, a performance evaluation method in which a mutual objective setting is added and evaluation is done based on these objective attainments (DeCenzo, Robbins, & Verhulst, 2016). An ideal performance appraisal should comprise the formation of performance standards, mentioning the expectations, measuring the real performance, discussing the results of appraisal with the employee and taking the correct actions where needed (DeCenzo et al., 2016).

3. Discussion and conclusion

This paper presents a viewpoint on best management practices for the better cyber-security environment in industry 4.0 organizations. The business strategies of industry 4.0 are discussed then the major cyber-security mistakes are discussed. The most important aspect of the industry 4.0 environment is to detect and respond to critical situations regarding Cyber-security (Techrepublic, 2004). Based on arguments that industry 4.0 success relies on innovation capabilities of enterprises (Lasi et al., 2014), this study offers a viewpoint on appropriate cyber-security practices according to the management perspectives which includes, leadership, organizational culture, and human resource practices. There is also limitations of this study as this is a viewpoint paper and it further needs empirical investigation and validation. Additionally, this paper also describes future directions for researchers in line with industry 4.0. the frame work in Figure 2 represents a direction to quantify the cyber-security effectiveness in Industry 4.0 environment by investigating the effect of cyber-security leadership, HR practices and organizational culture on individual cyber-secured behaviours of workers. A quantitative research approach is required to validate these arguments in the context of cyber-security and industry 4.0. The component of analysis should be the high-tech originalities with smart manufacturing and implementation of cyber-physical systems.

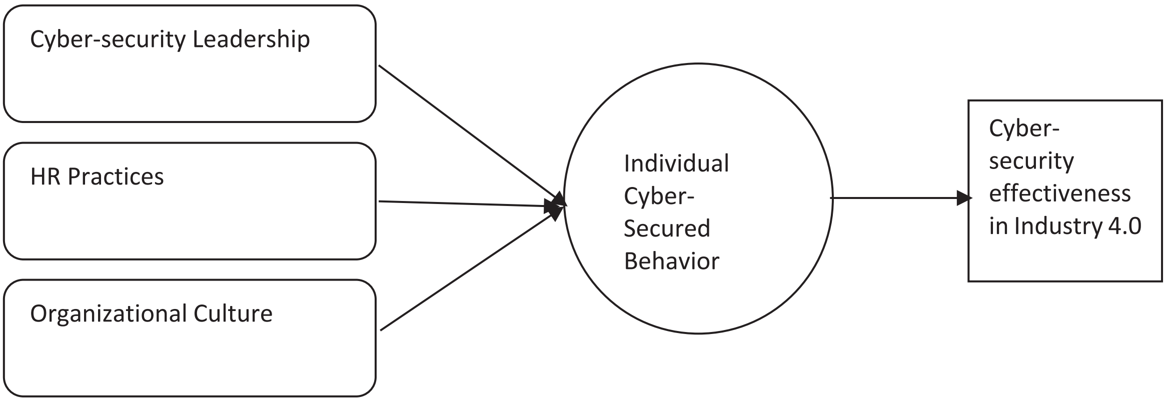

Cyber-security Leadership
HR Practices
Organizational Culture
Individual Cyber-Secured Behavior
Cyber-security effectiveness in Industry 4.0

Figure 2: Framework for future research

Cyber-security development is reliant on employee's capability which is enabled by best management practices. In this way, appropriate management practices can enhance the cyber-security skills of employees in any industry 4.0 organization. Therefore, for better cyber-security output, technological and high scientific techniques are also important aside from these management practices. The main objective of management practices is to reduce the individual behaviours that cause cybersecurity vulnerability and surge behaviours that protect the organization. Additionally, apart from keeping the technologies up to date and installing the latest cyber security software, the managers also make the decisions that affect behaviours, believes, values and attitudes of employees around the cybersecurity.

In conclusion this study integrates existing theories and logical believes and proposes that, for organizational level cyber security effectiveness it is important to encourage cyber-secured behaviour at employee individual level. We further argue that employee individual level cyber-secured behaviour can be promoted by applying appropriate management practices such as cyber-security leadership, HR practices and organizational culture.

Acknowledgements

Author acknowledge the financial support provided by Internal Grant Agency (IGA/FaME/2020/003) of FaME through Tomas Bata University in Zlín, Czech Republic

References

Agafonov, K., & Damkus, M. (2017). *Cyber Security Management Model for Crtitical Infrastrucure*. *4*(4), 559–573.

Baskerville, R., Spagnoletti, P., & Kim, J. (2014). Incident-centered information security: Managing a strategic balance between prevention and response. *Information & Management*, *51*(1), 138–151.

Bromiley, M. (2016). Incident Response Capabilities in 2016: The 2016 SANS Incident Response Survey. *SANS Institute, June*.

Burmeister, C., Lüttgens, D., & Piller, F. T. (2016). Business model innovation for Industrie 4.0: why the "Industrial Internet" mandates a new perspective on innovation. *Die Unternehmung*, *70*(2), 124–152.

Chang, S., Gong, Y., & Shum, C. (2011). Promoting innovation in hospitality companies through human resource management practices. *International Journal of Hospitality Management*, *30*(4), 812–818.

Chen, C.-J., & Huang, J.-W. (2009). Strategic human resource practices and innovation performance—The mediating role of knowledge management capacity. *Journal of Business Research*, *62*(1), 104–114.

Collins, C. J., & Clark, K. D. (2003). Strategic human resource practices, top management team social networks, and firm performance: The role of human resource practices in creating organizational competitive advantage. *Academy of Management Journal*, *46*(6), 740–751.

DeCenzo, D. A., Robbins, S. P., & Verhulst, S. L. (2016). *Fundamentals of Human Resource Management, Binder Ready Version*. John Wiley & Sons.

Dutta. (2002). *Management role in IS failure.pdf* (pp. 67–87). pp. 67–87.

Elkhannoubi, H., & Belaissaoui, M. (2016). A framework for an effective cybersecurity strategy implementation. *Journal of Information Assurance & Security*, *11*(4).

Emm, D. (2013). Security for SMBs: why it's not just big businesses that should be concerned. *Computer Fraud & Security*, *2013*(4), 5–8.

Fenz, S., Heurix, J., Neubauer, T., & Pechstein, F. (2014). Current challenges in information security risk management. *Information Management & Computer Security*, *22*(5), 410–430.

Gafsi, M., Ajili, S., & Hajjaji, M. A. (2020). *Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT'18), Vol.1* (Vol. 146). https://doi.org/10.1007/978-3-030-21005-2

Herath, H. S. B., & Herath, T. C. (2014). IT security auditing: A performance evaluation decision model. *Decision Support Systems*, *57*, 54–63.

House, R. J. (1971). A path goal theory of leader effectiveness. *Administrative Science Quarterly*, 321–339.

Huang, K., & Pearlson, K. (2019). For What Technology Can't Fix: Building a Model of Organizational Cybersecurity Culture. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 6398–6407. https://doi.org/10.24251/hicss.2019.769

Iansiti, M., & Lakhani, K. R. (2014). *Digital ubiquity: How connections, sensors, and data are revolutionizing business*.

Irwin, A. S. M. (2018). Double-Edged Sword: Dual-Purpose Cyber Security Methods. In *Cyber Weaponry* (pp. 101–112). Springer.

Isaacson, W. (2012). The real leadership lessons of Steve Jobs. *Harvard Business Review*, *90*(4), 92–102.

KESSEL, P. V, & Allan, K. (2014). Get ahead of cybercrime: EY's global information security survey. *Insights on Governance, Risk and Compliance*, 40.

Klahr, R., Shah, J., Sheriffs, P., Rossington, T., Pestell, G., Button, M., & Wang, V. (2017). *Cyber security breaches survey 2017: main report*.

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, *6*(4), 239–242.

Lee, C.-M., & Chang, H. (2014). A study on security strategy in ICT convergence environment. *The Journal of Supercomputing*, *70*(1), 211–223.

Limba, T., Plėta, T., Agafonov, K., & Damkus, M. (2019). *Cyber security management model for critical infrastructure*.

Ma Prieto, I., & Pilar Perez-Santana, M. (2014). Managing innovative work behavior: the role of human resource practices. *Personnel Review*, *43*(2), 184–208.

Mehrfeld, J. (2019). Cyber security and Industry 4.0. *At-Automatisierungstechnik*, *67*(5), 361–363. https://doi.org/10.1515/auto-2019-0026

Prislan, K., Lobnikar, B., & Bernik, I. (2017). Information Security Management Practices: Expectations and Reality. *CECC*, 5–22.

Rohn, E., Sabari, G., & Leshem, G. (2016). Explaining small business InfoSec posture using social theories. *Information & Computer Security*, *24*(5), 534–556.

Shah, T., & Mulla, Z. R. (2013). Leader motives, impression management, and charisma: A comparison of Steve Jobs and Bill Gates. *Management and Labour Studies*, *38*(3), 155–184.

Shamim, S., Cang, S., Yu, H., & Li, Y. (2016a). Management approaches for Industry 4.0: A human resource management perspective. *2016 IEEE Congress on Evolutionary Computation, CEC 2016*, 5309–5316. https://doi.org/10.1109/CEC.2016.7748365

Shamim, S., Cang, S., Yu, H., & Li, Y. (2016b). Management Approaches for Industry 4.0. *Evolutionary Computation (CEC), 2016 IEEE Congress*, 5309–5316. https://doi.org/10.1109/CEC.2016.7748365

Shamim, S., Cang, S., Yu, H., & Li, Y. (2017). Examining the feasibilities of Industry 4.0 for the hospitality sector with the lens of management practice. *Energies*. https://doi.org/10.3390/en10040499

Shamim, S., Zeng, J., Shariq, S. M., & Khan, Z. (2019). Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: A dynamic capabilities view. *Information & Management*, *56*(6), 103135.

Siponen, M., & Willison, R. (2009). Information security management standards: Problems and solutions. *Information & Management*, *46*(5), 267–270.

Soomro, Z. A., Shah, M. H., & Ahmed, J. (2016). Information security management needs more holistic approach: A literature review. *International Journal of Information Management*, *36*(2), 215–225.

Stewart, A. (2012). Can spending on information security be justified? Evaluating the security spending decision from the perspective of a rational actor. *Information Management & Computer Security*, *20*(4), 312–326.

Such, J. M., Vidler, J., Seabrook, T., & Rashid, A. (2015). *Cyber security controls effectiveness: a qualitative assessment of cyber essentials*. Lancaster University.

Sun, C.-C., Hahn, A., & Liu, C.-C. (2018). Cyber security of a power grid: State-of-the-art. *International Journal of Electrical Power & Energy Systems*, *99*, 45–56.

Wang, Y.-B., & Ho, C.-W. (2017). No Money? No Problem! The Value of Sustainability: Social Capital Drives the Relationship among Customer Identification and Citizenship Behavior in Sharing Economy. *Sustainability*, *9*(8), 1400.

Wei, D., Lu, Y., Jafari, M., Skare, P., & Rohde, K. (2010). An integrated security system of protecting smart grid against cyber attacks. *2010 Innovative Smart Grid Technologies (ISGT)*, 1–7. IEEE.

Zammani, M., & Razali, R. (2016). An empirical study of information security management success factors. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 904–913.

A Change Management Perspective to Implementing a Cyber Security Culture

Trishana Ramluckan, Brett van Niekerk and Isabel Martins
University of KwaZulu-Natal, Durban, South Africa
ramluckant@ukzn.ac.za
vanniekerkb@ukzn.ac.za
MartinsM@ukzn.zac.za

DOI: 10.34190/EWS.20.059

Abstract: There has been an increasing prevalence of global cyber-attacks. Because of the possible breaches in information security, it has become pertinent that organisations change organisational and individual cultures to become more secure. However, there are challenges regarding the implementation of these processes within organisations. Organisations have become dependent on information systems, which stores large quantities of data and can be considered as one of an organisation's greatest assets. Whilst employees are considered as the next important asset, their negligence, whether intentional or not, and due to their possible lack of knowledge regarding information security, have also become one of the biggest threats to information security. Employees often fall victim to phishing scams, malware and ransomware attacks. Whilst many consider the implementation of information security awareness initiatives as a solution to this impending threat, more often than not organisations utilize presentations to address information security awareness training. This approach has not been successful as the target audience has difficulty in retaining the knowledge, and this often hinders its proper implementation. Change management not only involves the change in processes and tools but also focuses on the techniques used to manage the cultural change within organisations. This paper 'encodes' awareness training for information security and cyber security from a change management approach and provides best practice approaches in changing an organisation's culture from an insecure culture to a secure one.

Keywords: Information security, change management, training and development, organisational culture, threat mitigation

1. Introduction

With reference to the SANS Institute (n.d.) information security can be defined as, "protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction" Information security consists of three main elements namely technology, processes, and people. Although security measures including passwords, firewalls and biometric controls are important, they are not sufficient when attempting to mitigate vulnerabilities of systems and threats to information. The general requirement for the prevention or mitigation of threats to information relies on a combination of measures. With the three main elements of information security being technology, people and processes, there is a shift towards the people-oriented element. This places importance on areas of training and governance (Da Veiga and Eloff, 2007).

Change management plays a vital role in instituting an information security culture. The change management process consists of various stages which includes everything from change identification through to post-implementation review and request closing (INFOSEC, 2020). The purpose of the change management process is to ensure that all changes follow a controlled and systematic implementation process. Furthermore, the key objective of change management within information security is the alignment of people and culture to information security best practices while at the same time mitigating any resistance to change (Kortan and von Solms, 2012)

This paper aims to propose a framework/model for the change management process to instil a cyber security culture within and organisation. Section 2 presents the literature review, which provides background information to secure cultures and introduces the Resource Based View (RBV) and ADKAR (Prosci, n.d.) change management frameworks/models. Section 3 provides a discussion on the culture and proposes the framework. The paper is concluded in Section 4.

2. Literature

The relevant literature is discussed in this section which includes information security culture, change management and a brief overview of the change management models.

2.1 Information Security Culture

The organisational culture within any business is important as it defines the way in which the employee views the organisation (Ulich, 2001, cited in Schlienger and Teufel, n.d; van Niekerk and von Solms, 2006; Kortan and von Solms, 2012). It can be described as a collective but adaptable phenomenon and can be easily influenced by company executives over time. The two main elements of organisational culture are assumptions and beliefs, which are often expressed in the collective values, norms and knowledge in an organisation's environment. The resulting factor that emerges is that these collective norms and values promote or instil a particular behaviour of the employees in the environment. Ultimately, the organizational culture has a crucial impact on the corporate success (Rühli, 1991, cited in Schlienger and Teufel, n.d.; van Niekerk and von Solms, 2006; Kortan and von Solms, 2012).

An organizational culture usually consists of varying subcultures founded on functions. Information security culture can be considered as a subculture or function of the organisation. Generally, the organisational culture should support all activities and with it information security as a subculture should become a normal element in employees' daily activities (Zaini, Masrek, Sani, and Anwar, 2018). With reference to Zaini et al., (2018) the three layers of Information Security Culture and their interactions are illustrated in Figure 1.

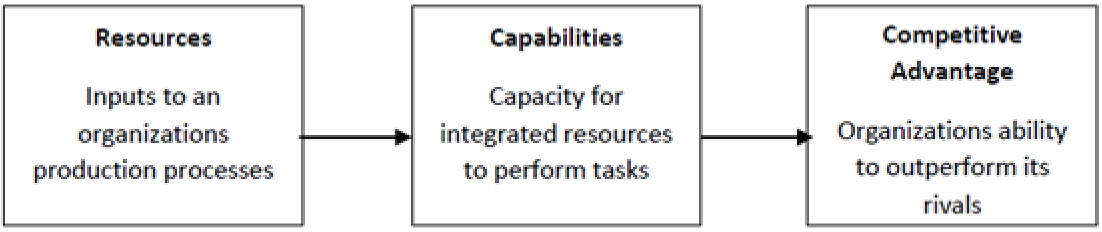

Resources
Inputs to an organizations production processes
Capabilities
Capacity for integrated resources to perform tasks
Competitive Advantage
Organizations ability to outperform its rivals

Figure 1: Three Layers of Information Security Culture

Source: Zaini, Masrek, Sani and Anwar (2018)

According to the International Telecommunications Union (2008), there is a new reality of cyber space and as such cybersecurity or information security requires the creation of an information security culture, with its foundations based on the norms of proper behaviour as well as the ability to prosecute online criminals. However, it must be noted that with the dependency on the cyber realm, more security challenges are created.

Zakaria (2006) indicates that employees require basic information security knowledge for a secure culture to successfully be instilled in the organisation. Ruighaver and Maynard (2006) indicate that the focus of security culture is on the end-user, however there is a management component, and it needs to be considered from a management perspective. Da Veiga (2016) also considers the individual and organisational levels of security culture, and also includes national and international culture. It is therefore important to consider both the organisational and individual levels when instilling a security culture. The Organisation for Economic Co-Operation and Development (2002) (cited by Kortan and von Solms, 2012) provides a few guidelines on the components that constitute a cybersecurity/information security culture. These include:

- The promotion of a secure online environment for all users
- Creating awareness about online risks and the counter measures available
- Increasing the confidence and knowledge of users in information systems and networks, as well as the way in which they are provided and utilised;
- Acting as a frame of reference for the development and implementation of cyber/information security measures
- Encouraging co-operation and information sharing, as appropriate, amongst all the participants in the development and implementation of security policies, practices, measures and procedures.
- Promoting the consideration of security as an important objective amongst all participants involved in the development or implementation of standards.

Ryttare (2019) provides key themes for change management related to cyber security, including: senior management awareness, involvement of all employees, security needs to be highlighted in a positive manner, there needs to be understanding of roles and consequences, customised training to ensure relevance, interactive learning, feedback channels, rewards, and continuous change efforts. Jobraj and van Niekerk (2015) reflected on an organisational cyber security awareness approach following the ADKAR change management model, where the focus was on the individual, departmental, and organisational levels. The main themes corresponded to those highlighted by Ryttare (2019), and the activities to achieve these considerations were: an interactive roadshow with activities, articles in internal publications, e-mail alerts, flyers and posters, screen savers and pop-ups, and table-top exercises for the executives (Jobraj and van Niekerk, 2015).

2.2 Change Management

Change management can be defined as the processes, tools and techniques involved in managing the people approach to change with the objective of achieving a successful business outcome (Prosci, 2020).

With reference to Burnes (2004) change is a continual characteristic of the organisational environment, at both the operational and strategic levels. It is important to note that organisational change and organisational strategy must be aligned to each other (Barney, 2004). According to Graetz (2000: 550) many would agree that the primary task for management today is the leadership of organisational change especially with increasing global deregulation and the rapidly changing nature of technological innovation. According to By (2005) are three main phases to change management which includes:

Phase 1: Preparing for Change
This Phase entails defining the change management strategy, preparing the change management teams and the development of the change model. The key objective to Phase 1 is to identify the points of resistance and establish mitigating mechanisms for the resistance to the change.

Phase 2: Managing the Change
This Phase involves the creation of change management plans and the action plan to be implemented. The main objective of Phase 2 is the development and implementation of the resistance to change management plan.

Phase 3: Reinforcing the Change
Phase 3 entails ensuring the success of the change that has been implemented by collecting and analysing feedback, identifying gaps while managing resistance as well as implementing any corrective actions.

The Resource Based View and the ADKAR frameworks/models conceptualise the ideals of the change management process and are discussed further in Sections 2.3.1 and 2.3.2.

2.3 The Change Frameworks/Models

There are two relevant frameworks/ models pertaining to change management that will be briefly described in Sections 2.3.1 and 2.3.2. The two frameworks/ models are the Resource Based View (RBV) model and the ADKAR model.

2.3.1 The Resource Based View (RBV) Framework/Model

RBV is often regarded as a common approach used by organisations to achieve competitive advantage. The approach emerged in the 1980s as its supporters believed that the sources of competitive advantage were based internal to the organisation. The supporters are of the opinion that it is much easier in the exploitation of external opportunities through existing internal resources. Within the RBV model resources become an imperative element in assisting companies achieve better organizational performance. According to Zaini, Masrek, Sani, & Anwar (2018), the foundation for competitive advantage lies in the application of valuable tangible or intangible resources available to a company. Tangible assets refer to the physical which includes property (land and buildings), equipment (machinery) and capital. These physical resources can simply be acquired and provide little to no real advantage to companies on a long-term basis, as competitors can easily obtain similar if not identical assets.

Intangible assets are considered as having no physical presence; however, these can still be owned e.g. brand reputation, intellectual property. These intangible assets accumulate over time and is not easily available in the open market. Intangible resources usually stay within a company and are the main source of sustainable competitive advantage. With reference to Zaini, Masrek, Sani, & Anwar (2018), information security is built on the notion that an organization has the capability to extend its competitive advantage by simply providing information resources confidentially and immobile through practices that would enhance its secrecy and security (Nelson & Romer, 1996). Within a cyber security context, there can be two different competitors: rival corporations in traditional business competitions, and those who have malicious intent against the organisation, for instance to steal funds through scams. Therefore, a security culture can form a competitive advantage against rival companies in that clients are more confident in the organisation's ability to maintain operations and protect their personal details compared to rival companies. In a pure security sense, improved security awareness or culture makes it more difficult for scams to be effective, providing improved competitive capability against hackers (i.e. greater resilience).

RBV consists of two main assumptions on the resources, which entail them being heterogeneous and immobile.

Heterogeneous refers to the assumption that the skills, capabilities and other resources that organizations own varies between organisations, thereby allowing varying strategies to promote competitive advantage. Companies usually reuse or replicate what each one does, allowing for perfect competition. However, in reality this does not occur as some companies, which are exposed to the same external and competitive forces (same external conditions), are able to implement different strategies and outperform each other. Therefore, RBV assumes that companies achieve competitive advantage by using their different bundles of resources.

Immobile refers to the assumption that resources are not mobile. Therefore, they cannot move between companies in a short period. Because of the resources' immobility, it becomes extremely difficult for a competitor to replicate strategies. Intangible resources including processes, knowledge and/or intellectual property are usually considered as immobile (Jurevicius, 2013). However, with basic information security knowledge, this may not strictly be the case as this knowledge is universal. Human resources moving corporations may benefit the new employer if the skills and knowledge allow for the new employer to improve its security capability.

2.3.2 ADKAR Framework/Model

The ADKAR model was created by Jeff Hiatt in 2003, with the intention of the developer for it to become a coaching and change management tool to assist employees with the change processes within organisations. The change process within organisations is generally met with contempt and resistance. The ADKAR model assists in the form of a management tool to identify the issues of the change process. ADKAR is an acronym representing the five building blocks instituting successful change illustrated in Figure 2. The building blocks are Awareness, Desire, Knowledge, Ability and Reinforcement.

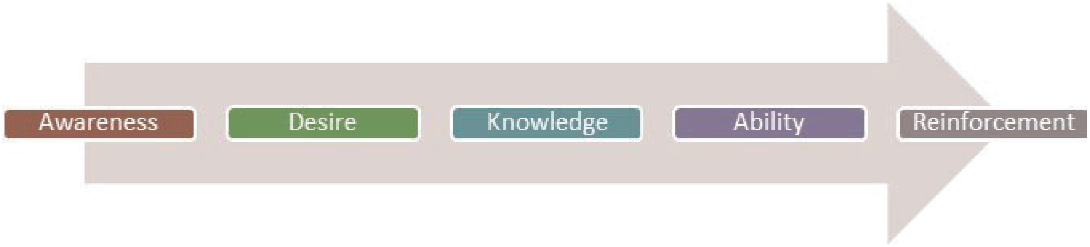


Figure 2: The Five Building Blocks of ADKAR

Source: Adapted from Prosci (2020)

The five stages can be briefly described as follows (Prosci, 2020):

- **Awareness** refers to the awareness of the need to change. This stage relates to the communication of the vision for the future as well as any other information about the change.
- **Desire** refers to the desire participate and support the change which is often influenced by a number of factors. These may include dissatisfaction in its various forms, personal experience, intrinsic forms of motivation and the possible acquisition of status and/or authority.
- **Knowledge** of how to change. Knowledge empowers staff to understand what the change is as well as the outcome of the change which may include their role in the change. Education and training are usually required to enhance the knowledge, skills, and behaviours of employees. Training and

education are important to ensure the smooth functioning in the “changed” environment. Training and education may be provided through a combination of activities e.g. workshops, retreats, on-line modules. Information provided should be easily available and accessible.

- **Ability** refers to the ability to implement the change on a daily basis and focuses on the application of new skills, processes and behaviours. This stage also takes into consideration the potential issues that may arise with the implementation. During this phase, staff are offered practice time while being coached. Areas of coaching may include for staff to develop an understanding of their new role and responsibilities, process and system changes, and new technology. Generally, activities that will support their function and team dynamics within the organisation. Managements’ role is important in this phase for assisting in the elimination or mitigation of any barriers, and in addition may provide expertise to eradicate any obstacles.
- **Reinforcement** to keep the change in place. This stage involves recognition of employees and rewards. The recognition and rewards for employees may include anything from personal recognition, to compensation, and benefits.

With reference to ADKAR change consists of two elements namely the organisation and the employees. The change process will only succeed if these two elements occurs work symbiotically. The stages of the ADKAR model function in a modular manner meaning the stages work independently of each other so if there is an issue with any particular stage, it can be resolved without causing much disruption to any of the other stages. It can be considered as a targeted approach as the ADKAR model focuses on the element/stage with the highest success rate (By, 2005). The ADKAR model can be considered as an evaluative model, providing feedback insight as well as hindsight of the stages.

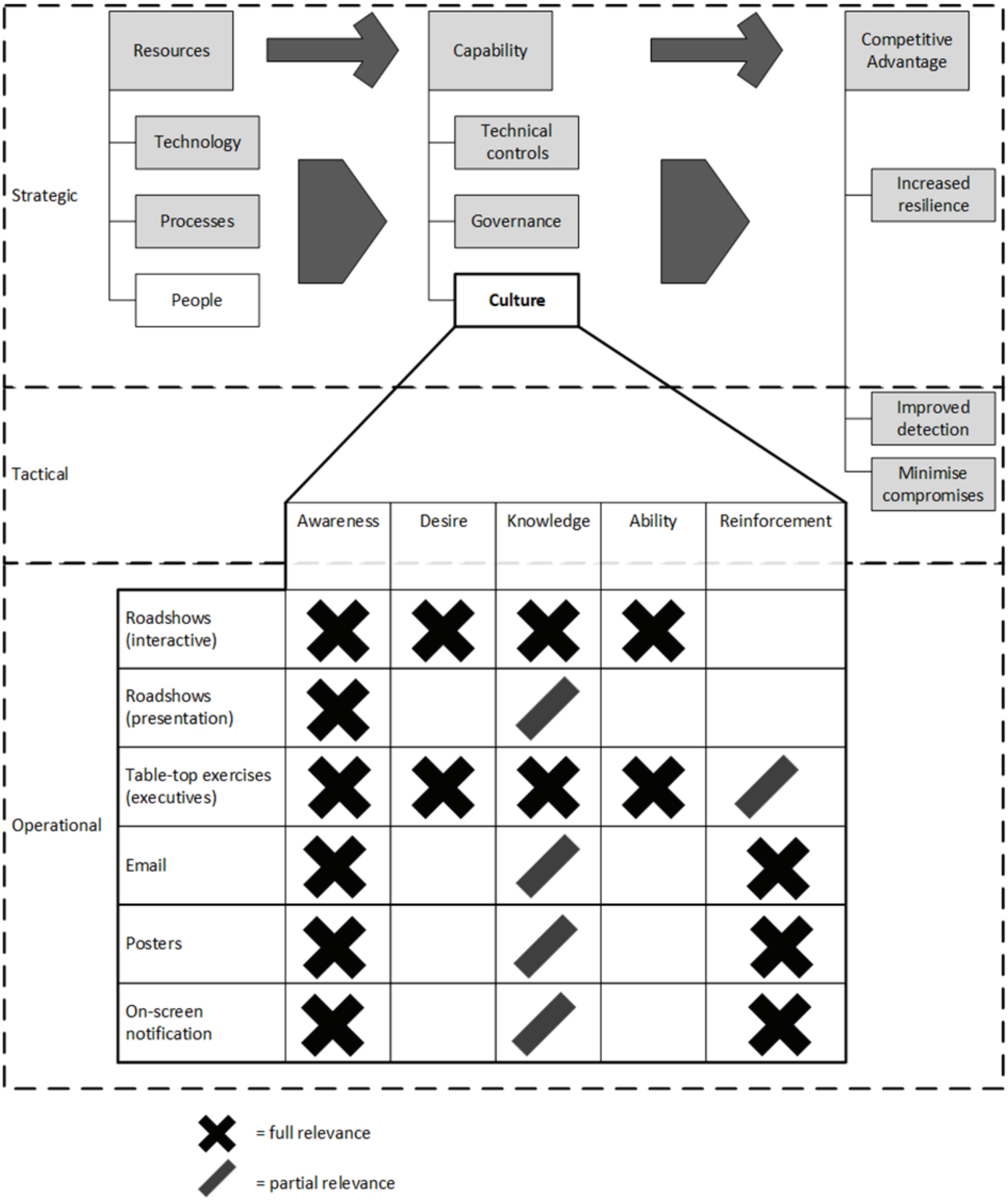
3. Discussion and Proposed Framework

As part of this paper, a framework has been developed which proposes a combination of the RBV and ADKAR models. The RBV and ADKAR frameworks/models have been considered in line with the three levels of management namely the strategic, tactical and operational levels. Further to this the three elements of information systems and information security are included, people, process and technology. These three can be aligned as the actual systems and corresponding technical controls as a mitigating element; governance, which acts as the control of processes; and, culture referring to the people-oriented element.

As illustrated in **Figure 3**, RBV falls within the ambit of strategic management and for the purposes of an information security culture consists of resources and capabilities which may lead to competitive advantage as previously illustrated in **Figure 1**. The Resources consist of technology, processes and people. The corresponding Capabilities consist of the technical controls, governance and security culture, while competitive advantage consists of increased resilience at the strategic level and improved detection and the minimalization of compromises at the tactical level (assuming the governance, technical controls and culture are effectively implemented). Culture, while pitched at the strategic level, branches down to the tactical level, where the stages (Awareness, Desire, Knowledge, Ability and Reinforcement) of the ADKAR are placed. The possible activities to enable a culture change are aligned to the ADKAR model and are placed in the operational level. The activities (following from Ryttare (2019) and Jobraj and van Niekerk (2015)) consisted of interactive roadshows, roadshows through presentations, table-top exercises for executives, email, posters and on-screen notifications.

At the operational level, a matrix was developed to illustrate the coverage of the elements of the ADKAR model at the tactical level with the activities at the operational level. The full cross represents full relevance, with the half cross representing only partial relevance. As is illustrated in the matrix, the table-top exercises for executives appears most successful in relation to culture as it is relevant for Awareness, Desire, Knowledge, Ability stages and partial relevance for Reinforcement of the ADKAR model due to the interactive nature of the activity. This is followed by the interactive roadshows (full relevance for Awareness, Desire, Knowledge, Ability), which due to the tasks performed as part of the roadshow increases enjoyment and the learning of skills. This is followed by emails, posters and on-line notifications (full relevance for Awareness and Reinforcement and partial relevance for Knowledge), which are particularly useful for reminders of what was covered in the Interactive Roadshows; these also are useful for alerts of specific threats that are imminent or underway. The Roadshows by presentation demonstrate least success with full relevance for Awareness and partial relevance for Knowledge; the impact of these is low due to ‘death by PowerPoint’.

The proposed matrix can be expanded with more relevant activities, assessed according to the specific audience(s) within an organisation. The matrix will assist in selecting a range of activities to cover all areas of the ADKAR model for all audience groups.



	Awareness	Desire	Knowledge	Ability	Reinforcement
Roadshows (interactive)	✖	✖	✖	✖	
Roadshows (presentation)	✖		╱		
Table-top exercises (executives)	✖	✖	✖	✖	╱
Email	✖		╱		✖
Posters	✖		╱		✖
On-screen notification	✖		╱		✖

Figure 3: Proposed Framework

4. Conclusion

This paper has provided some insight into the importance of the implementation of an information security culture within organisation. The proposed framework presented a combination of the RBV and ADKAR change management models that may assist in the implementation of an information security culture. An information security culture must consider the human factors in order for an organisation to improve its overall information security efforts. With reference to van Niekerk and von Solms (2006) an understanding of the various elements of an information security culture is required as the relationships between these elements can become problematic when considering corporate culture. Schein, quoted in Cecil (2013), defines organizational culture as "the formation of a pattern consisting of shared basic assumptions. These assumptions have been studied by a group on the basis that it has found resolutions to the problems of external adaptation and internal integration which can be considered as a product of joint learning". This definition can be used to aid understanding of an information security culture. From this definition, it can be understood that knowledge forms the foundation of corporate culture without which information security cannot be guaranteed. The RBV and ADKAR models combined provides the human factors or elements that need to be considered when implementing an information security culture.

References

Barney, J. B. (1991) "Firm Resources and Sustained Competitive Advantage", *Journal of Management*, 17(1), pp. 99–120.

Burnes, B. (2004) Managing Change: A Strategic Approach to Organisational Dynamics, 4th ed. Prentice Hall: Harlow

By, R. (2005) "Organizational Change Management: A Critical Review", *Journal of Change Management*, 5(1), pp. 369-380.

Cecil. (2013) "Edgar Schein: Organizational Culture and Leadership", #Hypertextual, [online], accessed 18 January 2020, https://thehypertextual.com/2013/01/17/edgar-schein-organizational-culture-and-leadership/

Da Veiga (2016) "A Cybersecurity Culture Research Philosophy and Approach to Develop a Valid and Reliable Measuring Instrument", SAI Computing Conference 2016, July 13-15, London, UK, pp. 1006-1015.

Da Veiga, A. and Eloff, J. H. P. (2007) "An Information Security Governance Framework", *Information Systems Management*, 24(4), pp. 361 – 372.

Graetz, F. (2000)" Strategic change leadership", *Management Decision*, 38(8), pp. 550–562.

INFOSEC Institute. (2020). "Change Management and The CISSP" [online], accessed 18 January 2020, https://resources.infosecinstitute.com/category/certifications-training/cissp/domains/security-operations/change-management/#gref

ICT Applications and Cybersecurity Division Policies and Strategies Department ITU Telecommunication Development Sector. (2009). "Understanding Cybercrime: A guide for Developing Countries" [online], accessed 18 January 2020, https://www.itu.int/ITU-D/cyb/cybersecurity/docs/itu-understanding-cybercrime-guide.pdf

Jobraj, T., and van Niekerk, B. (2015) "Information Security Awareness at a South African Parastatal: Challenges and Successes", presented at the 10th International Conference on Cyber Warfare and Security, 24-25 March, Kruger Park, South Africa.

Jurevicius, O. (2013) "Resource Based View: What makes your company unique?" *Strategic Management Insights*, 14 October, [online], accessed 18 January 2020, https://strategicmanagementinsight.com/topics/resource-based-view.html

Kortan, N. and von Solms, R. (2012) Fostering a cyber-security culture: a case of South Africa, *Proceedings of the 14th Annual Conference on World Wide Web Applications*, 7-9 November 2012, Durban, South Africa.

Nelson, R. and Romer, P. (1996) "Science, Economic Growth, and Public Policy", *Challenge*39(1), pp. 9–21.

PROSCI. (n.d) The PROSCI ADKAR Model, [online], accessed 18 January 2020, https://www.prosci.com/adkar

Ryttare, E. (2019) Change Management: A Key in Achieving Successful Cyber Security, Master's project, Luleå University of Technology, [online], accessed 27 January 2020, https://pdfs.semanticscholar.org/ea5b/1370d1919a50b8830661090df6376a2fe9dc.pdf

Ruighaver, A.B., and Maynard, S.B. (2006) "Organizational Security Culture: More Than Just an End-User Phenomenon", In: Fischer-Hubner, S., Rannenberg, K., Yngstrom, L., Lindskog, S. (eds.), *Security and Privacy in Dynamic Environments*, IFIP International Federation for Information Processing, Volume 201, Springer: Boston, pp. 425-430.

Ruhli, E. (1991) Unternehmungskultur - Konzepte und Methoden. In: E. Riihli and A. Keller, Eds. Kulturmanagement in schweizerischen Industrieuntemehmunnen. Bern und Stuttgart, Paul Haupt Verlag: 11-49, cited in Schlienger, T and Teufel, S. (n.d) Tool Supported Management of Information Security Culture [online], accessed 18 January 2020, https://link.springer.com/content/pdf/10.1007%2F0-387-25660-1_5.pdf

SANS Institute. (2020) Information Security Resources, [online], accessed 18 January 2020, https://www.sans.org/information-security/

Ulich, E. (2001) Arbeitspsychologie. Zurich, vdf, Hochschulverlag an der ETH Zurich,cited in Schlienger, T and Teufel, S. (n.d) Tool Supported Management of Information Security Culture [online], accessed 18 January 2020, https://link.springer.com/content/pdf/10.1007%2F0-387-25660-1_5.pdf

Van Niekerk, J and von Solms, R. (2006) "Understanding Information Security Culture: A Conceptual Framework", *Proceedings of the ISSA 2006 from Insight to Foresight Conference*, 5-7 July 2006, Balalaika Hotel, Sandton, South Africa

Zaini, M. K., Masrek, M. N., Sani, M. K. J. A., and Anwar, N. (2018) "Theoretical Modeling of Information Security: Organizational Agility Model based on Integrated System Theory and Resource Based View", *International Journal of Academic Research in Progressive Education and Development*, 7(3), pp. 390–400.

Zakaria, O. (2006) "Internalisation of Information Security Culture amongst Employees through Basic Security Knowledge", In: Fischer-Hubner, S., Rannenberg, K., Yngstrom, L., Lindskog, S. (eds.), *Security and Privacy in Dynamic Environments*, IFIP International Federation for Information Processing, Volume 201, Springer: Boston, pp. 437-441.

PhD Research Papers

Cyberattacks and the Prohibition of the Use of Force under Humanitarian Law with Reference to the Tallinn Manual

Khalil Akbariavaz[1], Pardis Moslemzadeh Tehrani[2] and Johan Shamsuddin bin Haj Sabaruddin[3]
[1]University Malaya, Malaysia
[2]Faculty of Law; Deputy Editor of the University Malaya Journal of Law and Policy (UMJLP) University of Malaya, Malaysia
[3]Dean of the UM Faculty of Law, Malaysia
tafakor.kh@gmail.com
pardismoslemzadeh@um.edu.my
johans@um.edu.my

DOI: 10.34190/EWS.20.508

Abstract: Cyberattacks can alter the outcome of modern war through the inclusion of governmental and non-governmental actors and new technologies. The abilities of the perpetrators to cause destruction of both material and non-material assets and even injury and death to individuals through cyberspace have challenged traditional definitions of the use of force. The present study aims to investigate approaches to answering novel questions encountered by jurists attempting to deliberate on such matters. Answers to these questions should be reached within existent legal frameworks or codified within a novel legal framework. This study poses the question as to whether these attacks are covered by the prohibitions stipulated in Article 2(4) of the UN charter. It concludes that due to the novelty of the phenomenon, existing legislation inadequately regulates the use of cyberspace for hostile acts. Preliminary efforts such as codification of the Tallinn Manual by NATO have not produced the desired legal outcomes. Thus, the general principles and rules of international humanitarian law should be utilized for regulating hostile relations in cyberspace. As stated by Marten Condition a century ago, the absence of special regulations should not bar the enforcement of general regulations and the application of principles, common rules and duties. Governments require recognition of their right to defend vital infrastructure. Assuming an attacked state can identify the source of a cyberattack and attribute it to another government, response options include referring the matter to the UN Security Council or the International Court of Justice or undertaking retaliatory cyberinterventions or instituting armed reactions, assuming the principle of proportionality is observed.

Keywords: Cyberattacks, use of force, humanitarian laws, Tallinn Manual

1. Introduction:

In 2009, a military international organization based in Tallinn, Estonia, known as the Supreme Cyber Defense Cooperation Center and instantiated by NATO, invited an international group of independent experts to author a set of guidelines for the creation of a law that could govern cyber-defense. The project, which was carried out by experts and international law researchers, sought to generalize legal and regulatory norms in modern warfare. International law guidelines regarding cyberwar stem from an expert-oriented process which created a non-binding document to expand cyberlaw and clarify existing documents about cyber-interventions and cyber hostilities.

The Tallinn manual can serve as an example of cooperative international law-making for the cyber-sphere, but due to its non-binding nature, it is inadequate as a basis for the current analysis. This discussion references other international documents and studies to more fully explore the creation and interpretation of existing international cyberlaw and justify interventions such as binding prohibitions and force. The prohibition of force is referred to in Paragraph 4 of Article 2 of the UN Charter, which states that "…all members should avoid threatening or applying force in their international relationships against the territorial integrity or the political independence of any country or any other method that is in contradiction to the United Nations' intentions". The prohibition is binding on members and non-members since these Articles express common rights. The only exception in the charter pertains to legitimate individual or collective defense.

A close examination of Paragraph 4 of Article 2 prompts the following questions:

- What does the concept of "force", as used in Article 2(4), connote exactly?
- Are economic or political embargoes considered "force"?
- Can "cyber force or power" be construed as force for the purpose of this provision?

- Can cyberwar be considered a use of force prohibited by Article 2(4)?
- Do "against territorial integrity" and "political independence" explain or constrain?

The prohibition on force laid out in Article 2(4) of the Charter has been asserted by the International Court of Justice in deliberations concerning armed activities inside the Congo to be the cornerstone of the UN Charter (ICJ, 2005). However, no clear-cut definition of "force" or "war" is offered either by the Charter or other international documents. Thus, the concepts of "war" and "force" as used by the Charter require elucidation. Cyberattacks and prohibition on the use of force are here considered in light of the provisions of humanitarian law with reference to the Tallinn Manual.

2. War

In its traditional sense, war is a state of combat between states in which the armed forces of the parties engage in violent interventions; armed contact is continuous and mutual (Oppenheim, 1906). Oppenheim defines war as confrontation between states through military forces with the intent of gaining and domination (Oppenheim, 1906). Aaron sees war as a social action stemming from the will of the organized political assemblies. Dinstein defined war as the clash between governing states realized in a technical form, to wit through the formal declaration of war by a state against another, or in a physical form, i.e. by the full-scale use of armed forces in the relationships between the two states disregarding any sort of formal announcements (Dinstein, 2016).

2.1 Cyberwar

Technological progress has transformed cyberspace into a potential battlefield which may be used alone or in combination with the other operation fields for hostile relationships (Yde, 2010). After land, sea, air and space, cyberspace is the fifth battle space (Tsagourias & Buchan, 2015). Generally, cyberspace is used for information exchange and can be defined as communication between computers. Cyberwarfare remains undefined, though invasive interventions intended to destroy or disrupt information on computers or networks may count. Emphasis is often also put on the effects of such attacks (Waxman, 2011).

2.2 Evolution

Technological progress has led states to conclude that cyberwar has many advantages, including anonymity, low cost, high impact, attribution difficulty and lack of effective international treaties and norms. Increased use of technology and the internet by states affords weak points, especially where security is lacking. Nongovernmental players, not only states, are potential actors. Precision in both attack and defense is difficult, and even the very presence of hostile operations in virtual battlefields may be hard to determine. Further, though attacks may be hidden, physical effects may follow (Waxman, 2011).

WWII has been considered a partial cyberwar due to the infiltration of airplanes' radio frequencies. The first cyberattack was the Morris Worm of 1988 which disrupted the cyber infrastructure of the USA. Other important recent cyber interventions include the 2007 Estonia intervention, the Stuxnet Virus which targeted Iran's uranium enrichment equipment and malware in Indonesia. The importance of viruses is in their physical effects. Cyberwar is more than just simple attacks on website or malicious email attachments (Kelsey, 2008). Between 2011 and 2013, Canada and the USA's defense and financial institutions in South Korea were attacked. "Red October" attacked countries in Eastern Europe and Central Asia. In 2018, the number of cyberattack victims reached 87 million individuals, later rising to 100 million, on Facebook alone. CyberGuard's cyberattack workgroup in 2019 reported largest cyberattack and robbery to date. It involved the compromise of 733 million emails and 21 million passwords and the theft of 87 GB worth of personal information. 991904772 emails and linked personal data was published online. Cyberattack goals include, for example, shutting down aerial defenses, infiltrating defense networks and disrupting power plants or media stations. Seriousness of damage caused ranges from minimal, as where systems are hijacked for use in attacking another party, to devastating, for example where damage to infrastructure results (Brenner & Clarke, 2010).A "simple" cyberattack might even result in a global war since it might endanger international peace and security.

Defining cyberattack and cyberwar poses difficulties. Some sources specify all kinds of online interventions as cyberattack; others restrict the term more. The American National Research Council in 2009 defined cyberattacks as "the use of intentional interventions for changing, disrupting, deceiving or reducing and/or eliminating the enemy's computer systems or networks or information and/or the programs existent in these systems and/or networks" (Hathaway et al., 2012). Rule 30 of the Tallinn Manual states: "cyberattack includes invasive or defensive cyber operations that is logically expected to cause injury or death to individuals or impose damage to the objects". Thus, it seems that cyberattack, when followed by effects similar those of ordinary weapon under war conditions, can be evaluated within the framework of the war laws. According to Paragraph 1 of Article 2 of the relevant UN charter, which refers to "the governance equality of all the members", governments are implicitly prohibited from taking any measures against other states that may cause it to become unable to administrate its … affairs". Note that cyber and traditional wars differ. Cyberwars occur in cyberspace as noted above, whereas traditional wars occur on or in land, sea, air or space and are armed clashes performed against the governance, territorial integrity or political independence of other states in specific battlefields (United Nations, 1974).

The problem relates to ways in which existing regulations can be enforced in a new and completely different domain. New technologies bring challenges for legal interpretation. This issue does not imply existing regulations are unenforceable, but it does make the interpretation of the regulations more challenging and riskier (United Nations, 1974).

3. Cyberattack in International Humanitarian Law

Applying humanitarian law to cyberattack is difficult. The main question is whether existing humanitarian law can be applied or whether new rules and regulations are needed.

Some have argued that since cyberattacks may lead to armed clashes they come within the scope of existing international rules and regulations, i.e. international humanitarian laws, relating to jus in bellum, which is a branch of international law that order the behavior of conflicting states. Jus in bellum has traditionally drawn on the Geneva and Hague laws. The Geneva laws include treaties relating to prisoners, captives, injured and sick individuals who cannot fighting. They are considered all-inclusive and imperative. The Hague regulations relate to 1899 and 1907 conventions regarding methods and instruments of war (International Committee of the Red Cross, 1989). At present, there is no consensus regarding the enforcement of humanitarian laws in cyberwar, and no universally accepted common law or procedure. Some jurists believe that the limitations originating from existing laws should be enforced for cyberattacks, but the applicability of humanitarian laws to these attacks is still unclear.

Much discussion exists regarding the insufficiency of current international laws on cyberspace. Some international jurists believe that humanitarian laws cannot be applied to cyberwar because there is no physical operation. According to Article 2 of the Geneva conventions and Paragraph 3 of Article 1 of the first Protocol, attack should be armed before international humanitarian laws can be enforced. Thus, if a cyberattack reaches the threshold of an armed clash, the international humanitarian laws are sufficiently flexible to cover it. Some researchers, like Schmitt, believe that where cyberattacks involve armed attack the victimized state can resort to legitimate defense as per Article 51 of the UN Charter. The Red Cross states humanitarian law applies to cyberattack involving armed attack since instruments and methods of war change over time. Article 2 of the Geneva conventions states "besides the regulations that should be enforced during peace time, the convention will be also enforced in case of the occurrence of formal war or any armed hostility between two or several states of the endorsing respected states even if one of them is found having confirmed the existence of a war state". Thus, regulations governing traditional wars be also applied to cyber clashes, though not to all.

Humanitarian laws may apply only to attacks that have reached an armed level or that have been conducted during armed clashes. Of course, there are discrepancies between jurists regarding the extent of intensity that constitutes international armed clash (Roscini, 2014). In terms of evaluating intensity, the type and number of warriors and weapons used, as well as the attack's scope and duration, can be taken into consideration (International Law Association, 2010).

The dominant theory regarding cyberattacks holds that existing international laws are sufficient for addressing them. This is supported, for example, by parts of the Tallinn Manual, which states that international law regarding public property can be enforced for cyberattack. Humanitarian laws were not formulated for cyberwars but are flexible enough to embrace them. This is reflected in Martens clause, as seen in the introduction to the second 1899 convention and in Paragraph 2 of Article 1 of the first Geneva protocol. When there is no international agreement, humanity principles, customary rules and public conscience can be enforced. Similar issues were considered by the IJC in respect of the legitimacy of using nuclear weapons. The

Red Cross has adopted a similar stance based on the argument that regulations are not limited to a certain method or weapon (Geiss, 2010).

If cyberattack is considered as armed attack, humanitarian law prohibits attacking civilians, nonmilitary properties, protected individuals and properties, and limits allowable material damage. Hostile actors are restricted from using the cyber infrastructures of neutral parties to wage war (Schmitt, 2012). Regulations resulting from a war rights perspective regulate the hostiles' behaviors (Geiss, 2010). International humanitarian laws can thus be enforced for cyberwars, although the manner of enforcement attention. Care should be used in discussion in this area to pave the way for the creation of internationally accepted common laws in the intended area.

4. Prohibition of Force

Although a principle prohibiting force is regarded as flowing from the UN Charter, the charter does not offer any definitions for such terms as "using force" and "armed attack". It is yet to be agreed whether cyberattack is a use of force. If it is accepted that it is, then threats of cyberattack will be also covered by the same prohibition. The phrase "using force" has been claimed to be ambiguous. The authors of the charter insisted that this expression has to be minimally interpreted. It has not been agreed to apply "force" to economic or political conflict. This may imply that the "force" provisions of Article 2(4) of the charter may be limited to military force. However, it cannot be concluded that the prohibition of force only applies to weaponry featuring explosive and firing power. It is globally accepted that chemical, biological or radiological attacks are considered armed attack too. Therefore, cyberattack can be considered armed attack if it attains the prescribed effect and scale. Failing to regulate cyberattack would reject the goals and spirit of the UN charter and the values of international society. The Tallinn Manual indicates that weapons are not necessary for the term "armed" in "armed attacks" to apply. However, when Estonia's presidency, congress, ministries, political factions and banks were subjected to cyberattack in 2007, NATO declared it did not recognize cyberattacks as military interventions triggering Article 5 of NATO's charter regarding collective legitimate defense.

Generally, for classifying an attack as large and important, three scales, the instrument-based approach, the target-based approach and the effects-based approach, are used. The ICJ (1996) did not consider the instrument-based approach appropriate for investigating a case before it at that time (Azzopardi, 2013). The Security Council of the UN confirmed this conclusion by accepting the legitimate defense right in response to the attacks on 11th of September 2001, where hijacked airplanes were used as weapons. The target-based approach has faced numerous problems in the justification stage. Despite the advantages of the first two approaches, therefore, the third one was followed due to problems with the first two.

The ICJ in the Nicaragua case (1986) adopted the criteria "scales and effects" approach while explicitly affirming the importance of forms of force by stating that it is the effect and scale test that distinguishes an armed attack from a mere frontier incident (ICJ, 1986). It was accordingly shown that although any armed attack involves force, not all uses of force can be considered armed attacks (ICJ, 1986). For example, in the oil platforms' case (2003) the ICJ held that "since the Iranian interventions did not reach the level of armed attack (ICJ, 2003), the USA had no right to react based on legitimate defense". It could be argued that cyberattacks targeting vital infrastructure and causing vast loss of life and financial damage should be placed within the limits of article 4(2) of the charter. The charter intended to apply to any intervention jeopardizing international peace and security. Article 4(2) of the charter prohibits the use of force in any way contradicting the charter's goals and the use of force against the territorial integrity of a state. The expression "territorial integrity" can be interpreted as encompassing the cyberspace of a state. Thus, extensive sabotage in this space is an invasion of the territorial integrity of that government. Schmitt offered a normative framework for the evaluation of attacks on computer networks which attempts to link the instrument- and target-based approaches which has been widely discussed by jurists. Schmitt lists seven characteristics necessary for a cyberattack to be classified as an armed attack. Where an attack has the attributes, it may be considered an armed attack; otherwise it falls short (ICJ, 2003).

These seven features are:

1. Extent of damage: An action resulting in the death of individuals or severe damage to property may be viewed as a military intervention. The less extensive the damage caused, the lower the likelihood of its being seen as a use of force.

2. Immediacy of results: When the effects of an attack come about within seconds or minutes, the operation causing it may be considered military. When it takes weeks or months for effects to be revealed, they are more likely to be envisioned as diplomatic or economic interventions.
3. Directness: If the intervention is the sole cause of the result, it is more likely to be considered as force. The weaker the relationship between the cause and effect, the weaker the military nature of the action will become.
4. Invasion: An invaded frontier is a sign of a military operation; actions carried out outside the borders of a target state are possibly diplomatic or economic.
5. Measurability: If the effect of the attack can be easily and objectively measured, such as by taking a photo of a burning place that has been targeted, it can be seen as a military operation. The more subjective the damage evaluation process; the more the action is likely to be diplomatic or economic.
6. Legitimacy: Governmental players alone may use legitimate physical force. The legitimacy of attacks in cyberspace may be difficult to determine.
7. Responsibility: If a government explicitly takes responsibility for sabotage actions, such actions are more likely to be considered military operations. Where responsibility is not unequivocally attributed, actions may be nonmilitary (Cambridge University Press, 2013).

These criteria, while useful, do not pay adequate attention to the unique characteristics of cyberattack. When an attack occurs under the current technological conditions, it may take weeks or months to determine the damage and attribute responsibility. In an era of long-range weapons and techniques, the invasion of physical borders loses its importance. In its advisory regarding nuclear weapons, the ICJ stated that the unique characteristics of nuclear weapons should be considered when interpreting rules on force and armed hostilities (ICJ, 1996). The ICJ commentary could be applied to computer network attacks. For example, the Stuxnet Worm, detected in 2010, contaminated users' computers, located files with SCAD extensions related to software and sent them to a certain server. Based on Schmitt's scale, the intensity of the 2010 cyberattack is below the threshold of armed attack. Although the effects were immediate, they were limited and were not followed by physical damage. Further, the attacks, although invasive, were not direct. Overall, in Schmitt's scale, they cannot be considered a use of force. Based on the perspectives of the other jurists, these attacks can be mostly considered as cyber turbulence rather than military attack (Waterman, 2007).

Despite the above, it could be argued that this cyberattack reached the threshold of use of force based on Schmitt's scale. A consensus has not been reached by the international community in this regard; however, the action recognized as "using force" is not required to have been necessarily committed by the armed forces of a country. For example, a cyber operation that is conducted by the armed forces can be also realized as an example of using force if performed by the intelligence organizations of a state or the contras. Concentrating on the scope and effect of an operation can be useful in determining force. Amount and effect are quantitative and qualitative factors can guide determinations. The international court of justice concluded in the Nicaragua case that providing weapon to and training militia who later took part in hostile operations against another country was a use of force. Providing malware and instructions for performing cyberattacks against another country may also be considered as using force (Waterman, 2007).

5. The Tallinn Manual and the Principle of the Prohibition on Force

Tallinn Manual is comprehensive and complete in terms of criminalization and offering of solution even with its copying of some articles from the conventions on humanitarian issues and although it has been able to criminalize the interventions made in international cyberspace based on exact studies by a group of experts (Jensen, 2016). The Tallinn Manual has succeeded in elaborating the permissibility and prohibition of each of the cyberauctions that, as believed by experts, can be considered components of international armed conflict. However, since cyberattack does not have any kinetic effect in itself it is open to question how much of an effect is the result of the intervention. For instance, if a cyber intervention leads to instability and people become angry and attack police officers and individuals get killed, an argument could be made that the cyber action is equivalent to armed attack. If a water treatment plant is subjected to a cyberattack and a group of individuals become sick or die as a result of water contamination, it would be certainly considered as an armed attack (Schmitt, 2013).

Tallinn Manual includes both the international and non-international armed conflict. In brief, unlike what is inferred about its common use, nothing has been written about cyber security. Cyber espionage, stealing of intellectual properties and a vast spectrum of the penal interventions in virtual space are considered as serious and real threats against countries, organizations and private individuals. Sufficient reaction to such measures demands national and international relations that are neglected by Tallinn Manual due to the fact that the international laws do not play their required role regarding the use of force and armed conflict. Cyber espionage, the theft of intellectual property and other virtual infractions are serious threats and demand a national and international response. International laws lack the required efficiency for confronting threats in cyberspace. The Tallinn Manual emphasizes cyber interventions in their exact sense but provides insufficient protection against many interventions (Schmitt, 2013).

6. Expansion of the Cyberattack Rules

There is no credible and legal definition of cyberwar or cyberattack in general international laws, specific treaties, customary laws or extant doctrine. However, descriptions based on technology and performance can be found. The Defense Ministry of the US has offered definitions of cyber interventions such as "using a computer for creating disruption, degeneration or destruction of the information existent in computers and computer networks or the computers and networks themselves" (ScienceDirect, 2019).

The International Committee of the Red Cross revealed its stance at least in two formal texts regarding the enforceability of international humanitarian law on cyber interventions in armed attacks. Available documentation indicates that the Committee believes that even if given military activity is not specifically systematized, that does not mean that it can be limitlessly used. From this perspective, war instruments and methods such as cyber technology have been used as weapons or systems of hostility and can thus become the subject of international humanitarian laws. While there is no legal article in the international humanitarian laws for explicit prohibition of cyber interventions, their use in armed hostilities can be accommodated within the framework of existing law. The documentation indicates that the Committee favors the characterization of uses of cyberspace which have a clear military intent as uses of force and thus subject to regulation (ICRC, 2010).

The most important issue in cyberwarfare is the confusion about the limits of existing provisions and the definition of terms such as "attack". If "attack" is held equivalent to "military intervention", attacks on cybernetworks may be regulatable. However, uncertainly continues to exist as to the extent to which cyber interventions can be regarded as attacks" (Saxon, 2013).

7. Conclusion and recommendation:

As mentioned, there are two essential viewpoints proposed as regards the interpretation of the term "force" as used in Article 2(4). The first is instrument-based and the second is target-based. Western states believe that the prohibitions outlined in Articles 2(4) and 51 of the Charter are directed at military attacks or armed violations via traditional armed technologies. The simple and clear appearance and meaning of Articles 41, 42 and 51, as well as the Introduction to the Charter, which states that "… armed force cannot be used for any other case than the public interest", support this interpretation. Hence, only interventions that have reached the limit of armed attack are viewed as clear violation of Article 2(4) of the Charter unless they have been stipulated by the Security Council or justified as legitimate defense.

Cyberattacks should be placed within the inclusion circle of the prohibition put forth by Article 2(4) so they can be regulated. Instead of underlining the type of instrument and mechanism of action, emphasis should be on the effects and instruments, as well as international laws governing the use of force by countries as a means of national policy and international laws regulating armed engagement. The guidelines deal with matters recognized as relating to laws of war, laws of armed hostility and international humanitarian law. Elements related to international law such as the responsibilities of states and maritime law are also canvassed by the manual.

The common perspective is that an action should have a physical effect such as an explosion or physical force. Ian Brolin, however, proposed a target-based approach to the determination of force. This would expand the scope of force to include the use of, for example, chemical and biological weapons which lead to destruction of life and property.

Pikte's scale and effect criterion, which emphasizes intent and the targeting of vital infrastructure, allows cyberattack to be seen as armed attack. Viewed as falling within the purview of the sevenfold criteria proposed by Michael Schmitt, a target-based approach theorist, cyberattacks can be defined as uses of force for the purpose of international law.

In order for liability for a cyberattack to be ascribed to a state, the action should be firstly determined to be a breach of an international commitment, after which it can be attributed to the state. Each of these conditions faces numerous challenges due to the undeveloped nature of discussions on the legal aspects of cyberattack in international law. The unique character of cyberspace implies that it is difficult to detect and assign responsibility for cyberattacks, which adds to the complexity of the situation.

References

Azzopardi, M. (2013). The Tallinn Manual on the international law applicable to cyber warfare: A brief introduction on its treatment of jus ad bellum norms. *ELSA Malta Law Review, 3*, 179.

Brenner, S. W., & Clarke, L. L. (2010). Civilians in cyberwarfare: Conscripts. *Vanderbilt Journal of Transnational Law, 43*, 1011.

Dinstein, Y. (2016). *The Conduct of Hostilities under the Law of International Armed Conflict.* Cambridge, England: Cambridge university press.

Geiss, R. (2010). The legal regulation of cyber-attacks in times of armed conflict. (pp. 141-145). Geneva: ICRC.

Hathaway, O. A. (2012). The law of cyber-attack. *California Law Review, 100*, 817.

ICJ. (1986). *Nicaragua v. United States of America.* The Hague, Netherlands: International Court of Justice.

ICJ. (1996). *Nuclear Weapons .* The Hague, Netherlands: International Court of Justice.

ICJ. (2003). *Oil Platforms Case.* The Hague, Netherlands: International Court of Justice.

ICJ. (2005). *Armed activities on the territory of Congo (Congo V. Oganda).* Netherlands: ICJ.

ICRC. (2010, OCTOBER 29). *Cyber warfare*. Retrieved from https://www.icrc.org/en/war-and-law/conduct-hostilities/cyber-warfare

International Committee of the Red Cross. (1989). *Rules of International Humanitarian Law and Other Rules relating to the conduct of hostilities.* Geneva: ICRC.

International Law Association. (2010). *Final report on the meaning of armed conflict in International Law.* Hague, Netherlands: International Law Association.

Jensen, E. T. (2016). The Tallinn Manual 2.0: Highlights and Insights. *Georgetown Journal of International Law*, 735.

Kelsey, J. T. (2008). Hacking into international humanitarian law: The principles of distinction and neutrality in the age of cyber warfare. *Michigan Law Review, 106*, 1429.

Oppenheim, L. (1906). *International Law, War and Neutrality, ed. Lauterpacht* (7 ed.). Edinburgh: Longmans, Green.

Roscini, M. (2014). *Cyber operations and the use of force in international law.* USA: Oxford University Press.

Saxon, D. (2013). *International humanitarian law and the changing technology of war.* Leiden, Netherlands: Martinus Nijhoff Publishers.

Schmitt, M. N. (2012). Cyberspace and international law: the penumbral mist of uncertainty. *Harvard Law Review Forum, 126*, 176.

Schmitt, M. N. (2013). *Tallinn manual on the international law applicable to cyber warfare.* Cambridge, United Kingdom: Cambridge University Press.

ScienceDirect. (2019). *computer network attack*. Retrieved from https://www.sciencedirect.com/topics/computer-science/computer-network-attack

Tsagourias, N., and Buchan, R. (2015). *Research Handbook on International Law and Cyberspace.* Cheltenham, United Kingdom: Edward Elgar Publishing.

United Nations. (1974, December 14). *Definition of Aggression*. Retrieved from United Nations,: https://unispal.un.org/UNISPAL.NSF/0/023B908017CFB94385256EF4006EBB2A

Waterman, S. (2007, June 11). *Analysis: Who cyber smacked Estonia*. Retrieved from United Press International: http:// www. space war .com/response/who-cyber- smacked- Estonia- 999.Html

Waxman, M. C. (2011). Cyber-attacks and the use of force: Back to the future of article 2 (4). *The Yale Journal of International Law, Vol. 36: 421-459, p 432, 36*, 421.

Yde, I. (2010). *The Law of Cyber Armed Conflicts: Translating Existing Norms of International Humanitarian Law into Cyber Language.* Denmark: Royal Danish Defense College. Retrieved from http://www.fak.dk/en/publications/Documents/The%20Law%20of%20Cyber%20Armed%20Conflict

A Study on Problems of Behaviour-based User Attribution in Computer Forensic Investigation

Francis Xavier Kofi Akotoye[1], Richard Ikuesan Adeyemi[2] and Hein S. Venter[1]
[1]University of Pretoria, South Africa
[2]Cyber and Network Security, School of Information Technology, Community College of Qatar, Qatar
u18349872@tuks.co.za
rikuesan@ieee.org
hein.venter@up.ac.za

DOI: 10.34190/EWS.20.117

Abstract: The origin of an attack is essential for implementing mitigating measures and pursuing legal redress for the victim or affected entity. Identifying human agents of cybercrime, termed user attribution in computer forensic investigation, is important for purposes of achieving prosecution at the law court and to deter further attacks. Attaining user attribution during digital investigation remains largely elusive due to the volatile nature of digital artefacts and the susceptibility of such artefacts to deliberate or non-deliberate manipulations. The ease with which attackers can hide their identity through employing evasive and deceptive methods, as well as strict legal requirements, further worsens the task of achieving user attribution. There is the need, therefore, to explore techniques proposed in the domain to measure their effectiveness and identify the gaps that exist and motivate further directions toward addressing user attribution problem. This work presents a systematic review of behaviour-based user attribution techniques proposed for digital forensic investigation. It highlights features adopted for recognition and evaluates their effectiveness in distinguishing between users. The methods are categorized into application dynamics, online dynamics, mouse dynamics and keystroke dynamics, which are evaluated based on recognition rate. The observation presented in this study is relevant to researchers and professionals who seek clarity of insight into the user attribution challenge. Furthermore, the findings in this study provide a baseline for the development of a robust user attribution solution that can scale through legal rigors and scrutiny.

Keyword: Attribution, User Attribution, Digital Forensics, Cyber Crime, Biometric Systems

1. Introduction

The role of digital technology in human life cannot be underestimated. It has permeated every aspect of human engagement be it social, economic, entertainment, academic, medical, or security. This trend can be attributed to the dynamic capabilities of digital systems in capturing, processing and generating meaningful human-centric results. The growth in Artificial Intelligence (AI), cloud computing and 5th Generation (5G) networks has resulted in the emerging 4th Industrial Revolution (4IR), which would further push for the integration and adoption of digital devices (Chen et al., 2017). Though this growing trend can be harnessed for positive and productive purposes, the reverse is also true for negative and/or criminal intent. Computers are used to perpetrate crimes such as information theft, fraud, espionage, denial of service attacks, hacking, cyberstalking, spamming, etc. which result in huge financial and/or material loss to individuals and businesses and the loss of confidence and trust by the victim. Cybercrime events are investigated to identify what was hacked, how it was hacked, when it was hacked and who did the hacking. This is termed as Digital Forensic Investigation (DFI).

DFI involves gathering and analysing evidence in order to build a timeline of events, and/or make a claim or otherwise of criminal intent and action. In traditional forensic investigation, the crime scene is carefully scouted, and pieces of potential evidence documented and collected for analysis. The result of the analysis is evaluated and interpreted for causal and/or attributional purposes. This procedure can be extended to cybercrime by conducting "*post-mortem*" assessment of cyber incidents in order to identify loopholes in the system, how the attacks were carried out, the level of damage caused and the origin of the attack.

One important aspects of digital forensic investigation is to identify the perpetrator (usually, a human actor) of cybercrime, referred to as user attribution. Attackers often adopt evasive and deceptive mechanisms or simply deny their involvement in a crime to escape blame (Boebert, 2010), making it difficult for investigators to establish a link between a crime and the probable perpetrator. Strict legal requirements coupled with jurisdictional bureaucracy present further challenges for reliable user attribution. It is important, therefore, to examine the field and ascertain what has been done and evaluate the effectiveness of the strategies proposed in addressing user attribution problem.

The subsequent sections of this work are outlined as follows: section 2 presents review methodology while section 3 offers a background to digital forensics. Definition of forensic attribution and how it relates to digital forensic attribution, followed by the categorization of the various proposed user attribution techniques into four main methods are captured in section 4. In section 5, a discussion of the methods is given and the paper concludes in section 6.

2. Methodology

To review user attribution methods proposed for computer forensics investigation, the following query terms ("user attribution", "user attribution" AND "digital forensics", "forensic user attribution", "user identification" "user identification" AND "digital forensics", "forensic user identification") were executed on the following publication repositories: Scopus, IEEE, ACM, Springer link, Taylor & Francis and DFRWS conference proceedings. The results were further filtered by publication year defined by the period year 2010 to the present and the mention of the term "user attribution" or "user identification" in the publication's abstract. The methods are categorized and discussed in section 4.

3. Background to Digital Forensics

Digital Forensics (DF) involves the identification, retrieval and investigation of incriminating materials found on digital devices. Investigating cyber incidents require adherence to clear and well-defined procedures recognized as DFI process model (Cohen, 2010a, Kohn et al., 2013, Mir et al., 2016, Palmer, 2001, Valjarevic and Venter, 2012). Although a widely accepted DFI process model remains elusive, one common theme that runs through the various works is that DFI should be established on well-proven scientific principles (Olivier, 2016). Generally, DFI processes are condensed into the stages of identifying, recovering, analysing and reporting of the digital evidence acquired (Mir et al., 2016).

Investigation initiates with securing and preserving the crime scene (digital or physical) by applying for and securing legal warrants that empower investigators to gain control over devices involved in an incident. This is necessary to ensure that potential digital evidence (PDE) is not contaminated or erased either by deliberate or non-deliberate activity. The next step requires PDEs to be duly documented, stored and retrieved in forensically sound manner. It is necessary to conduct this in a manner which does not introduce unwanted artefacts into the original stream of bits. Storage media for backup storage need to be carefully sanitized before bit-by-bit imaging of captured potential evidence. The norm is to use write blockers during data imaging to safeguard against contamination. Establishing a chain of custody can further enhance the integrity of the imaged data (Hauger and Olivier, 2015). Hashing schemes such as Message Digest 5 (MD5) or Secure Hash Algorithm (SHA) (Schmitt and Jordaan, 2013, Roussev, 2009, Kumar et al., 2012) are recommended to guarantee that data are exact copies of the original. Collected artefacts are analysed and outcomes interpreted and presented as evidence for inculpating or exculpating purposes.

4. State of the Art in User Attribution

This section defines forensic attribution and outlines techniques adopted by forensic investigators in linking perpetrators to incidents. The section further discusses uniqueness of digital domain and how this influences forensic investigation and concludes with a review of methods proposed for user attribution.

4.1 Forensic Attribution

Attribution, defined as "the action of regarding something as being caused by a person or thing" (Dictionary and Idioms, 1989). In a cause-and-effect scenario when an event occurs, the necessary information about the event is gathered and studied in order to identify and explain the cause of the event (McLeod, 2019, Fiske and Taylor, 2013). Such information may be defined by the internal characteristics of the causal agent and/or external factors that impact the agent (Heider, 2013, McLeod, 2019). For instance, a person's internal characteristics can be used to attribute the person's behaviour based on their personality, intentions or beliefs. Conversely, when behaviour is based on environmental influences beyond the person's direct control, external characteristics are considered. This distinction impacts the evidence gathering process during investigation. Investigators normally comb for artefacts that could be used to tie a person/thing to the crime scene such as fingerprints, hair, blood or body fluid (for DNA sequencing), teeth marks, bullet marks or gunpowder residue (for ballistic analyses), shoe-sole prints, knife marks, etc. which are utilized as internal characteristics to differentiate between persons or objects. External markers such as residues from soil, mud, paint, fabric, or metal shavings are used to place agents at crime scenes. In addressing attribution, two questions need to be answered – what needs to be attributed? (Shamsi et al., 2016) and why the need for attribution? (Tran, 2018, Tsagourias, 2012, Shamsi et al., 2016). The first question helps in identifying what artefacts are necessary to connect the agent to the crime, which informs the attributional technique to use. The latter question addresses the rationale for undertaking attribution, i.e. how would the outcome of the attribution be utilized. Successful attribution discloses the identity of the agent thus probably shedding more light on their method of operation. This outcome could be captured in a repository for historical purposes (e.g. criminal database), and/or to serve as guidelines or early warning systems against similar occurrences. Attribution may lead to institution of legal or retaliatory moves against the identified agent (Tran, 2018, Tsagourias, 2012, Shamsi et al., 2016) or to deter other potential saboteurs (Nicholson, 2015). These questions are significant in determining how the process of attribution would proceed; if court charges must be preferred, then human attribution must be achieved and the artefacts that were used to arrive at such conclusions must satisfy the level of proof required for such purpose. However, if the intention is to mitigate the effect of an incident, then the reliability of the mechanics adopted for attributional purpose may not be as stringent as required for legal prosecution.

4.2 Digital Forensic Attribution

In digital domain the question of what needs to be attributed has been concisely categorized by Hauger and Olivier (2018) into two realms – digital and physical. Digital realm encapsulates digital devices directly or remotely involved in an incident (Cohen, 2010b). Digital devices per their operational mechanisms tend to embellish created objects with certain "footprints" that when examined identifies the device or its class (or model). Shamsi et al. (2016) argued that the boundaries of digital realm attribution encompass what they termed "*cyberweapon*", which identifies the code or application responsible for invoking the cyber-incident. Implying that the result of attribution in digital realm may point to a physical device or a virtual one (represented in a code form). Early attempts in printer attribution involved correlating physical properties of ink and paper to link printer with document (LaPorte, 2015). Other considerations ranged from exploiting printer defects (Chiang et al., 2010, Shang and Kong, 2015) or geometric distortions of printed matter (Bulan et al., 2009, Ju et al., 2012, Wu et al., 2009) as well as textual pattern formulations (Ferreira et al., 2015, Mikkilineni et al., 2011, Navarro et al., 2018). In camera attribution researchers have exploited inherent camera image processing techniques, camera component and model features (Orozco et al., 2013) as well as camera metadata, image environment, and fine-grained image parameters for device identification (Boutell and Luo, 2005, Romero et al., 2008). Appreciable successes have been achieved in categorizing cameras based on sensor pattern noise which highlights differences among cameras from different vendors and even models from the same manufacturer (Costa et al., 2012, Lukáš et al., 2006, Van Lanh et al., 2007). In network forensics evaluation of packet traceback and logging have been exploited for attributional effects (Shamsi et al., 2016, Nicholson et al., 2013).

Physical attribution attempts to uncover persons or locality of an attack. Physical attribution is challenging since the real world cannot be described by features that define cyberspace (Hauger and Olivier, 2015). Physical attribution requires artefacts that can associate humans to activities that occurred in cyberspace *i.e.* digital features that correlate with human activities. This problem is further exacerbated by the fact that an actor may have the capacity to delete, modify or plant artefacts that may redirect attention from them to some innocent party. In other cases, political or ideological differences especially at national levels present challenges to attaining effective physical attribution. Human attribution in digital domain exists mainly as author attribution or user attribution. The distinction lies in the artefacts examined. Author attribution relies chiefly on textual data by evaluating text stylometric properties (Rudman, 1997, Stamatatos, 2009). Textual data such as books, news articles, emails, short messages, etc. are evaluated to discern the statistical behaviour of the grammatical, lexical, syntactic, structural, stylistic and semantic qualities of such materials. Though these features possess reliable discriminative properties (Rudman, 1997, Stamatatos, 2009), their limitation to text-domain confines their application to code authorship attribution, hate speech attribution, etc. User attribution, however, is encountered in a variety of digital forensic domains ranging from network, database, mobile devices, software, computer forensics, etc. The purpose of this discussion is to review user attribution methods that exist for computer forensic investigation.

4.3 User Attribution Methods

In mainstream forensic investigation crime scenes are normally laden with physical and tangible artefacts deposited or disturbed by activities of the perpetrator. Such artefacts when lifted and examined by investigators potentially recreate the chain-of-events and/or identification of agent(s) of the act. In some instances objects from the crime environment are deposited on the perpetrator serving as plausible proof of their involvement or presence at the scene of the incident. Digital forensics operates on a similar premise since users interact with computer systems to either manipulate the system to execute a task or to retrieve information hence leaving a trail of PDEs.

4.3.1 Application Dynamics (GUI, Command line)

Khosmood et al. (2014) argued that user attribution was possible from analysis of command line histories of a Unix system user. Their technique classified 100 lines of users' command history based on command name, semantic labelling, word and character N-grams; and created a machine learning model which produced performance rate of 93%. El Masri et al. (2014) also proposed a GUI-based user identification technique anchored on MS Word command dynamics. They evaluated MS Word sessions of 24 users and reported recognition rate of 95.43%. The application specificity of their technique limits its generalizability, furthermore the dataset employed (spanning 1997 and 1998) may not be representative of current advances in application engineering (they fail to state which version of MS Word was tested) and command behaviour. In (Bailey et al., 2014) features from keyboard, mouse and GUI interactions were combined for user identification. The study examined duration and latency parameters of keystroke, mouse and GUI usage features to demonstrate the feasibility of multi-modal behavioural biometric for user attribution. Ingo Deutschmann and colleagues (2013) combined users' keystroke, mouse movements, application usage dynamics with system footprints to continuously identify system users. They proposed an evaluation metric based on similarity distance between measured behaviour and expected behaviour, which was projected onto a probabilistic trust model based on established threshold score. The authors claimed the technique addresses temporal dimensionality inherent in continuous authentication of users. The strategy exhibited stable performance for sizeable user population but failed to scale when population increased. The study reported high-performance rating for keystroke dynamics as compared to mouse or system footprints. The study explored various modalities in a unimodal context though a multimodal integration may have posted more robust performance throughput. Probably, the less restrictive performance requirement of binary-class classification as compared to the tedious task of identifying a user from a group of known or unknown subjects (multi-class classification) justifies their approach.

4.3.2 Online Dynamics (Web)

Studies show that human behaviour can be culled from online activities thus creating a structure for identity generation or confirmation. In (Ikuesan et al., 2017) the study explored the probability of using online behavioural signatures to profile users based on Polychronic-Monochronic Tendency Scale (PMTS). Extracted features were analysed with six supervised machine learning algorithms, which exhibited performance of over 80% accuracy especially in the case of Logistic Model Tree technique. Though experimental results were promising the feasibility of the system hinged on server-side feature harvesting which was not scalable with system deployment, especially in cross-jurisdictional context. Adeyemi and colleagues (2014) proposed a user identification framework based on extraction and classification of users' psychographic data extracted from online activities. The first phase of the framework segment users into clusters based on internet web surfing patterns. In the second phase psychosocial markers (personality trait, thinking style, and temporal style) of users' activities were subsequently adopted to segment users into unique classes. Phases I and II were fed into the analytical evaluator in phase III to generate (dis)similarity measures of the candidates. Extracted features exhibited statistical relevance however, the feature space could be further expanded. In (Adeyemi et al., 2016b) the study used Logistic Model Tree (LMT) with bagging technique to probabilistically categorize online users into respective groups based on thinking style signatures. The study indicated that users' thinking style were reflective in web request characteristics and page visitation dynamics. Though the study demonstrated acceptable results, individualizing users based on the technique remains an open challenge. In (Adeyemi et al., 2016a) the study showed personality traits extractable from network activities or online request patterns to be definitive of individual dynamism. Six machine learning techniques including Support Vector Machine were adopted for 3-class classification of High-, Moderate- and Low-class personality traits. Performance review pointed to a probabilistic distinction of users' personality traits, though extending the technique to identify individual users is non-trivial. The online context of individualizing features may present a weak link especially if the identified subject's machine cannot be explicitly tied to the human perpetrator. In (Gresty et al., 2016), Jaccard Similarity Coefficient was used to estimate the similarity or otherwise of users in a multi-user environment based on internet history session-to-session artefacts. The technique detected one-time and/or repetitive event patterns that were user attributable. Though accuracy rate of 90% was reported, the study was limited by the removal of "spoiler" components which in fact denoted real-life instances in internet history evaluation.

4.3.3 Mouse Dynamics

The popularity of GUI shell systems coupled with ease of use and intuition makes the mouse a common user agent for computer interaction. The interactive domain defined by mouse usage creates patterns which are descriptive of a person using the mouse at any given point in time. Details of mouse movements (drag, scroll, XY orientation) and mouse events (click, double click, clicking speed, etc.) can be measured to discriminate between persons. Shen et al. (2012) were able to verify computer users by extracting user-marking identifiers mined from procedural mouse features extracted from basic mouse operations. The scheme performs well in user authentication context but fails to address user identification challenges. Feher et al. (2012) proposed the use of independent mouse action for user characterization. Their technique reduced verification time drastically whilst enhancing the accuracy of the system by aggregating lower-level mouse features into high-level features rich in discriminative qualities. Mouse dynamics was used to segment users into age and gender classes (Kratky and Chuda, 2016) by leveraging mouse movement velocity, position and path behaviour from online interactions to classify users. Though the proposed model failed to achieve reliable identification performance, it afforded age and gender segregation. Ernsberger et al. (2017) explored mouse activities on various online news sites to model and visualize users' mouse signatures. The study, though, failed to establish reliable recognition performance (~80% accuracy) probably due to relatively small sample size and limited experimental duration however; it underscored the potential of mouse dynamics for forensic user attribution. It further argued that generating visual cues from users' mouse path activity supports user activity extrapolation from event reconstruction, while encouraging further exploration of mouse features enriched with discriminatory power to address the challenge of psychological imbalance that generates variability in mouse dynamics measurement.

4.3.4 Keystroke Dynamics

This technique measures users keystroke behaviour during data input such as keypress pattern and deflection, speed and sound pattern that emanates from keypresses. These characteristics are exploitable to create models that categorize users. Tsimperidis and Katos (2013) investigated text origins – i.e. does the text originate from a desktop or a laptop? – by exploiting the temporal features of digrams. Desktop-based keystroke latency exhibited different characteristics from laptop-based latency due to key depth deviations between the two types of keyboards. Latency variations that were statistically improved into scoring system per keyboard type and evaluation resulted in converged accuracy rate of 77.11%. The study, although exposed profiling capacity of keystroke artefacts necessary for forensic investigation, its generalizability was limited by the small number of subjects and the use of fixed text as reference model which deviates from real-life scenario. Besides, typing Greek text on a standard keyboard may capture non-keystroke discrepancies that may influence model accuracy. In this study (Roth et al., 2014) keystroke acoustics was proposed as an identification technique that uses sounds generated from striking pattern of individual keys to distinguish users. The sounds sampled into "virtual letters" and aggregated as "virtual alphabet" described the acoustic signature of individuals at the keyboard. An ingenious proposal, however, the feasibility of this technique for open-set identification is unaddressed. Besides, the technique may be highly susceptible to distortion from the capture of extraneous sound during feature registration therefore inhibiting the reliability of the scheme for attributional purposes. The ever-changing state of user disposition (Mohlala et al., 2017) may also pose challenges to the stability of acoustic features that are captured. In fact, Mohlala et al. (2017) argued that for keystroke dynamics to be forensically effective, it should be proactively captured in a forensic readiness context to provide for adequate capture of features. Shute et al. (2017) proposed a technique for recognizing the handedness of a user by partitioning the keyboard into 6 zones of 2 regions mapped to the 2 handed typing dynamics of a normal keyboard user. Temporal features of keystroke events occurring in each zone were registered and evaluated with machine learning techniques. The scheme delivered accuracy rates of 94.5% on handedness detection. However, due to the imbalance of left-hand/right-hand proportion (with global average for western population at ~10%), the study adopted Kappa (0.79) statistics to contextualize model performance. Shen et al. (2016) also explored the issue of handedness using timing intervals of keystroke through patterns of typing fixed alphanumeric text of 11 characters for classification. Candidate handedness were measurable 87.75% of the time, though the size of the reference text and the experimental environment did not necessarily reflect real situations for forensic quality.

Generally, keystroke characteristics present great potential for computer forensic user attribution due to its ubiquity and transparent data capture. Keystrokes are easily captured in both command- and graphical-based environments in contrast to mouse dynamics. Feature dimensions of keystroke are more extensive than those presented by mouse dynamics; hence they possess greater discriminatory characteristics that can be leveraged to establish user distinction. However, keystrokes are influenced by certain behavioural tendencies that limit their effectiveness. Users' emotional state during typing may influence normal typing behaviour (Cui et al., 2018, Mohlala et al., 2017). Therefore introducing within-class variability which negatively affects keystroke accuracy. Other factors such as physiological and psychological imbalances of users and differences in keyboards and their layout (Zhong and Deng, 2015) also account for deviations in the normal keystroke characteristics of users.

5. Discussion

Differences in applications regarding user interface and feature versioning present challenges for adopting application dynamics for user attribution. Different application features and user agents demand that the forensic investigator be knowledgeable in a variety of applications in order to extract data from low-level implementation for user identity recreation. Rate of technological advancement however, renders it infeasible for investigators to possess such extensive knowledge. Perpetrators tend to leverage zero-day vulnerabilities of applications as well as other obfuscating mechanisms to hide their trail. Investigators often uncover such techniques after the attack and in most cases are unlikely to encounter them again since vendors do deliver patches and updates to plug the loophole.

Attribution based on online behaviour also exhibits inherent shortcomings which decrease the feasibility of such techniques. The open architecture of the internet, which drives most online interactions, encourages anonymization or encryption of online activities to safeguard user privacy hence hardening the work of the forensic investigator. Users’ online activities are mostly kept server-side, which generates questions of jurisdictional control during investigation hence limiting the level of user attribution achievable. The lack of control over user data residing on the server may raise questions of integrity and privacy that could be raised by a user under investigation (Michalsons, 2018), limiting the value of the attribution. Advances in anti-forensic techniques such as cryptography, steganography, trail obfuscation (Kessler, 2007) further hinder the work of the forensic investigator in achieving attribution. Modern web browsers are laden with anonymity-ensuring features such as “incognito” browsing that wipe off user information (password, cookies, history, etc) necessary for linking users’ online activity.

The artefacts space for attribution in computer forensics is quite limited and volatile. However, the characteristics that underpin user interactions on the computer are grounded in the behaviour of the user. The argument for employing behavioural biometrics for user attribution is now gaining traction, though the digital forensic requirement of mouse/keystroke dynamics is more stringent (Ikuesan and Venter, 2019). Unlike authentication challenges, which are mostly binary class problems user attribution is a multi-class problem (Ikuesan and Venter, 2019) involving users present or absent from the system under review. The current set of mouse and keystroke features available for attributional purposes may not be rich enough for individuality assessment resulting in inefficiencies of behaviour-based user attribution strategies. There is the need to further explore the feature boundaries to uncover features that possess the probability of delivering reliable user attribution outcomes, specifically in the following trajectory:

1. The need for a methodology capable of identifying and capturing significant features rather than reliance on artificial mining techniques, which are susceptible to behavioural noise.
2. The need to explore other features that may result from fusing keystroke/mouse dynamics with physiological modalities.
3. The need for effective noise exfiltration techniques in order to clean noise-induced traits.
4. The need for feature-rich data sets to aid research and serve as a testing benchmark.

6. Conclusion

One main reason for digital forensic investigation is to uncover the perpetrator of a cybercrime. In a cyber incident, a perpetrator may be a device (digital/system attribution), a human or the locality where the operation originated (physical attribution). Conducting digital and physical attribution involves identifying and collecting unique artefacts that would assist in extrapolating identity. User attribution points to the person who was operating the computer (desktop, laptop) at the time of an incident. The authors explored the various existing methods for achieving user attribution in computer forensic investigation and how effective the methods are in delivering reliable outcomes. Evaluation of recognition outcomes points to the fact that there is need for further feature exploration that meets user attributional requirements for digital forensic investigation.

Reference

ADEYEMI, I. R., ABD RAZAK, S. & SALLEH, M. 2016a. Understanding Online Behavior: Exploring the Probability of Online Personality Trait Using Supervised Machine-Learning Approach. *Frontiers in ICT,* 3**,** 8.

ADEYEMI, I. R., RAZAK, S. A. & SALLEH, M. A psychographic framework for online user identification. 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 2014. IEEE, 198-203.

ADEYEMI, I. R., RAZAK, S. A. & SALLEH, M. Individual difference for HCI systems: examining the probability of thinking style signature in online interaction. 2016 4th International Conference on User Science and Engineering (i-USEr), 2016b. IEEE, 51-56.

BAILEY, K. O., OKOLICA, J. S. & PETERSON, G. L. 2014. User identification and authentication using multi-modal behavioral biometrics. *Computers & Security,* 43**,** 77-89.

BOEBERT, W. E. A survey of challenges in attribution. Proceedings of a workshop on Deterring CyberAttacks, 2010. 41-54.

BOUTELL, M. & LUO, J. 2005. Beyond pixels: Exploiting camera metadata for photo classification. *Pattern recognition,* 38**,** 935-946.

BULAN, O., MAO, J. & SHARMA, G. Geometric distortion signatures for printer identification. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 19-24 April 2009 2009. 1401-1404.

CHEN, B., WAN, J., SHU, L., LI, P., MUKHERJEE, M. & YIN, B. 2017. Smart factory of industry 4.0: Key technologies, application case, and challenges. *IEEE Access,* 6**,** 6505-6519.

CHIANG, P.-J., KHANNA, N., MIKKILINENI, A. K., SEGOVIA, M. V. O., ALLEBACH, J. P., CHIU, G. T. & DELP, E. J. 2010. Printer and scanner forensics: models and methods. *Intelligent Multimedia Analysis for Security Applications.* Springer.

COHEN, F. Toward a science of digital forensic evidence examination. IFIP International Conference on Digital Forensics, 2010a. Springer, 17-35.

COHEN, F. B. Attribution of messages to sources in digital forensics cases. 2010 43rd Hawaii International Conference on System Sciences, 2010b. IEEE, 1-10.

COSTA, F. D. O., ECKMANN, M., SCHEIRER, W. J. & ROCHA, A. Open set source camera attribution. 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images, 2012. IEEE, 71-78.

CUI, S.-X., EKWONNAH, O. & WANG, Y.-H. 2018. The Analysis of Emotions over Keystroke Dynamics. *DEStech Transactions on Computer Science and Engineering*.

DEUTSCHMANN, I., NORDSTRÖM, P. & NILSSON, L. 2013. Continuous authentication using behavioral biometrics. *IT Professional,* 15**,** 12-15.

DICTIONARY, O. E. & IDIOMS, E. 1989. Oxford References Online. Oxford: Oxford University Press.

EL MASRI, A., WECHSLER, H., LIKARISH, P. & KANG, B. B. Identifying users with application-specific command streams. 2014 Twelfth Annual International Conference on Privacy, Security and Trust, 2014. IEEE, 232-238.

ERNSBERGER, D., IKUESAN, R. A., VENTER, S. H. & ZUGENMAIER, A. A Web-Based Mouse Dynamics Visualization Tool for User Attribution in Digital Forensic Readiness. International Conference on Digital Forensics and Cyber Crime, 2017. Springer, 64-79.

FEHER, C., ELOVICI, Y., MOSKOVITCH, R., ROKACH, L. & SCHCLAR, A. 2012. User identity verification via mouse dynamics. *Information Sciences,* 201**,** 19-36.

FERREIRA, A., NAVARRO, L. C., PINHEIRO, G., DOS SANTOS, J. A. & ROCHA, A. 2015. Laser printer attribution: Exploring new features and beyond. *Forensic science international,* 247**,** 105-125.

FISKE, S. T. & TAYLOR, S. E. 2013. *Social cognition: From brains to culture*, Sage.

GRESTY, D. W., GAN, D., LOUKAS, G. & IEROTHEOU, C. 2016. Facilitating forensic examinations of multi-user computer environments through session-to-session analysis of Internet history. *Digital Investigation,* 16**,** S124-S133.

HAUGER, W. & OLIVIER, M. Determining trigger involvement during Forensic Attribution in Databases. IFIP International Conference on Digital Forensics, 2015. Springer, 163-177.

HAUGER, W. & OLIVIER, M. S. 2018. NoSQL databases: forensic attribution implications. *SAIEE Africa Research Journal,* 109**,** 119-132.

HEIDER, F. 2013. *The psychology of interpersonal relations*, Psychology Press.

IKUESAN, A. R., RAZAK, S. A., VENTER, H. S. & SALLEH, M. 2017. Polychronicity tendency-based online behavioral signature. *International Journal of Machine Learning and Cybernetics***,** 1-16.

IKUESAN, A. R. & VENTER, H. S. 2019. Digital behavioral-fingerprint for user attribution in digital forensics: Are we there yet? *Digital Investigation,* 30**,** 73-89.

JU, Y., SAXENA, D., KASHTI, T., KELLA, D., SHAKED, D., FISCHER, M., ULICHNEY, R. & ALLEBACH, J. P. Modeling large-area influence in digital halftoning for electrophotographic printers. Color Imaging XVII: Displaying, Processing, Hardcopy, and Applications, 2012. International Society for Optics and Photonics, 82920X.

KESSLER, G. C. 2007. Anti-forensics and the digital investigator.

KHOSMOOD, F., NICO, P. L. & WOOLERY, J. User identification through command history analysis. 2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS), 2014. IEEE, 1-7.

KOHN, M. D., ELOFF, M. M. & ELOFF, J. H. 2013. Integrated digital forensic process model. *Computers & Security,* 38**,** 103-115.

KRATKY, P. & CHUDA, D. Estimating gender and age of web page visitors from the way they use their mouse. Proceedings of the 25th International Conference Companion on World Wide Web, 2016. International World Wide Web Conferences Steering Committee, 61-62.

KUMAR, K., SOFAT, S., JAIN, S. & AGGARWAL, N. 2012. SIGNIFICANCE of hash value generation in digital forensic: A case study. *International Journal of Engineering Research and Development,* 2**,** 64-70.

LAPORTE, G. M. 2015. Chemical analysis for the scientific examination of questioned documents. *Forensic Chemistry: Fundamentals and Applications***,** 318-353.

LUKÁŠ, J., FRIDRICH, J. & GOLJAN, M. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security,* 1**,** 205-214.

MCLEOD, S. A. 2019. *Attribution theory* [Online]. Available: https://www.simplypsychology.org/attribution-theory.html [Accessed September 25 2019].

MICHALSONS. 2018. *Data privacy or data protection in South Africa* [Online]. Available: https://www.michalsons.com/blog/data-privacy-in-south-africa/150#targetText=In%20terms%20of%20South%20African,most%20probably%20be%20fairly%20difficult.&targetText=bodily%20privacy%2C,the%20privacy%20of%20communications%2C%20and [Accessed September 28 2019].

MIKKILINENI, A. K., KHANNA, N. & DELP, E. J. Forensic printer detection using intrinsic signatures. Media Watermarking, Security, and Forensics III, 2011. International Society for Optics and Photonics, 78800R.

MIR, S. S., SHOAIB, U. & SARFRAZ, M. S. 2016. Analysis of digital forensic investigation models. *International Journal of Computer Science and Information Security,* 14**,** 292.

MOHLALA, M., IKUESAN, A. R. & VENTER, H. S. User attribution based on keystroke dynamics in digital forensic readiness process. 2017 IEEE Conference on Application, Information and Network Security (AINS), 2017. IEEE, 124-129.

NAVARRO, L. C., NAVARRO, A. K., ROCHA, A. & DAHAB, R. 2018. Connecting the dots: Toward accountable machine-learning printer attribution methods. *Journal of Visual Communication and Image Representation,* 53**,** 257-272.

NICHOLSON, A. 2015. Wide spectrum attribution: Using deception for attribution intelligence in cyber attacks.

NICHOLSON, A., JANICKE, H. & WATSON, T. An Initial Investigation into Attribution in SCADA Systems. ICS-CSR, 2013.

OLIVIER, M. 2016. Digital forensic science: A manifesto. *South African Computer Journal,* 28**,** 46-49.

OROZCO, A. S., GONZÁLEZ, D. A., CORRIPIO, J. R., VILLALBA, L. G. & HERNANDEZ-CASTRO, J. Techniques for source camera identification. Proceedings of the 6th international conference on information technology, 2013. 1-9.

PALMER, G. A road map for digital forensic research. First Digital Forensic Research Workshop, Utica, New York, 2001. 27-30.

ROMERO, N. L., CHORNET, V. V. G., COBOS, J. S., CAROT, A. A. S., CENTELLAS, F. C. & MENDEZ, M. C. 2008. Recovery of descriptive information in images from digital libraries by means of EXIF metadata. *Library Hi Tech*.

ROTH, J., LIU, X., ROSS, A. & METAXAS, D. 2014. Investigating the discriminative power of keystroke sound. *IEEE Transactions on Information Forensics and Security,* 10**,** 333-345.

ROUSSEV, V. 2009. Hashing and data fingerprinting in digital forensics. *IEEE Security & Privacy,* 7**,** 49-55.

RUDMAN, J. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities,* 31**,** 351-365.

SCHMITT, V. & JORDAAN, J. 2013. Establishing the validity of MD5 and SHA-1 hashing in digital forensic practice in light of recent research demonstrating cryptographic weaknesses in these algorithms. *International Journal of Computer Applications,* 68.

SHAMSI, J. A., ZEADALLY, S., SHEIKH, F. & FLOWERS, A. 2016. Attribution in cyberspace: techniques and legal implications. *Security and Communication Networks,* 9**,** 2886-2900.

SHANG, S. & KONG, X. 2015. Printer and Scanner Forensics. *Handbook of Digital Forensics of Multimedia Data and Devices***,** 375-410.

SHEN, C., CAI, Z., GUAN, X., DU, Y. & MAXION, R. A. 2012. User authentication through mouse dynamics. *IEEE Transactions on Information Forensics and Security,* 8**,** 16-30.

SHEN, C., XU, H., WANG, H. & GUAN, X. Handedness recognition through keystroke-typing behavior in computer forensics analysis. 2016 IEEE Trustcom/BigDataSE/ISPA, 2016. IEEE, 1054-1060.

SHUTE, S., KO, R. K. & CHAISIRI, S. Attribution Using Keyboard Row Based Behavioural Biometrics for Handedness Recognition. 2017 IEEE Trustcom/BigDataSE/ICESS, 2017. IEEE, 1131-1138.

STAMATATOS, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology,* 60**,** 538-556.

TRAN, D. 2018. The Law of Attribution: Rules for Attribution the Source of a Cyber-Attack. *Yale JL & Tech.,* 20**,** 376.

TSAGOURIAS, N. 2012. Cyber attacks, self-defence and the problem of attribution. *Journal of Conflict and Security Law,* 17**,** 229-244.

TSIMPERIDIS, I. & KATOS, V. Keystroke forensics: are you typing on a desktop or a laptop? Proceedings of the 6th Balkan Conference in Informatics, 2013. ACM, 89-94.

VALJAREVIC, A. & VENTER, H. S. Harmonised digital forensic investigation process model. 2012 Information Security for South Africa, 2012. IEEE, 1-10.

VAN LANH, T., CHONG, K.-S., EMMANUEL, S. & KANKANHALLI, M. S. A survey on digital camera image forensic methods. 2007 IEEE international conference on multimedia and expo, 2007. IEEE, 16-19.

WU, Y., KONG, X., YOU, X. G. & GUO, Y. Printer forensics based on page document's geometric distortion. 2009 16th IEEE International Conference on Image Processing (ICIP), 2009. IEEE, 2909-2912.

ZHONG, Y. & DENG, Y. 2015. A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics***,** 1-22.

Defining the Workforce and Training Array for the Cyber Risk Management and Cyber Resilience Methodology of an Army

Lisa Davidson
UNSW Canberra Cyber - University of New South Wales, Canberra, Australia
lisadavidson73@hotmail.com

DOI: 10.34190/EWS.20.114

Abstract: As 'middle-power countries' mature National Cyber Security policies and practices, many have identified a requirement for a military option for governments to either assist or lead cybersecurity responses. For a military force, this could place their operating methodology in a diarchy; the first being the protection and resilience of internal capability networks, the second having the capability to integrate into external networks, including critical infrastructure to defend national interests. Gaining this ability as an Army is not simply a role for the cyber specialist. For this to occur, organisational cyber resilience responses and cyber risk management practices must be established well before a cyber-incident. To meet such a national strategy, an Army must design, educate and train those forces that will undertake these roles both internally and externally to the organisation. Compared to civil industries, this could be akin to being a cybersecurity consumer and a cybersecurity service provider simultaneously; both requiring different business operating models ranging from Human-in-the-Loop (HITL) to Industrial Control System (IDS) security operations against a broad and varying threat array. This paper outlines research to define the nature of Cyber Risk Management and Cyber-Resilience for a Middle-Power Army, noting its likely internal and external requirements. It discusses those intrinsic attributes required for Cyber Risk Management and Cyber Resilience to inform workforce design and training. Finally, this paper proposes a workforce and skilling array design that is required to meet the various methodologies throughout Cyber Risk Management and Cyber-Resilience life-cycles for further experimentation.

Keywords: Cybersecurity, Resilience, Risk Management, Military, Training, Workforce Design

1. Introduction

Understanding the impact of a cyber-vulnerable network in an organisation can inform the system owners of a risk profile and derives a risk management strategy to be applied to mitigate cyber-attacks. For large-complex networks comprising of multiple platforms using the technology derived from multiple sources, the provenance and coding of each piece is difficult to risk mitigate without complexity (U.S. Government Accountability Office, 2018). This type of network architecture conflates an inevitable reality that the network will be compromised. Where risk management fails, the network must be resilient in design and in use so as to be adaptable and recoverable. Cyber Risk Management and Cyber Resilience strategies have a symbiotic relationship that collectively enable system function. Understanding cyber risk, building cyber-resilient network architectures and obtaining recovery after a cyber-attack takes strategy, training and practice (Christensen, 2017).

The gaining of this mastery is not simply a role for the cyber specialist (Schreider, 2017), as evident in for example cybersecurity in the health sector (Argaw, Bempong, Eshaya-Chauvin, & Flahault, 2019; Sher, Talley, Yang, & Kuo, 2017; Zafar, 2016). Further, both resilience and risk management practices must be established well before a cyber-incident (Christensen, 2017; S Fowler, Sweetman, Ravindran, Joiner, & Sitnikova, 2017). It needs to be considered and enabled by other requirements such as network design, equipment procurement process, governance and maintenance schedules and an understanding of the impact to the organisation whilst degradation is ongoing (Alberts, Haller, Wallen, & Woody, 2017; Brennan, Joiner, & Sitnikova, 2019).

For the Australian Army and wider Australian Defence Force, Cyber Risk Management and Resilience is more than the internal response to a cyber-related incident. It has two roles to undertake for Cyber Risk Management and Cyber-Resilience. The first is ensuring the continuity of their internal operating networks, the second is the possibility of providing resilience and risk reduction measures in support of sustaining national power in responses both nationally and internationally. This is as a result of the Australian military being an arm of government and enabler to national power. To meet a national strategy and guided by policy, an Army must design, educate and train those forces that will undertake these roles both internally and externally to the organisation. Compared to civil industry this could be akin to being a Cybersecurity consumer and a Cybersecurity service provider simultaneously; both requiring different business operating models ranging from user to industrial control system security against a broad and varying threat environment.

This essay will define the nature of Cyber Risk Management and Cyber Resilience for a Middle-Power Army, noting its likely internal and external requirements. It will discuss those intrinsic attributes required for Cyber Risk Management and Cyber Resilience to inform workforce design and training. Finally, it will propose a workforce and skilling array that is required to meet the various methodologies throughout Cyber Risk Management and Cyber Resilience life-cycles for further experimentation.

This paper is part of wider PhD research into developing a Cybersecurity Workforce for a Middle Power Army.

2. Cyber Risk Management and Cyber-Resilience: Definitions

This chapter will utilise the definitions below to bound Cyber Risk Management and Cyber-Resilience for the research. These definitions acknowledge the requirement of cyber risk management; however, caveat the requirement for an organisation to successfully operate in a reduced capacity where risk is realised.

Cyber Risk Management. In 2015 (Refsdal, Solhaug, & Stolen, 2015) define Cyber-risk as 'a risk that is caused by a Cyber-threat'. They further refine Cyber-risk Management as the treatment of these risks, defining two broad groups of likely risk being malicious and non-malicious cyber risks. The refining of Cyber-risk types assists in the approach and organisation posture in the management of the risk. This definition nests with approaches of risk management defined by NIST and those such as table-topping in the treatment of risk(Christensen, 2017).

Cyber Resilience. Is defined as 'the ability to continuously deliver the intended outcome despite adverse cyber events' (Björck, Henkel, Stirna, & Zdravkovic, 2015). Resilience is the system to countering those events that cannot be addressed through risk management. In a military context, Cyber Resilience is defined as the ability of a military organisation to (a) avoid complete destruction in face of a technological surprise, and then to (b) recover and return to effective performance against the adversary' (S. Fowler & Sitnikova, 2019; Kott, 2018).

3. Symbiosis of Risk Management and Resilience

Risk Management and Resilience are intrinsically linked and provide the ability for a system to be robust. Conventional risk management approaches are based on hindsight knowledge, failure reporting and risk assessments calculating historical data-based probabilities, resilience engineering looks for ways to enhance the ability of organisations to be resilient in the sense that they recognise, adapt to and absorb variations, changes, disturbances, disruptions and surprises (Steen & Terje, 2011) . Risk Management is a quantitative approach to mitigating risk through measurable actions to deny network degradation through cause and effect mitigation. Whereas Resilience is executed through agility and flexibility to meet shocks not expected through Risk Management. Cyber Resilience is an enhancement of Cyber Risk Management by providing a response and mitigation to those items that are beyond Cyber Risk Management bounds. Cyber Resilience without an under pinning Cyber Risk Management plan would overwhelm the management of the network in a degraded state, with it never in a constant base line. Therefore the symbiosis between the two approaches is realised.

For Cyber Risk Management, it should be used as a business as usual approach to system security. It deals with the known or likely circumstance of network vulnerabilities. However, by 'their very nature, technological surprises are hardly amenable to risk-management approaches. Risk-based approaches require identification of threat, vulnerability and consequences (Kott, 2018)'. Cyber Resilience is used when the culmination of cascading risk realisation has occurred. This identifies that Cyber Risk management is intrinsically linked to Cyber Resilience, as it defines those networks not critical to function and can be isolated or quarantined to ensure business continuity is sustained. Risk management should define an enterprises 'crown jewels' so as resilience can be overlayed to protect them. For a military, this concept is akin to being able to absorb a breach by an enemy force in the tactical battle, encounter the enemy force and continue to execute the desired mission outcomes.

As Cyber Risk Management and Cyber Resilience are scaled to meet larger and more complex networks; control in executing robust symbiosis is difficult. Levin offers ' a fundamental issue in assessing the robustness of any system is that of *systemic risk*, namely the potential for localized disturbances to spread contagiously and become global problems (May, Levin, & Sugihara, 2008; Frank et al., 2014).' This could only be achieved through a thorough understanding of the network as a whole; which is not feasible on global scale. However a federated coalition of networks and resilience practices may be of use when protecting governments and national systems.

Managing the protection of a federated system would require a Whole of Government policy and coordination function.

4. Military roles in Cyber Risk Management and Resilience

As previously discussed, cyber risk management and resilience are intrinsically linked. The interplay is constantly maturing. This maturity when considering critical infrastructure is moving from threat-based thinking 'towards a more systemic and holistic perspective...this is an all-hazard approach to resilience analysis and strategy-making, wherein exposures and failure likelihoods are integrated with concepts such as networked vulnerability and coping capacity' (Theocharidou, Galbusera, & Giannopoulos, 2018). Many middle-power countries (L. Davidson, 2017) are expanding their military's role beyond traditional bordered threats to aid in national security. With maturity, this may naturally point middle power Army's towards cyber effects not directly associated with conventional warfare, but in support of sustaining National Power.

In 2016 Sweden identified a more Whole-of-Government approach to security, including national cybersecurity. An increased budget of the Sweden's Defence enabled an evolution to allow their total Defence concept to be 'concerned with trans-boundary threats and risks as well as with traditional territory defence'(Gruber, 2016). Sweden's approach is not isolated. In 2012 the United Nations assessed 47 of their membership states had 'assigned the role of national cyber-security, to their sovereign military forces to execute' (Sabillon, Cavaller, & Cano, 2016). Sabillion, Cavaller and Cano go on to define that the allocation of these roles is not limited countries of great national power, such as the United States of America and the United Kingdom, but also are seen as roles in Middle powers such as South Korea and Australia. These assessments highlight the dual requirement of military in ensuring their capability for dealing with traditional warfare is risk managed and resilient; but also, able to respond to national requirements for resilience.

5. Australian Army Context

The2016 Australian Department of Defence Whitepaper details the 'government will strengthen Defence's cyber capabilities to protect itself and other critical Australian government systems from malicious cyber intrusion and disruption' (Department of Defence, 2016). Supporting this whitepaper is the Integrated Investment Program that provides detail on the range of capabilities that will be developed and delivered out to 2026. It encompasses those infrastructure, platform and workforce projects that will be delivered as a 'whole-of-capability and whole-of-life approach to investment' (Government, 2016).

Although these policy's deal with the near future, the Australian Department of Defence is at a strategy cross-road of building the military that the government needs beyond 2026. Due to Australia's middle power complexity, it will always be challenged by what roles it refers to its Defence Force. The Australian Strategic Policy Institute assesses the Australian Government is at a crucial time to decide on what its military forces will do in the future. Stating if the 'government doesn't clearly articulate its priorities in relation to competition, limited war and major war, Australia will continue to remain without a clear and coherent strategic policy framework'(Fruhling, 2020). Although these military postures appear somewhat removed from cyber risk management and resilience; any strategic shifts in force posture or employment requires the Army to re-train and re-role workforces and evolve capability. This is particularly important for those specialist jobs requiring long training and deep understanding of capability evolutions of which cybersecurity, cyber risk management and cyber resilience are all inputs. The tensions described here have already been realised in the United States of America. In 2017 the Commander of U.S. Cyber Command Admiral Rogers described to the Senate of the Armed Services Committee the nexus between organisational requirements and training impacts. He highlights that without clear policy, priorities and direction, any training in technical skills is wasted as staff are relocated between different technical roles as tasking. Although the United States scale of requirement is vastly different from middle powers, the value of their lessons is apparent in this circumstance (*Navy Adm. Mike Rogers, commander of U.S. Cyber Command and director of the National Security Agency, testifies before the Senate Armed Services Committee, May 9, 2017.*, 2017).

Over the last 15 years, the Australian Army has been evolving its operational methodology, force structure and capability into the digital age through the Hardened Network Army, Plan Beersheba and now Army in Motion; the Army is accelerating its capability development and delivery to meet complex threats, coalition operations and domestic requirements in Defence of Australia (Fruhling, 2020). The interconnectedness in growing the Army's capability provides flexibility and agility in its employment, but it also brings with it vulnerability as described more broadly for the West by (Keith F Joiner & Tutty, 2018). The 'greater interdependence among different infrastructure networks is increasing the scope for systemic failures – whether from cyber-attacks, software glitches, natural disasters or other causes – to cascade across networks and affect society in unexpected ways' (Theocharidou et al., 2018). An Army is not immune to this risk and therefore the Australian Army must consider the treatment of cyber risk management and cyber resilience as it accelerates development. The final consideration of the Australian Army context is that the land forces will always deploy on joint operations (Department of Defence, 2016); indeed, they will likely operate primarily as part of a coalition or joint force. Therefore, the Army must be interoperable with those elements entailing the further considerations of not subjecting cyber-risk into the joint networks and aiding in cyber-resilience solutions of the joint capability. The system-of-systems implications of such cascading of networked forces is outlined by (Keith F Joiner & Tutty, 2018), with a significant issue being the federation of modelling and simulation support infrastructures for representative cyber-attack surfaces. The cascading of joint networks is also a significant issue in defending civilian networks, like electronic health (Kruse, Smith, Vanderlinden, & Nealand, 2017) human services, banking (Keith F Joiner et al., 2018) and logistics. This highlights the complexity in training the workforce in the design, management and protection of these systems; of which all touch government enablement.

Considerable academic analysis of the types of threats and response actions Australia will need to undertake has occurred. In general this has been a tension of what is more appropriate for Australia as a middle power; in that a military provides a response of troops being either a regular defence member or part of a ready militia (Austin, 2016). Tension and complexity arrives potentially due to a mix of Australian policy gaps (K. F. Joiner, 2017) and wider social acceptance of key cybersecurity concepts (Atkinson, Smallhorn, Caldwell, & Tolhurst, 2016). The work force options may mature once the Australian Army understands its role and training requirements to be able to support the Nation in enabling cybersecurity.

6. Training

Training a workforce in cyber risk management and resilience will require multiple approaches. This is due to the range of networks a military is responsible for and the ever-changing threat environment. Kott offers 'in order to be resilient to a technological surprise, a military organization must invest in developing skills and competencies of its personnel and leaders. It requires investments in training and exercises' (Kott, 2018). This identifies that both risk management and resilience is not just a role for the cybersecurity specialist but requires buy in and function from other roles such as operational planners and senior leadership. Understanding what level of training is required for a military also depends on what levels of risk management and resilience they are empowered to undertake. Analysis of the U.S. shows skills and techniques, such as cyber-table-topping and security development operations (i.e., portmanteau SecDevOps), that where developed primarily by, and for, Defense, are readily needed and adapted for civilian defence of critical infrastructure (Chan, 2018; Christensen, 2017; Matteson, 2017). As such, military training has proven to lead more than the training of its own cybersecurity force, irrespective of where it is employed.

Analysis of both the Australian Government's 2016 Whitepaper and Integrated Investment program highlight the roles of the military in capability development and the effect of providing a cyber-secure environment to ensure this defence force can operate effectively (Government, 2016). This implies a cyber-risk management approach needs to be applied to the capability development and operation of military platforms and networks. A suite of training is required for the workforce to be able to do this. Coupled with this training requirement is the ability to effect cyber resilience on military networks and as required in support of government requirement. When delivering cyber-resilience, it has been highlighted that for cyber-resilience to be 'effective and efficient it needs to be addressed holistically and on several levels in parallel'(Björck et al., 2015). This is further refined to expose cyber resilience needs to be possible at the technical, organisational, regional and through to supranational levels. These are all layers and environments where a military will operate on both internal and external networks.

All of these requirements in both cyber risk management and cyber resilience convey a vast array of skilling and education requirement to sustain these environments. Figure 1 provides a suggested array of workforce roles a middle-power Army requires to fulfil its function. It suggests the layers and relative workforce size to provide an overall cyber workforce effect to the Australian Army (Lisa Davidson, 2016). Defining the level of training and skilling requires an understanding of volumes of roles an Army will fill and the threat or cyber-attack it needs to respond to. This definition includes those roles external to an Army that support national defence responses. This research seeks to define the volumes of roles and the skilling required.

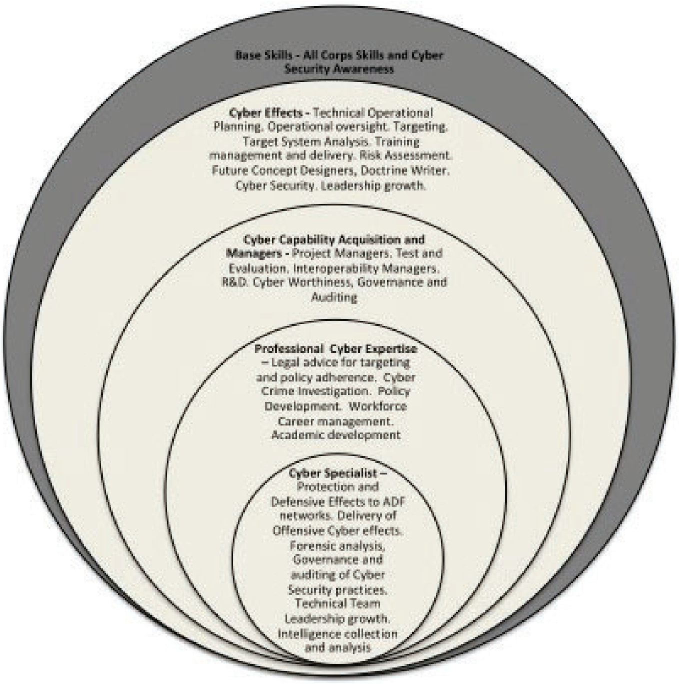

Base Skills - All Corps Skills and Cyber Security Awareness
Cyber Effects - Technical Operational Planning. Operational oversight. Targeting. Target System Analysis. Training management and delivery. Risk Assessment. Future Concept Designers, Doctrine Writer. Cyber Security. Leadership growth.
Cyber Capability Acquisition and Managers - Project Managers. Test and Evaluation. Interoperability Managers. R&D. Cyber Worthiness, Governance and Auditing
Professional Cyber Expertise – Legal advice for targeting and policy adherence. Cyber Crime Investigation. Policy Development. Workforce Career management. Academic development
Cyber Specialist – Protection and Defensive Effects to ADF networks. Delivery of Offensive Cyber effects. Forensic analysis, Governance and auditing of Cyber Security practices. Technical Team Leadership growth. Intelligence collection and analysis

Figure 1: Proposed Cyber Workforce Array

Although workforce roles and technical training can be defined through an understanding of credentials provided by government and commercial agencies. There is a body of analysis that highlights organisational and individual requirement beyond this. Broadly they could be defined as ‘soft skills’, but also highlight particular organisational, cultural and operating methodologies that inculcate effective cyber risk management and resilience approaches. Table 1 analyses three views of desired cyber resilience requirements of a government organisation. Beyond technical skilling, these functions are pivoted from dynamic organisational design and operating methodologies coupled with leadership approaches. Although these could be considered organisation approaches and not technical cyber-skills-related; they highlight the holistic systems approach required to execute cyber resilience. For a middle power military, this is a vast environment to be trained in.

Table 1: Organisational requirements for resilience to cyber attacks

Requirement 1 **(Kott, 2018)**	**Requirement 2** **(Theocharidou et al., 2018)**	**Requirement 3** **(Stojadinovic, 2018)**
Flexibility	Agility	Information sharing
Agility	Diversity	Collaboration
Breadth	Technology	Critical Infrastructure Technology
Diversity	Governance (auditing)	Agility

7. Research

Training for Cyber Risk Management and Cyber Resilience for a military force will require multiple approaches to encapsulate the needs of figure 1 and table 1. This is due to the skills requirement in design, operating and maintaining complex networks both internally and externally to the organisation. To further define the workforce and training requirements for the Australian Army; the following research recommendations are provided.

1. Define the likely systems and threats that will require cyber risk management and cyber resilience from the Australian Army.

2. Define the strategy of sustaining the dual roles of internally protecting operating networks and aiding the defence of the Nation.
3. Define the skills of providing cyber risk management and cyber resilience in capability design.
4. Confirm agility, diversity and collaboration is vital to organisational approaches in the recovery of an Army system.
5. How is an Army capability designed to ensure cyber risk management practices and support cyber resilience? How is this design methodology trained?
6. What skilling investment in critical infrastructure protection is required by an Army.

Through a mixed-method approach to research, these questions will be addressed. This will enable mapping of layers of skills type, multiple instances of those skills between Cyber Risk Management and Cyber Resilience and include the technical levels required to meet internal Army requirements and those tasked of government.

8. Conclusion

The domain of Cyber Risk Management and Cyber Resilience for a Middle Power military force is complex. Firstly, that force must understand all of its systems and ensure it secure from vulnerability through Cyber Risk Management and Cyber Resilience. Those systems can very between tactical digital radios, vehicle fleets, and business systems to those that rely on national infrastructure such as Global Positioning Systems, health and finance networks. By virtue of providing the function of defence to a country, Middle Power military's may also need to provide Cyber Resilience responses to parts of national infrastructure as requested by government in pursuit of national security. Understanding those networks will require deep technical skill, positive engagement with government agencies and a culture of agility.

The design and subsequent skilling of that response requires an understanding of the policy behind response, the possible threat level to counter and the level of technical complexity required of its military force. As discussed, for the Australian Army and wider Australian Defence Force, Cyber Risk Management and Resilience is more than the internal response to a cyber-related incident. It has two roles to undertake for Cyber Risk Management and Cyber-Resilience. In order to meet a future national strategy, an Army must design, educate and train those forces that will undertake these roles both internally and externally to the organisation. Ensuring both Cyber Risk Management and Cyber Resilience is foundation to design to capability is possibly the only way to ensure it is considered in detail. However this cannot be done until a realisation of the fundamental organisational and technical skills requirements is completed. This research seeks to define that to ensure middle power military's are capable of achieving this in the future.

References

Alberts, C., Haller, J., Wallen, C., & Woody, C. (2017). Assessing DoD System Acquisition Supply Chain Risk Management. *CrossTalk, 30*(3), 4-8.

Argaw, S. T., Bempong, N. E., Eshaya-Chauvin, B., & Flahault, A. (2019). The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review. *BMC Medical Informatics and Decision Making*.

Atkinson, S. R., Smallhorn, C., Caldwell, N., & Tolhurst, G. (2016). *Identification and Classification Designing and Engineering Synthetic Ecologies*. Paper presented at the Systems Engineering, Test and Evaluation, Melbourne, Australia.

Austin, G. (2016). Australia Rearmed! Future Needs for Cyber-enabled Warfare. In *ACCS Discussion Paper No.1* (pp. 33). Australian Centre for Cyber Security: University of New South Wales.

Björck, F., Henkel, M., Stirna, J., & Zdravkovic, J. (2015). *Cyber Resilience - Fundamentals for a Definition* (R. A., C. A., C. S., & R. L. Eds. Vol. 1). Sweden: Springer, Cham.

Brennan, G., Joiner, K. F., & Sitnikova, E. (2019). Architectural Choices for Cyber Resilience. *Australian Journal of Multidisciplinary Engineering, 15*, 68-74.

Chan, S. (2018). *Prototype Orchestration Framework as A High Exposure Dimension Cyber Defense Accelerant Amidst Ever-Increasing Cycles of Adaptation by Attackers: A Modified Deep Belief Network Accelerated by A Stacked Generative Adversarial Network for Enhanced Event Correlation*. Paper presented at the CYBER 2018: The Third International Conference on Cyber-Technologies and Cyber-Systems, Athens, Greece.

Christensen, P. (2017). Cybersecurity Test and Evaluation: A Look Back, Some Lessons Learned, and a Look Forward! *ITEA Journal, 38*(3), 221–228.

Davidson, L. (2016). ***Considerations for Establishing the Australian Army's Cyber Work Force***. Paper presented at the SecureCanberra@MilCIS 2016, Canberra, Australia.

Davidson, L. (2017). *Analysing the Characteristics of Middle Power Cyber Capability*. Paper presented at the European Conference on Cyber Warfare and Security.

Department of Defence. (2016). *Defence White Paper 2016*. Canberra, Australia: Commonwealth of Australia

Fowler, S., & Sitnikova, E. (2019). *Toward a framework for assessing the cyber-worthiness of complex mission critical systems*. Paper presented at the Military Communications and Information Systems Conference, Canberra, Australia.

Fowler, S., Sweetman, C., Ravindran, S., Joiner, K. F., & Sitnikova, E. (2017). Developing cyber-security policies that penetrate Australian defence acquisitions. *The Australian Defence Force Journal, 102*.

Fruhling, S. (2020). Reassessing Australia's defence policy (part 2): What are our strategic priorities? *The Strategist*.

Government, A. (2016). *Integrated Investment Program*. Canberra, Australia: Commonwealth of Australia Retrieved from https://www.defence.gov.au/WhitePaper/Docs/2016-Defence-Integrated-Investment-Program.pdf

Gruber, B. (2016). Going Side by Side: Defence and Resilience in Swedish Security Policy. *Social Science Open Access Respositry, Working Paper / Austrian Institute for International Affairs*.

Joiner, K. F. (2017). How Australia can catch up to U.S. cyber resilience by understanding that cyber survivability test and evaluation drives defense investment. *Information Security Journal: a Global Perspective, 26*, 74-84.

Joiner, K. F., Ghildyal, A., Devine, N., Laing, A., Coull, A., & Sitnikova, E. (2018). Four testing types core to informed ICT governance for cyber-resilient systems. *International Journal of Advances in Security, 11*.

Joiner, K. F., & Tutty, M. G. (2018). A tale of two allied defence departments: new assurance initiatives for managing increasing system complexity, interconnectedness and vulnerability. *Australian Journal of Multi-Disciplinary Engineering, 14*(1), 4-25.

Kott, A. (2018). Technological Surprise and Resilience in Military Systems. In B. Trump, M. Florin, & I. Linkov (Eds.), *IRGC Resource Guide on Resilience: Domains of resilience for complex interconnected systems* (Vol. 2, pp. 89-93).

Kruse, C. S., Smith, B., Vanderlinden, H., & Nealand, A. (2017). Security Techniques for the Electronic Health Records. *Journal of Medical Systems, 41*(127).

Matteson, S. (2017, 3 April). DevSecOps: What it is and how it can help you innovate in cybersecurity. *ZDNet*. Retrieved from https://www.zdnet.com/article/devsecops-what-it-is-and-how-it-can-help-you-innovate-in-cybersecurity/

Navy Adm. Mike Rogers, commander of U.S. Cyber Command and director of the National Security Agency, testifies before the Senate Armed Services Committee, May 9, 2017., (2017).

Refsdal, A., Solhaug, B., & Stolen, K. (2015). Cyber-risk Management. In C.-r. Management (Ed.), *Cyber Risk Management* (pp. 33-47): Springer, Cham.

Sabillon, R., Cavaller, V., & Cano, J. (2016). National Cyber Security Strategies: Global Trends in Cyberspace. *International Journal of Computer Science and Software Engineering, 5*(5), 67-81.

Schreider, T. (2017). *Building Effective Cybersecurity Programs: A Security Manager's Handbook*: Rothstein Publishing.

Sher, M.-L., Talley, P. C., Yang, C.-W., & Kuo, K.-M. (2017). Compliance with Electronic Medical Records Privacy Policy: An Empirical Investigation of Hospital Information Technology Staff. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 54*, 1-12.

Steen, R., & Terje, A. (2011). A risk perspective suitable for resilience engineering. *Safety Science, 49*(2), 292-297.

Stojadinovic, B. (2018). The case for Systemic Resilience; Urban Communites in Natural Disasters. In B. Trump, M. Florin, & I. Linkov (Eds.), *IRGC resource guide on resilience (vol. 2): Domains of resilience for complex interconnected systems.* (Vol. 2, pp. 141-146).

Theocharidou, M., Galbusera, L., & Giannopoulos, G. (2018). Resilience of Critical Infrastructure Systems: Policy, Research Projects and Tools. In B. Trump, M. Florin, & I. Linkov (Eds.), *IRGC Resource Guide: Domains of resilience for complex interconnected systems* (Vol. 2, pp. 147-159).

U.S. Government Accountability Office. (2018). *Weapon Systems Cybersecurity: DOD Just Beginning to Grapple with Scale of Vulnerabilities*. Washington

Zafar, H. (2016, August 11-14). *Cybersecurity: Role of behavioral training in healthcare.* Paper presented at the 22nd Americas Conference on Information Systems: Surfing the IT Innovation Wave, San Diego, California.

Designing Attack Infrastructure for Offensive Cyberspace Operations

Gazmend Huskaj[1, 2], Ion A. Iftimie[3, 4] and Richard L. Wilson[5,6]
[1]Swedish Defence University, Stockholm, Sweden
[2]University of Skövde, Skövde, Sweden
[3] Eisenhower PhD Fellow, NATO Defense College, Rome, Italy
[4] Senior Advisor, European Union Research Center, George Washington School of Business, Washington, D.C.
[5] Towson University, U.S.A.
[6] Senior Research Scholar, Hoffberger Center for Professional Ethics, University of Baltimore
gazmend.huskaj@fhs.se
iftimie@gwu.edu
wilson@towson.edu
DOI: 10.34190/EWS.20.077

Abstract: This article addresses the question 'what considerations should be taken by cyber commands when designing attack infrastructure for offensive operations?'. Nation-states are investing in equipping units tasked to conduct offensive cyberspace operations. Generating 'deny, degrade, disrupt, destroy or deceive' effects on adversary targets requires to move from own ('green'), through neutral ('grey'), to adversary ('red') cyberspace. The movement is supported by attack infrastructure for offensive cyberspace operations. In this paper, we review the professional and scientific literature identifying the requirements for designing an attack infrastructure. Next, we develop and define the concepts for attack infrastructure. Finally, we explain and describe the considerations for designing attack infrastructure. The research question is answered by proposing a framework for designing attack infrastructure. This framework is vital for military and civilian commands designing attack infrastructure for offensive cyberspace operations.

Keywords: Attack Infrastructure, Cyber Commands, Offensive Cyberspace Operations, Operational Security.

1. Introduction

Knowing how to design attack infrastructure supporting offensive operations is a relevant topic for cyber commands around the world. Nation-states are investing in civilian and military units to conduct offensive cyberspace operations. Generating 'deny, degrade, disrupt, destroy or deceive' effects on adversary targets requires to move from own ('green'), through neutral ('grey'), to adversary ('red') cyberspace. The movement is supported by attack infrastructure for offensive cyberspace operations. The case of WannaCry has demonstrated the impact uncontrolled weaponized code can have on civilian infrastructure (Chen and Bridges, 2017; Kao and Hsiao, 2018; Mohurle and Patil, 2017). The cases of Operation Cloud Hopper (Bird, 2015; Galinkin et al., n.d.; Vincent, 2019) and Operation Glowing Symphony (Iftimie, 2019; Jacobsen, 2019) have demonstrated the impact of controlled weaponized codes and required infrastructures to support these operations. Probing scientific databases reveals 15 articles on the topic. Probing red team literature also reveals numerous thoughts and ideas on attack infrastructure. Based on this review of literature, this study seeks to answer the following research question: what considerations should be taken by cyber commands when designing attack infrastructure for offensive operations?

The main contributions are summarized as follows:

1. attack infrastructure is described by conducting a review of the scientific and professional literature;
2. based on 1), concepts are developed and defined;
3. then, the considerations for designing attack infrastructure are explained and described;
4. finally, the design considerations for attack infrastructure are provided, followed by operational security and ethical considerations, and policy recommendations.

This study begins with a review of the professional and scientific literature identifying the requirements for designing an attack infrastructure. This results in a conceptual framework presented in Table 1. The research approach for this research is described in Section 3, and the attack infrastructure considerations are presented in Section 4. Section 5 develops theory for attack infrastructure. Sections 6 and 7 discuss Cybersecurity and

Ethical considerations, while Section 8 provides policy recommendations. Conclusions and future work are presented in Section 9.

2. The attack Infrastructure from practitioner and scientific points of view

Attack infrastructure is a requirement for conducting offensive cyberspace operations. Offensive cyberspace operations are defined as "a sequence of planned actions executed by an organized group of people with a defined purpose in and through hardware and software which are used to create, process, store, retrieve and disseminate information in different types of interconnected networks that build a large, global network, built and used by people" (Huskaj & Wilson, 2020). Offensive methods include, but are not limited to, obfuscation, redirection, and social engineering. The organized group of people requires infrastructure which enables the commands to reach the intended target(s). The design requirements for the attack infrastructure consist of segregation of duty, i.e. segregate Homebase or the forward operating base (FOB) from the C&C-server, phishing and payload servers.

The requirements for attack infrastructure consist of, domains, redirectors, servers for command & control (C2), phishing and payload delivery (Dimmock & Borosh, 2018). The domain should blend with the targets "baseline web traffic or fit with the social engineering campaign" (Dimmock, 2017). Redirectors are used to obfuscate and protect the Homebase or FOB from response actions. Two types of C2-servers exist: one for C2-actions on the target, and one to ensure permanence. The C2-server for permanence should never be used for actions on the target due to the risk of losing permanence (Mudge, 2014). They can also be used to redirect a target to the web site holding the payload (Dimmock, 2017). Social engineering and phishing attacks are used for payload delivery to the target (Dimmock & Borosh, 2018). Communication between the organized group, their C2-server to their other assets is always encrypted (Dimmock & Borosh, 2018; Feroze, 2018). These insights are supported by scientific research as depicted in Table 1.

Table 1: Summary of the required infrastructure, tools, techniques and procedures for offensive operations

Concept	Definition	Author
Abuse legitimate services	Creating malicious domain names	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018)
Advanced Persistent Threat (APT)	Operate many domains and subdomains	(Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017)
Backdoors	Malware that opens ports and calls the C&C, enabling an attacker to take actions in the target system	(Hrad O., Kemppainen S., 2016; Leontiadis N., Moore T., Christin N., 2014)
Brute-forcing	To gain access into a system	(Hrad O., Kemppainen S., 2016; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017)
Buffer overflow	Enables putting more data in memory causing execution of code	(Mimura M., Tanaka H., 2017)
C&C	Send commands to the target	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018; Hrad O., Kemppainen S., 2016)
Collection and threat collection	Collecting information about the target system, but also exfiltrating information once inside the target	(Antonatos S., Akritidis P., Lam V.T., Anagnostakis K.G., 2008; Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Kim N., Lee S., Cho H., Kim B.-I., Jun M., 2018)
Content delivery network (CDN)	Use them to deliver malware	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018)
Cyber sanctions	The actual or threatened restriction of digital transactions to affect a behavioral change by the target through the introduction of psychological pressure against its political leaders and populace	(Iftimie I., 2019)
DDoS - NTP, TCP	Various protocols may be used to conduct DDoS-attacks	(Berti-Equille L., Zhauniarovich Y., 2017; Collier B., Thomas D.R., Clayton R., Hutchings A., 2019; Karami M., Park Y., McCoy D., 2016; Krupp J., Backes M., Rossow C., 2016; Mezzour G., Carley K.M., Carley L.R., 2017)
Directory traversal	Enables access to restricted directories in a web-server	(Mimura M., Tanaka H., 2017)

Concept	Definition	Author
DNS misconfiguration, reliable, malicious resolver	DNS are reliable infrastructure to use for beacon-sending and C&C	(Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018; Karami M., Park Y., McCoy D., 2016)
Domains	Using domains, domain-names, sub-domains and domain generation algorithms to create new infrastructure and increase success rate of attacks which can resemble well-known legitimate domains	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018; Kim N., Lee S., Cho H., Kim B.-I., Jun M., 2018; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017; Lu L., Perdisci R., Lee W., 2011)
Example infrastructure	Brute force into a target, pass info to C&C, which then downloads additional software	(Hrad O., Kemppainen S., 2016; Mezzour G., Carley K.M., Carley L.R., 2017)
Exploits, Kit, Angler, Blackhole Toolkit	Tools to conduct actions on the target	(Leontiadis N., Moore T., Christin N., 2014; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017; Mezzour G., Carley K.M., Carley L.R., 2017)
Fake applications	Deceiving the user to install it, and once installed, asks user to pay premium, or ransomware	(Mezzour G., Carley K.M., Carley L.R., 2017)
HTTP	A protocol to distribute exploits or malware and for C&C	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018; Mimura M., Tanaka H., 2017)
Information systems - malicious that launch attacks	Own systems or hijacked systems used for further attacks	(Mezzour G., Carley K.M., Carley L.R., 2017)
Injection	Injecting scripts into parent web-documents	(Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012)
IP addresses, and malicious	Many IP-addresses are required to conduct operations because some may be burned while conducting operations	(Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Lu L., Perdisci R., Lee W., 2011; Kim N., Lee S., Cho H., Kim B.-I., Jun M., 2018)
Internet Service Providers (ISPs)	Get IP-addresses from many different ISPs to conduct attacks and for resilience	(Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012)
Malware, ransomware, trojan	Software to achieve effects. Malicious from the victim's perspective	(Hrad O., Kemppainen S., 2016; Kim N., Lee S., Cho H., Kim B.-I., Jun M., 2018; Leontiadis N., Moore T., Christin N., 2014; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017)
MitM	Man-in-the-middle attack to extract credentials used for future attacks	(Almeshekah M.H., Atallah M.J., Spafford E.H., 2015)
Obfuscation	Making it difficult for the target/defenders to identify the attackers motives	(Antonatos S., Akritidis P., Lam V.T., Anagnostakis K.G., 2008)
Persistence	Have presence on many different targets that enable access if some targets are identified and blocked by the defenders	(Leontiadis N., Moore T., Christin N., 2014; Lu L., Perdisci R., Lee W., 2011)
Phishing - Harvesting credentials	Deceiving the user to download and run a file, or click on a link which enables access to a target	(Almeshekah M.H., Atallah M.J., Spafford E.H., 2015; Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017)
Privilege escalation	Using the privileges of newly gained access in a target for future actions	(Hrad O., Kemppainen S., 2016)
Re-registration	Re-registering domain names to increase chance of success of attacks	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018)
Redirection - and click link	Use various techniques to redirect users to website that enables attack success	(Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Leontiadis N., Moore T., Christin N., 2014; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017; Lu L., Perdisci R., Lee W., 2011; Mezzour G., Carley K.M., Carley L.R., 2017)
Reflection attack	Using various protocols to conduct amplify attack traffic	(Karami M., Park Y., McCoy D., 2016)
Rootkits	Software hidden from the target, enables persistence	(Hrad O., Kemppainen S., 2016)

Concept	Definition	Author
Search poison - engine poisoning	Use popular keywords to cause user interaction which then are redirected to the attackers site	(Leontiadis N., Moore T., Christin N., 2014; Lu L., Perdisci R., Lee W., 2011)
Sinkholing	Redirecting network traffic from the intended destination to the attackers information system	(Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018)
Social engineering	Various techniques to deceive the user into various actions; download file, click on link, give credentials	(Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017)
Spoofing	Spoofing the target's source address overwhelming them with network traffic denying access to services	(Karami M., Park Y., McCoy D., 2016)
Triggering	Strings that trigger browser-specific code, such as redirection	(Antonatos S., Akritidis P., Lam V.T., Anagnostakis K.G., 2008; Leontiadis N., Moore T., Christin N., 2014)
Web exploitation	Techniques that enable access to a target. Web architecture-compromising server, HTML, Javascript, browser, hosting sites for phishing, exploit browser to deliver malware via advertising ecosystem, cross-site scripting, man-in-the-browser, malicious iFrames, Drive-by-download	(Almeshekah M.H., Atallah M.J., Spafford E.H., 2015; Alrwais S.A., Dunn C.W., Gupta M., Gerber A., Spatscheck O., Osterweil E., 2012; Antonatos S., Akritidis P., Lam V.T., Anagnostakis K.G., 2008; Chiba D., Akiyama M., Yagi T., Yagi T., Mori T., Goto S., 2018; Liu D., Li Z., Du K., Wang H., Liu B., Duan H., 2017; Mezzour G., Carley K.M., Carley L.R., 2017; Mimura M., Tanaka H., 2017)

3. Research approach

Academic research on attack infrastructure for offensive operations is limited. The identified research is primarily focused on defensive measures by first noting attacker-tools, tactics, techniques and procedures (herein modus). Then, after the threat's modus is described, the researchers discuss various defensive measures. The search for academic articles, used the string "attack infrastructure." It was done in Elsevier's Scopus® database, the ACM Digital Library, and IEEE Xplore®, resulting in 14, 22, and 8 document results respectively. Next, all of the documents were manually reviewed by GH. The requirements for an article to be within scope for inclusion were: it must discuss attack infrastructure, methods, tools, tactics, techniques and procedures. The exclusion criteria were articles beyond scope and doubles. Therefore, the final articles for review amounted to 15, which are also depicted in Table 1.

The same criteria applied when searching for professional documents on the topic. However, the search string used in a search engine (Google), was "red team attack infrastructure" without the apostrophes. The listed results were manually reviewed by GH, and snowballing began with Leibowitz & Timzen (2019). The included professional sites are: Kohlenberg (2018); Feroze (2018); Gimmick & Borosh (2018); Dimmock (2017); and Mudge (2014).

The results of the collected data were twofold: the research dataset is already mentioned above, while the "professional" dataset discussed design requirements on attack infrastructure, operational security, and behavior of red team operators when conducting offensive operations. These two perspectives are noted in section 2.

4. Components in Attack Infrastructure

The domain name system (DNS) is a critical part of attack infrastructure. It is a critical function of the Internet: it maps the IP-address of an information system to a name. Information systems are better at managing numbers while humans are better at managing names. For example, the IP-address `127.0.0.1`, "assigned for use as the Internet host `loopback` address" (Cotton & Vega, 2010), is also known as "`localhost`", i.e. this computer (Cheshire & Krochmal).

DNS is a distributed system of databases and "is based on a hierarchical database containing resource records (RRs) that include the name, IP address, and other information about hosts" (Stallings, 2006, p.777). The RRs are many, but those of primarily importance are the A and NS records. The A record maps the IP address to a the system name, e.g.: `127.0.0.1 localhost`. The NS record points to the authoritative name server for this particular domain (Stallings, 2006). `example[.]com` is an example domain with the NS-record `ns1.example[.]com`. The A record is used to point domains to redirectors (more on redirectors below) and team servers (Mudge, 2014).

Therefore, using domains, domain-names, sub-domains to create new infrastructure that resembles well-known legitimate domains increases the success rate of attacks. In the 'Operation Cloud Hopper' case, attackers used domains such as `mailserever[.]com`, `mailsserver[.]com`, and `mailvserver[.]com` (PwC, 2017). These are domain names where a user may be tricked into believing are legitimate mail servers, but in fact are not. The second way to generate domains quickly are domain generation algorithms (DGA). A DGA is implemented in the weaponized code to constantly provide it with new domains on the fly for communications to the command and control server (Arntz, 2016; Scarfo, 2016). Examples include `uqhucsontf[.]com`, `myypqmvzkgnrf[.]com`, `ocufxskoiegqvv[.]com` (Scarfo, 2016). In the 'Operation Cloud Hopper' case, 1375 domain names were used in that operation (PwC & BAE Systems, 2017). This shows that it is recommended to have multiple domain names for operations.

4.1 Command and control servers (C2)

C2 servers ensure that communication flows between Homebase and the target. The complexity of the Internet, various protocols and services, makes it possible to utilize the same for C2. The first step however, is to establish a front domain. The front domain serves the purpose of being a legitimate channel enabling communications to/from the target, and to/from the C2-server. Using a legitimate domain makes it easier to blend in with other communications in the target's infrastructure and reduces the likelihood of being spotted. Examples of legitimate domains are DropBox, and OneDrive (Sharma & Singh, 2018), or Google hosts, e.g. `google[.]com`, `gmail[.]com`, `mail.google[.]com` (Nahari, 2017). The target-type decides which front domain is used based on the collection phase on the target.

4.2 Redirectors

Redirectors are servers that, as the name implies, redirect communications from/to the target to/from the Homebase. Two versions exist: one for ongoing operations (known as short haul) and one for maintaining presence in the target (long haul). The Homebase can have one or several redirectors in front for operation security, by making it more difficult for the adversary to find Homebase and related infrastructure. The short haul servers are used to take actions on the target (Mudge, 2014; Sharma & Singh, 2018). The short haul servers are protected by one or more redirectors, and if a defender takes action, only the identified short haul server is destroyed. This enables conducting actions on the target through another redirector while simultaneously deploying a new redirector to take the place of the previously destroyed one. The long haul servers are used to ensure permanence in the target (Mudge, 2014; Sharma & Singh, 2018). The difference with the short haul servers is that long haul servers have longer callback timers to C2 indicating they're alive. The pre-determined time (days, weeks, months, years) of the offensive operation determines the callback timer, which could be once every 24 hours, to more (Mudge, 2014).

4.3 Phishing and payload delivery

Deceiving the user to download and run a file, or click on a link which delivers the payload and enables access to a target's information system is the purpose of phishing. Collection on the target decides which type of payload(s) should be chosen: "HTML Applications (HTAs), Embedded OLE objects in Office documents, Office macros, and Windows shortcut (LNK) files" (Dimmock, 2017). Two possible courses of action (COAs) exist: 1) based on collection, choose one or several of the payload types and deploy them from own owned servers, cloud providers, or hijacked servers; 2) based on collection, create a test-bed environment, conduct tests on that, and based on the results, adjust payload and payload delivery. Just as with other infrastructure, one or several redirectors should be put in front of the payload delivery server to protect it from defenders.

4.4 Encryption

Encryption should be a standard feature to ensure communications are not eavesdropped by a third party. OpenSSL with a custom generated certificate, or a legitimate one, or by impersonating another vendor, can be used to encrypt communications to/from the target (Sharma & Singh, 2018). Encryption is also needed to make it difficult for defenders to detect payloads by using payloads that offer encryption (Sharma & Singh, 2018).

4.5 Protection

C2 servers, stagers and payload deliverers are high-value infrastructure. Protecting them by deploying redirectors in front of them has already been mentioned. Additional actions to be taken include enabling firewall rules on these assets. The firewall rules should only allow the infrastructure under the organized group's control the ability to connect to C2, stagers, redirectors and payload deliverers (Sharma & Singh, 2018).

5. Design Considerations: Framework Development

America's International Technology Education Association defines design "as an iterative decision-making process that produces plans by which resources are converted into products or systems that meet human needs and wants or solve problems" (Banach and Ryan, 2009, p. 105). Within the context of an attack infrastructure for offensive cyberspace operations, design is a conceptual stage, which precedes the more concrete planning and building/execution stages of the attack infrastructure. Figure 1 gives an overview of the design considerations for attack infrastructure supporting offensive operations. This is supported by scientific literature and professional literature. The policy level tasking the organized group to conduct operations is not considered. It is assumed that the policy level have given the green light/"GO" to conduct offensive operations.

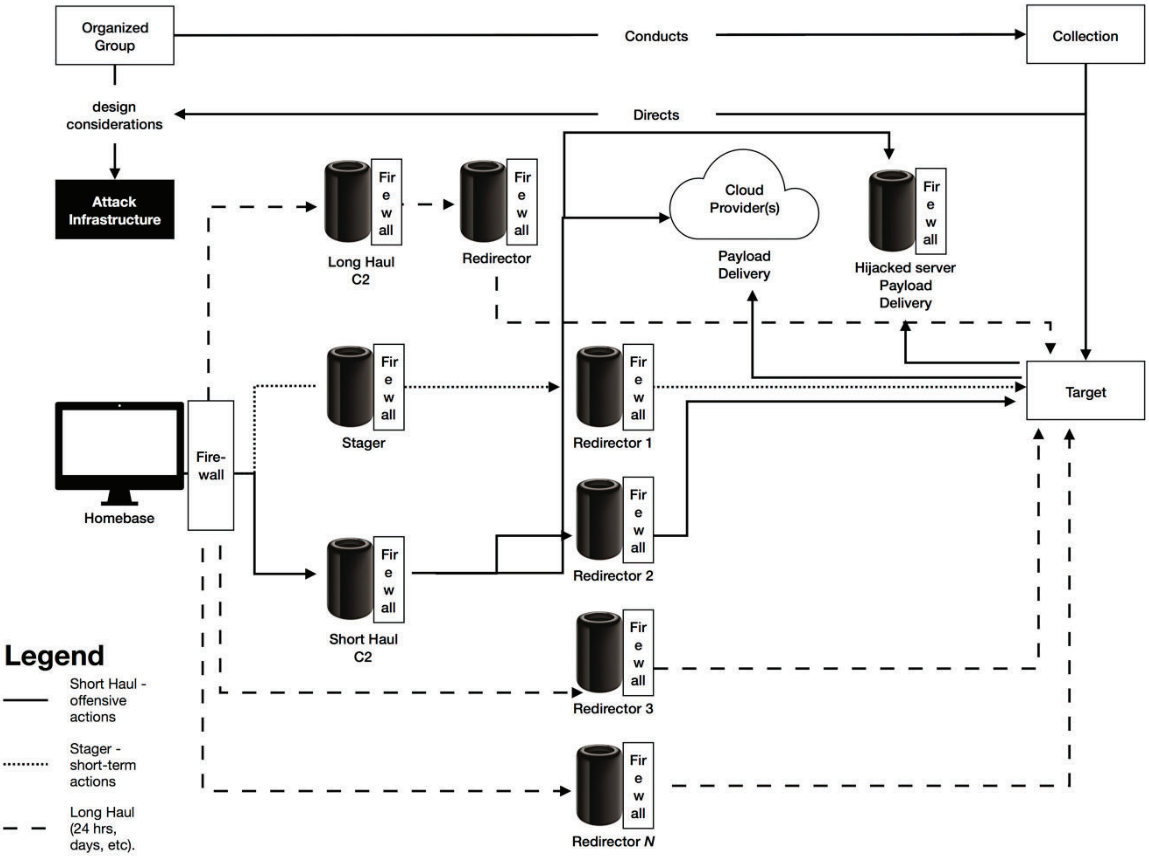


Figure 1: Simplified example of design considerations for attack infrastructure

The organized group begins by conducting intelligence collection on the target organization. Collection can target numerous open sources like social media, the organization's website(s), and `whois`. Collection also determines what kind of operating system(s), web pages, services (e.g. OneDrive, DropBox, etc.), sensors (e.g. intrusion detection systems and anti-virus systems) the target is using. Many tools exist that can automate this stage. This information results in a list of assets which directs collection to vulnerability databases and exploit databases, directing offensive methods. Furthermore, applying machine learning algorithms on employees in the target organization enables generating near-perfect profiles. These profiles may then be used to generate spear-phishing e-mails, with attachments or links. The attachments or links will then redirect the target user to one or several payload sites. These sites are either hosted on cloud provider's infrastructure, or in hijacked servers. One key requirement is encrypting the payload to avoid detection by various sensors in the target, such as intrusion detection systems and anti-virus systems.

Designing the infrastructure requires many domain names, either close to impersonation of legitimate ones, or hiding behind legitimate ones. To register domain names, scripts can be used to generate human readable domains like `mailserever[.]com`, `mailsserver[.]com`, and `mailvserver[.]com`. Another option is to employ a DGA to generate domains like `uqhucsontf[.]com`, `myypqmvzkgnrf[.]com`, and `ocufxskoiegqvv[.]com`. The servers which the domain names point to are used as redirectors, long haul C2, short haul C2, and for staging attacks. Each of these high-value assets have a redirector in front of them with a firewall. The purpose is to make it harder for defenders to trace the operation. Finally, once the payload is delivered to the target, and activated, the stager is used to establish permanence by adding additional software which calls back to the long haul C2 every 24 hours or more; and software which calls back to the short haul C2. The operators will then use the short haul C2-server to conduct further actions. These actions can be to conduct more collection to learn the infrastructure, position additional software for long haul C2, but also create escalate privileges of existing users to avoid detection by creating new users.

6. Operational Security Considerations

Similarly to the concept of protection, which uses firewalls to conceal attack infrastructure, the primary operational security considerations are those of using hardware and software that insures security of the attack infrastructure as well as tactics, techniques, and procedures (TTPs) that conceal the attack infrastructure's design, functions, and attribution of users.

First, software and hardware used in the attack infrastructure must be secure. This means that when considering updates to the software and hardware of the attack infrastructure, a serious discussion should take place whether to implement the updates or not and weigh the risks about potential effects on the infrastructure, and ongoing offensive operations.

Secondly, the issue of attribution of attack infrastructure must be addressed. If any node (hardware or software) or TTP of the attack infrastructure can be attributed to it, this endangers not only the attack infrastructure, but also the success of offensive cyberspace operations and the safety of the military and/or civilian units conducting them. For this reason, operational security is of outmost significance for the design of attack infrastructure for offensive cyberspace operations. Lack of operational security during the design of attack infrastructure for offensive cyberspace operations has led in the past not only to attribution of attack infrastructures, but also to the identities of their users. For example, attribution of many Chinese and Russian intelligence officers that ended up on the FBI Cyber's Most Wanted list was made possible by failed Chinese and Russian designs of attack infrastructure for cyberspace operations.

7. Ethical Considerations

The primary ethical issues here are using cloud service provider's infrastructure for payload deliver, hijacking servers to do the same, and impersonating legitimate actors. Furrow (2005) states that ethics is related to evaluating actions done by those who are capable of being moral agents. The focus can be on the person, intent, motive the act or the impact. The ethical issues are identified by asking four questions: 1) what actions are performed when taking actions to deploy a payload and/or impersonate a legitimate actor; 2) what is the character of the person(s) conducting those actions; 3) what are their intentions; and 4) what are the consequences of the payloads. These questions can be analyzed by applying a set of rules developed by a community of scholars. A set of 5 rules have been developed by a community of scholars to help guide thoughts about computing artifacts which in this case are taken to be the result of the development of attack infrastructure. (Miller, Keith, Moral Responsibility for Computing Artifacts: Five Rules, Version 24). These 5 rules were developed to help guide how technological artifacts should be designed and used. In this analysis attack infrastructure for offensive cyberspace operations are taken to fall into this category. In order to apply the 5 rules we assume that attack infrastructure for offensive cyberspace operations, and associated technologies and artifacts, are all sophisticated technological artifacts.

The first rule states, "The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artifact as part of a sociotechnical system." The application of this rule would be that the creators of the applications of the technologies related to attack infrastructure for offensive cyberspace operations need to be aware of the impact of their operations on users and that users should be aware of the impact of attack infrastructure technology and they are responsible for employing attack infrastructure technology in a socially responsible way.

The second rule of the 5 states, “The shared responsibility of computing artifacts is not a zero-sum game. The responsibility of an individual is not reduced simply because more people become involved in designing, developing, deploying or using the artifact. Instead, a person’s responsibility includes being answerable for the behaviors of the artifact and for the artifact’s effects after deployment, to the degree to which these effects are reasonably foreseeable by that person.” This second rule would state that those engaged in the development of artifacts employing attack infrastructure for offensive cyberspace operations should be responsible for the creation of this infrastructure and for what the application of this technology can accomplish, as well as for the social impact of the attack infrastructure employed for offensive cyberspace operations.

The fifth rule of the 5 rules is, “People who design, develop, deploy, promote, or evaluate a computing artifact ‘such as attack infrastructure for offensive cyberspace operations’ should not explicitly or implicitly deceive users about the artifact or its foreseeable effects, or about the sociotechnical systems in which the artifact is embedded.” One application of this rule would be that the designers of attack infrastructure should make it transparent to users and to members of society how applications have a social impact. It needs to be made clear to stakeholders how attack infrastructure is being used, and designers should provide a simplistic way for users to understand how the technology works. The purpose is to guide thoughts on attack infrastructure for offensive cyberspace operations and how these activities affect sociotechnical systems attacked by offensive cyber operations.

The conclusions that can be drawn from this ethical analysis are that 1) the organized group conducting offensive operations are doing so under the orders given by the policy level in a nation state. Rule-based nation states adhere to International law and the law of armed conflict. Next, 2) the character of personnel conducting those actions fit within a nation state’s predetermined set of criteria, excluding those who do not fit. Then 3), the intentions are those of the nation state. Finally, 4) the consequences of the payloads are those which have gone through a thorough internal national review. The biggest challenge is rule five, and deception. However, as mentioned above, the nation state directs the developer, and if this conflicts with the developer’s own moral compass, the developer is replaced.

8. Policy Recommendations

The conclusions of the ethical considerations form the basis of anticipatory ethical considerations/policy. It is anticipated that designing attack infrastructure for offensive operations require major collection efforts on human targets, on the target’s information systems, and infrastructure. Next, the organized group should demonstrate to policy makers how infrastructure is designed to increase their knowledge so they get more comfortable on this supports offensive operations.

9. Conclusions

The answer to the research question of ‘what design considerations should be taken when designing attack infrastructure for offensive operations?’ is noted in the framework development section. The ethical considerations are that developers and designers are morally responsible for the computer artefacts they generate, and their consequences. However, rule-based nation states adhere to international law and the law of armed conflicts. Therefore, people in the organized group that are designing attack infrastructure and developing payloads have legal support and policy support. The challenge is nation states who do not adhere to international law, the law of armed conflict, and employ proxies like cyber criminals, or patriotic hackers. Future research could include to develop an attack infrastructure in a virtual environment separated from the Internet. The environment can be used to train operators in the organized group, but also develop tools, tactics, techniques and procedures.

References

Alrwais, S. A., Dunn, C. W., Gupta, M., Gerber, A., Spatscheck, O., & Osterweil, E. (2012). Dissecting Ghost Clicks: Ad fraud via misdirected human clicks. ACM International Conference Proceeding Series, 21–30. https://doi.org/10.1145/2420950.2420954.

Antonatos, S., Akritidis, P., Lam, V. T., & Anagnostakis, K. G. (2008). Puppetnets: Misusing web browsers as a distributed attack infrastructure. ACM Transactions on Information and System Security, 12(2), 1–38. https://doi.org/10.1145/1455518.1455524.

Arntz, P. (2016). Explained: Domain Generating Algorithm. Retrieved from https://blog.malwarebytes.com/security-world/2016/12/explained-domain-generating-algorithm/.

Banach SJ and Ryan A (2009) The art of design: A design methodology. apps.dtic.mil. Available at: https://apps.dtic.mil/docs/citations/ADA518192.

Berti-Equille, L., & Zhauniarovich, Y. (2017). Profiling DRDoS attacks with data analytics pipeline. International Conference on Information and Knowledge Management, Proceedings, Part F1318, 1983–1986. https://doi.org/10.1145/3132847.3133155.

Bird J (2015) NATO's Role in Counter-Terrorism. *Perspectives on Terrorism* 9(2). Available at: http://www.terrorismanalysts.com/pt/index.php/pot/article/view/419/html (accessed 1 September 2019).

Chen Q and Bridges RA (2017) Automated Behavioral Analysis of Malware: A Case Study of WannaCry Ransomware. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, December 2017, pp. 454–460. ieeexplore.ieee.org.

Cheshire, S., Krochmal, M. (2013). Special-Use Domain Names. Retrieved from https://tools.ietf.org/html/rfc6761.

Chiba, D., Akiyama, M., Yagi, T., Hato, K., Mori, T., & Goto, S. (2018). DomainChroma: Building actionable threat intelligence from malicious domain names. Computers and Security, 77, 138–161. https://doi.org/10.1016/j.cose.2018.03.013.

Collier, B., Clayton, R., Thomas, D. R., & Hutchings, A. (2019). Booting the booters: Evaluating the effects of police interventions in the market for denial-of-service attacks. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, 50–64. https://doi.org/10.1145/3355369.3355592.

Connolly, L., Lang, M., & Tygar, J. D. (2015). ICT Systems Security and Privacy Protection. IFIP Advances in Information and Communication Technology, 455, 283–296. https://doi.org/10.1007/978-3-319-18467-8.

Cotton, M., Vegoda, L. (2010). Special Use IPv4 Addresses. Retrieved from https://tools.ietf.org/html/rfc5735.

Dimmock, J. (2017). Designing Effective Covert Red Team Attack Infrastructure. Retrieved from https://bluescreenofjeff.com/2017-12-05-designing-effective-covert-red-team-attack-infrastructure/.

Dimmock, J., & Borosh, S. (2018). Red Team Infrastructure Wiki. Retrieved from https://github.com/bluscreenofjeff/Red-Team-Infrastructure-Wiki.

Feroze, R. (2018). RedTeaming from Zero To One – Part 1. Retrieved from https://payatu.com/redteaming-from-zero-to-one-part-1.

Furrow, D. (2005). Ethics (Key Concepts in Philosophy). Bloomsbury.

Galinkin E, Jenko Hwong AS, Estep C, et al. (n.d.) The Future of Cyber Attacks and Defense is in the Cloud. Available at: https://www.researchgate.net/profile/Raymond_Canzanese/publication/336592029_The_Future_of_Cyber_Attacks_and_Defense_is_in_the_Cloud/links/5da78744299bf1c1e4c82fb0/The-Future-of-Cyber-Attacks-and-Defense-is-in-the-Cloud.pdf (accessed 27 January 2020).

Hrad, O., & Kemppainen, S. (2016). Honeypot utilization for analyzing cyber attacks. In Proceedings of the 10th European Conference on Software Architecture Workshops - ECSAW '16 (Vol. 222, pp. 1–2). New York, New York, USA: ACM Press. https://doi.org/10.1145/2993412.2993415.

Huskaj, G., Wilson, R.L., (2020). An Anticipatory Ethical Analysis of Offensive Cyberspace Operations. Proceedings of 15th International Conference on Cyber Warfare and Security (ICCWS), Norfolk, Virginia, U.S.A.

Iftimie IA (2019) Cyber Sanctions: The Embargo of Flagged Data in a Geo-Cultural Internet. In: *18th European Conference on Cyber Warfare and Security*, Reading, UK, 2019. Academic Conferences and Publishing International Limited.

Jacobsen JT (2019) NATOs offensive cyberspaceoperationer. Muligheder og udfordringer ved NATOs forespørgselsdrevne og effektbaserede tilgang. *Internasjonal Politikk* 77(3). tidsskriftet-ip.no: 241–251.

Kao D and Hsiao S (2018) The dynamic analysis of WannaCry ransomware. In: *2018 20th International Conference on Advanced Communication Technology (ICACT)*, February 2018, pp. 159–166. ieeexplore.ieee.org.

Karami, M., Park, Y., & McCoy, D. (2016). Stress testing the booters: Understanding and undermining the business of DDoS services. 25th International World Wide Web Conference, WWW 2016, 1033–1043. https://doi.org/10.1145/2872427.2883004.

Kim, N., Lee, S., Cho, H., Kim, B. I., & Jun, M. S. (2018). Design of a Cyber Threat Information Collection System for Cyber Attack Correlation. 2018 International Conference on Platform Technology and Service, PlatCon 2018, (2016), 1–6. https://doi.org/10.1109/PlatCon.2018.8472775.

Kohlenberg, T. (2018). Red Teaming. A Conference for Defence. Retrieved from https://www.youtube.com/watch?v=P4zIUQQo6Hg.

Krupp, J., Backes, M., & Rossow, C. (2016). Identifying the scan and attack infrastructures behind amplification DDoS attacks. Proceedings of the ACM Conference on Computer and Communications Security, 24-28-Octo(i), 1426–1437. https://doi.org/10.1145/2976749.2978293.

Leibowitz, M., & Timzen, T., (2019). Attack Infrastructure for the Modern Red Team. Retrieved from https://speakerdeck.com/tophertimzen/attack-infrastructure-for-the-modern-red-team.

Leontiadis, N., Moore, T., & Christin, N. (2014). A Nearly Four-Year Longitudinal Study of Search-Engine Poisoning Categories and Subject Descriptors. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 930–941. https://doi.org/10.1145/2660267.2660332.

Liu, D., Li, Z., Du, K., Wang, H., Liu, B., & Duan, H. (2017). Don't Let One Rotten Apple Spoil the Whole Barrel. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17 (pp. 537–552). New York, New York, USA: ACM Press. https://doi.org/10.1145/3133956.3134049.

Lu, L., & Lee, W. (2011). SURF : Detecting and Measuring Search Poisoning Categories and Subject Descriptors. Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS '11), 467–476. https://doi.org/10.1145/2046707.2046762.

Mezzour, G., Carley, K. M., & Carley, L. R. (2017). Global variation in attack encounters and hosting. ACM International Conference Proceeding Series, Part F1271, 62–73. https://doi.org/10.1145/3055305.3055306.

Miller, K., Moral Responsibility for Computing Artifacts: Five Rules, Version 24, IT Professional, 13(3):57 – 59, July 2011.

Mimura, M., & Tanaka, H. (2018). Long-Term Performance of a Generic Intrusion Detection Method Using Doc2vec. Proceedings - 2017 5th International Symposium on Computing and Networking, CANDAR 2017, 2018-Janua, 456–462. https://doi.org/10.1109/CANDAR.2017.

Mohurle S and Patil M (2017) A brief study of wannacry threat: Ransomware attack 2017. *International Journal of Advanced Research in Computer Science* 8(5). International Journal of Advanced Research in Computer Science. Available at: https://sbgsmedia.in/2018/05/10/2261f190e292ad93d6887198d7050dec.pdf.

Mudge, R. (2014). Infrastructure for Ongoing Red Team Operations. Retrieved from https://blog.cobaltstrike.com/2014/09/09/infrastructure-for-ongoing-red-team-operations/.

Nahari, S. (2017). Red Team Insights on HTTPS Domain Fronting Google Hosts Using Cobalt Strike. Retrieved from https://www.cyberark.com/threat-research-blog/red-team-insights-https-domain-fronting-google-hosts-using-cobalt-strike/.

PwC., BAE Systems., (2017). Operation Cloud Hopper - Indicators of Compromise. Retrieved from https://www.pwc.co.uk/cyber-security/pdf/cloud-hopper-indicators-of-compromise-v3.pdf.

Scarfo, A. (2016). Domain Generation Algorithms - Why so effective?. Retrieved from https://umbrella.cisco.com/blog/2016/10/10/domain-generation-algorithms-effective/.

Sharma, H., & Singh, H. (2018). Hands-on red team tactics. Birmingham: Packt Publishing Ltd.

Stallings, W. (2006). Data and computer communications (8th ed.). Pearson Prentice Hall.

Vincent A (2019) Don't feed the phish: how to avoid phishing attacks. *Network Security* 2019(2). Elsevier: 11–14.

Applications of Epidemiology to Cybersecurity

Jessemyn Modini, Timothy Lynar, Elena Sitnikova and Keith Joiner
University of New South Wales, Canberra, Australia
j.modini@adfa.edu.au
t.lynar@adfa.edu.au
e.sitnikova@adfa.edu.au
k.joiner@adfa.edu.au

DOI: 10.34190/EWS.20.057

Abstract: Epidemiology provides a novel approach to understanding cybersecurity risk. It provides a systematic model for the analysis of likelihood, consequence, management and prevention measures. While current research exists on the analysis of individual cybersecurity risk factors, there is a significant research gap on the collective interaction of these risk factors and their impact on the risk of cybersecurity compromise. Effective cybersecurity risk management requires the estimation of the probability of infection, based on a comprehensive range of historical and environmental factors, including system or network configurations and characteristics. The application of epidemiology highlights two fundamental approaches to increasing the efficiency and potency of cybersecurity; the requirement for comprehensive analysis of all cybersecurity risk factors, not just specific network vulnerabilities or uses, and the requirement for a centralised reporting, monitoring and data centre for cybersecurity incidents to inform this analysis, and facilitate a collective community response to mitigating cybersecurity risks. This paper discusses these applications of epidemiology to cybersecurity, to highlight the importance of research which combines these macro and micro-level approaches to provide a definitive evaluation of cybersecurity risk.

Keywords: Epidemiology, Cybersecurity, Risk Factors, Epidemiological Applications, Epidemiologic Security Analysis

1. Introduction

This paper explores the novel approach of applying epidemiology to cybersecurity. Epidemiology provides a novel approach to understanding cybersecurity risk, and has a plethora of similarities and applications, as detailed in Figure 1.

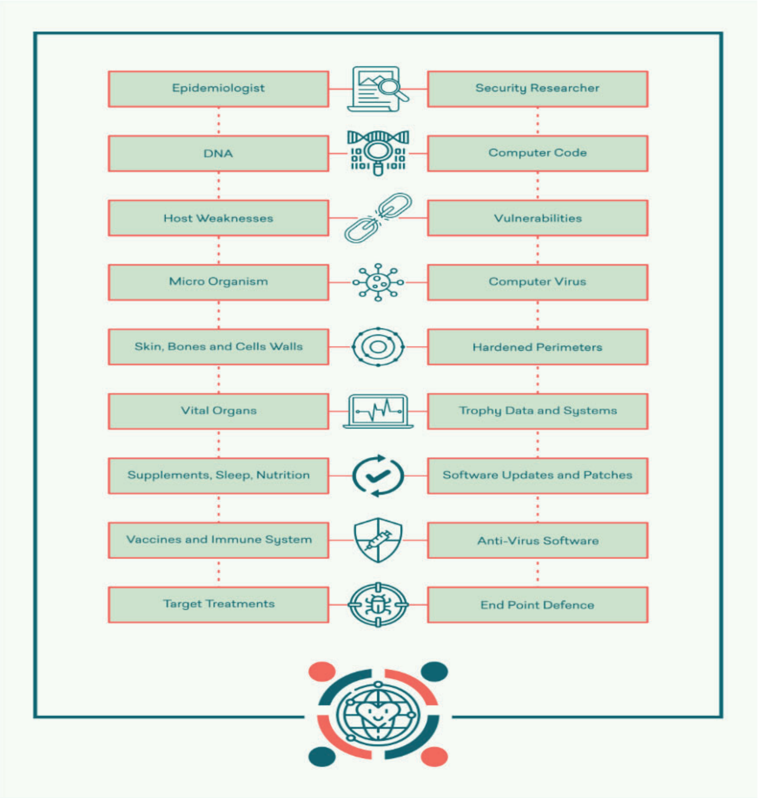


Figure 1: Thematic Applications of Epidemiology to Cyber Security

It provides a systematic model for the analysis of likelihood, consequence, management and prevention measures. This paper provides a review of literature, which identifies that whilst current research exists on the analysis of individual cybersecurity risk factors, there is a significant research gap on the collective interaction of these risk factors and their impact on the risk of cybersecurity compromise. This paper explores how epidemiological principles can be applied to estimate the probability of infection it based on a comprehensive range factors, and how a comprehensive repository of this data could be utilised to manage cyber security risk in the same way in which health is. This paper discusses these applications of epidemiology to cybersecurity, to highlight the importance of research which combines these macro and micro-level approaches to provide a definitive evaluation of cybersecurity risk.

2. Epidemiology and Cybersecurity

Historically as humans have aggregated into communities, the requirement for increasingly complicated infrastructure has emerged to help prevent the spread of disease. The concept of epidemiology emerged as a philosophical musing from Hippocrates, who earned his name as the "Father of Western Medicine" for "shifting the study of public health away from mysticism to patient-oriented empiricism" (Porter, 1999, p. 15) (Brantly, 2017). Epidemiology is defined as "the study of the distribution, and the determinants of disease frequency" (Rothman, 2012). As a discipline, it is focussed on "comparisons in order to establish cause-effect relationships, evaluate information, and make good decisions that will improve outcomes" (LaMorte, 2017). The core objectives of epidemiology are to identify the causes of a disease, the burden of the disease in a community, the history and impact of the disease, as well as the methods of prevention, management and response should an outbreak occur. Disease causes are determined by particular risk factors, which predetermine susceptibility to diseases. These risk factors predominantly comprise genetic profiles, and environmental factors (Gordis, 2013), but can also encompass elements such as nutrition. LaMorte (2017) contends that two key assumptions underpin epidemiology. The first of these is that "human disease does not occur at random; there are factors or determinants which can increase or decrease the likelihood of developing disease". The second of these assumptions is that "these causal and preventive factors, the determinants of disease, can be identified through systematic investigation of populations". This assumption can be contrasted to cybersecurity with ease, by replacing 'human disease' with 'computer disease or compromise', where 'populations' refers to networks or devices with cyber-physical features. Epidemiology is fundamentally a way of thinking and a means of problem solving that can be expanded outside the traditional application of public health and applied to the study of cybersecurity.

Cybersecurity events are rapidly increasing in prevalence and can have devastating consequences; on-par with a large-scale 'disaster'. Epidemiology can be used to monitor for and prepare for 'disasters' as a tool to assess and minimise vulnerabilities. According to Novick (2005, p. 162) "epidemiologic analysis can help in allocating limited resources to obtain maximal benefits in disaster situations". David Heyman, director of the homeland security program at the Center for Strategic and International Studies, a Washington DC-based research organisation, stated, ''we have a great sense of vulnerability, but no sense of what it takes to be prepared'' (Lipton, 2005). This disturbance was true in 2005 and is still accurate today and is directly applicable to the concept of a cybersecurity disaster, where an event causes a serious disruption over a short amount of time.

Considerable research has been invested in methods for the detection and aggregation of software vulnerabilities; however, there is a significant lack of research into the higher-level determination of actual risks and probability of compromise (Gil, Kott and A. L. Barabási, 2014) (Scarfone and Mell, 2008)(Frigault *et al.*, 2008). A consensus exists in the cybersecurity scholar community that while there is significant attention placed upon node and link network vulnerabilities, there is "limited systematic understanding of the factors that determine the likelihood that a node (computer) is compromised" (Gil, Kott and A. L. Barabási, 2014). While sophisticated tools exist for vulnerability detection, these tools are artificially limited by the threat intelligence, definitions and signatures to which they have access. They fundamentally rely on heuristic metrics, to evaluate risks of compromise. Effective cybersecurity risk management requires the estimation of probabilities of infection, based on the system or network configurations and characteristics (Yang, Holsopple and Sudit, 2007). A systematic approach to understanding the risk of compromise will inform strategic resource management for efficiency gains. An approach such as this bypasses existing assumptions, and as such, ''agnostic'', data-driven approaches are particularly valuable for making predictions in real-world scenarios, as they bypass the need for mechanistic explanations and contextual knowledge" (Gil, Kott and A.-L. Barabási, 2014).

In 1983, the technical virus was named after the convention of a biological virus, and defined as:

> *"a program that can 'infect' other programs by modifying them to include a possibly evolved copy of itself. With the infection property, a virus can spread throughout a computer system or network using the authorizations of every user using it to infect their programs. Every program that gets infected may also act as a virus and thus the infection grows"* (Cohen, 1987).

From this, biological themed cybersecurity terminology has emerged through terms such as 'worms' as a form of standalone malware which replicates itself to spread across a network.

The impact of cybersecurity incidents can be measured using epidemiological terminology, through "prevalence" and "cost of illness", where prevalence is the "number of existing cases of a disease in a population at a given time" (Pirc et al., 2016) and cost of illness is likened to the cost for remedy including lack of productivity, costs of replacing hardware, software, potential reputational damage, etc. These parallels can be utilised to provide a systematic framework for the application and analysis of disease causes, spread and consequence, which can then be assessed to inform effective prevention and management methodologies.
One of the largest challenges faced by cybersecurity experts is the inherent diversity of factors involved in attacks. These exist across a broad spectrum from hardware and software to human behaviour and social engineering. A novel approach is required to analyse these divergent elements and to provide a systematic approach for cyber-security risk management. Applications of epidemiology to cybersecurity are called out loudly in the academic community, as a critical gap in much need of analysis. Camp et al. (2019) specifically note the "need for a "cyber epidemiology" that treats individuals as highly distinct, independent, and important agents within a socio-technical system", and "advocate an approach to understanding how cybercrime thrives due to a failure to develop the understanding needed for effective behavioural control measures that are presented at the right place and the right time" (Camp et al., 2019). This research is critical to inform effective and efficient prevention, detection and incident response for cybersecurity attacks.

3. Technical Applications of Epidemiology to Cybersecurity

There is a multitude of technical applications of epidemiology; predominantly focussed on the spread of malware type 'diseases' through software and hardware collectives. The concept of the internet as a biological system first emerged in 1999 (White, Pagurek and Bieszczad, 1999), and was then specialised for the use case of email (Newman, Forrest and Balthrop, 2002). Given the inherent similarities between the "propagation of pathogens (viruses and worms) on computer networks and the proliferation of pathogens in cellular organisms (organisms with genetic material contained within a membrane- encased nucleus)" (Goel & Bush, 2004, p49), many parallels can be identified between the spread of biological, and technical diseases. This research has spurred a range of studies that apply the spread of biological pathogens to replicate and analyse how viruses spread through computer networks.

Biological networks can inform the understanding of how certain structures affect the spread of an epidemic (Stegehuis, van der Hofstad and van Leeuwaarden, 2016). Existing literature explores the analysis of network host vulnerability factors, malware and virus spread across peer-to-peer networks and Wi-Fi routers. This research has been conducted within recent decades and is largely focussed on specific applications. The community structure of a network is identified as a key factor for predicting how events flow across them. In a study that modelled complex networks with community structures, Stegehuis et al. (2016) were able to "capture the behaviour of epidemics or percolation on real-world networks accurately" and found that "mesoscopic community structure is vital for understanding epidemic spreading". Community structures are, therefore, critical factors for enforcing or inhibiting percolation of disease.

In an analysis of network traffic on a University Network, Uri and Jaladhanki (2017) was able to apply general epidemiological principles to "simulate the spread of infection on the dynamic bipartite graph inferred from observed external and modelled unobserved internal web-browsing traffic and evaluate the susceptibility of URI nodes to threats initiated by random clients and clients from specific countries". Higher rates of infection were identified for client nodes, compared to servers, "with maximum rates achieved when an infection is initiated randomly" (Uri and Jaladhanki, 2017). Further to this, it was found that infection rates were not proportionate to the intensity of traffic. A short list of 'most influential countries' was also identified, highlighting a degree of estimated pre-defined potency for each attack.

The biological concept of immune systems can also be applied to cybersecurity. This application was identified by Kephart, and based on factors comprising "(A) periodic monitoring of a data processing system (10) for anomalous behaviour that may indicate the presence of an undesirable software entity such as a computer virus, worm, or Trojan Horse; (B) automatic scanning for occurrences of known types of undesirable software entities and taking remedial action if they are discovered; (C) deploying decoy programs to capture samples of unknown types of computer viruses; (D) identifying machine code portions of the captured samples which are unlikely to vary from one instance of the virus to another; (E) extracting an identifying signature from the executable code portion and adding the signature to a signature database; (F) informing neighbouring data processing systems on a network of an occurrence of the undesirable software entity; and (G) generating a distress signal, if appropriate, to call upon an expert to resolve difficult cases" (*United States Patent (19) Arnold et al.*, 1993). In his early research collaborations, Kephart (1995) used epidemiological approaches to model a "neural network virus detector that learned to discriminate between infected and uninfected programs, and a computer immune system that identified new viruses, analysed them automatically, and used the results of its analysis to detect and remove all copies of the virus that were present in the system". In contemporary hindsight, this approach is effectively an Intrusion Detection System (IDS) or Intrusion Prevention System (IPS) implementation. As artificial immune systems effectively comprise a range of detection and response elements that can identify threat signatures and respond to them to 'neutralise' or decontaminate the network from the threat. IDS and IPS are still based on perimeter defences, which unfortunately are not reliably effective at mitigating threats. Goel and Bush (2004) contend that "it is imperative to complement these existing security models with reactive systems that can detect new strains of pathogens rigorously and can destroy them before they can cause damage and propagate further". The authors argue that "reactive models provide a scalable, resilient, and cost-effective mechanism that may keep pace with constantly evolving security needs". In order to achieve this, however, there must be a detailed understanding of the network, to enable collective defence.

The hardware and software configuration of hosts on a network can also be utilised to determine their potential vulnerability to cybersecurity threats. In a paper titled "A genetic epidemiology approach to cyber-security", Gil et al. (2014) identify that through a technical analysis of network hosts, a statistical relationship can be identified between specific host services, and vulnerability to threats (Gil, Kott and A. L. Barabási, 2014). This framework allows the determination of "probabilities of infection directly from observation", that offers an "automated high-throughput strategy to develop comprehensive metrics for cyber-security"(Gil, Kott and A. L. Barabási, 2014). The authors utilised network hosts and patter behaviour as a specific case study for the utilisation of epidemiological-based approaches and found this type of approach highly accurate and meaningful.

In medical epidemiology, genetic variants can be analysed to pre-determine susceptibility and estimation of health-related risks for individuals. A parallel application of this can be used to "establish associations between the network services a host is running and the kinds of threats to which it is susceptible", through the correlation of IPS and IDS threat log data, and threat activity for hosts within a network (Gil, 2016, p. 89). This application supports a model for a highly accurate estimation of infection risk, based on existing correlation data. Gil (2016, p. 90) notes that while there is a prevalence of research on modelling relationships between software security vulnerabilities and methods of compromise, however "studies in which actual probabilities of compromise are derived and compared with threat data in the wild are largely absent from the literature". It is also noted that due to the inherent complexities and the numerous factors involved in cybersecurity compromises, there are challenges in attaining large-scale data models that have been fundamental in understanding the 'underpinnings' of complex diseases. This challenge supports the requirement for a centralised reporting and data repository for cybersecurity incidence, reflecting a similar model to the U.S. Centre for Disease Control (CDC). Another similar model is the Financial Services Information Sharing and Analysis Centre (FS-ISAC) which is used as a "dedicated forum for business, trade and commerce to reduce cyber risk in the global financial system"(Ernst and Young, 2019). The framework utilised by this study could be scaled to assess other network or system characteristics and configuration items as risk factors, to determine the impact they have on risk. These configuration items could include applications, hardware and devices, accounts, privilege and administrator rights and levels, drivers, operating system, and manufacturer (Yang *et al.*, 2019). It is noted that significant data is required to expand the study to encompass these factors. Additionally, a range of data is required to assess how these elements interact as risk factors, to determine the aggregate risk of a system or network being compromised, rather than the risk of a specific element being compromised.

In a study of peer-to-peer virus propagation, Ebrahim, Khan and Khalid (2014) utilise epidemiological approaches defined by a key set of parameters comprising damage level, distribution level, the difficulty of removal, threat containment, geographical distribution, count of source sites, count of infections and wild level. Epidemiology has been used to create mathematical models for virus propagation modelling. This approach has two key models; the "susceptible-infected-susceptible (SIS) model" or the "susceptible-infected-removed (SIR) model". A third, less popular model exists, called the "susceptible-infected-exposed-removed (SIER)" model. These models are defined as (Ebrahim, Khan and Bin Khalid, 2014), (1) the SIS model, where "each host stays in one of two states: susceptible or infectious. Each susceptible host becomes an infectious one at a certain rate. At the same time, infectious hosts are cured and become again susceptible at a different rate. This model system where having the infection and being cured does not confer immunity", the SIR model, where "a node cannot be infected more than once" and the SIER model, where "a node cannot be infected more than once which cause an individual to be able to infect others immediately upon their infection".

Biological paradigms, including genomics (RNA interference), proteomics (pathway mapping), and physiology (immune system) can be used to study the spread of pathogens, across and within networks (Goel and Bush, 2004). In similar studies, other scholars have effectively modelled virus propagation in peer-to-peer networks, using similar epidemiological based approaches (Thommes and Coates, no date) (Ramachandran and Sikdar, 2006) (Zhou *et al.*, no date)(Feng *et al.*, 2008). Epidemiological methods have also been used to assess the spread across Internet of Things (IoT) hosts, due to the "inadequate security design of IoT devices and the obstacles associated with improving the overall security of the IoT" (Jerkins and Stupiansky, 2018). The authors found that as "IoT malware is often transmitted from host to host similar to how biological viruses spread in populations", "biological viruses and computer malware may exhibit epidemic characteristics when spreading in populations of vulnerable hosts" (Jerkins and Stupiansky, 2018). The concept of inoculation of key hosts was prototyped to demonstrate the parallels between select immunity and disease propagation. In a similar approach to biological vaccination development, Jerkins and Stupiansky (2018) explore whether infection vectors can be identified, "either from a known vulnerability in IoT hosts or by reverse engineering samples of IoT malware". From this, a "neutered, or benign variant, of the malware is created that exploits the identified infection vector"(Jerkins and Stupiansky, 2018) which can then be used to teach the infected devices to build immunity to the malware variant.

Epidemiology can also be applied to mathematical modelling of generalised virus propagation across networks. The SLBS model is a mathematical formula for the spread of computer viruses across a network, developed by Yang and Yang (2012). It captures the main elements, including uninfected internal computers, latent internal computers, breaking-out internal computers, uninfected external computers, latent external and breaking-out external computers (Yang and L.-X. Yang, 2012). These elements are quantified by assumptions based around connectivity, susceptibility, and probability factors. Like all mathematical formulas, it is general and based on fully connected networks. Therefore, it cannot effectively capture "the effect of the topological structure of the Internet on the spread of computer viruses" (Yang and L. X. Yang, 2012), nor evaluate the disease spread at any level of detail below computers as risk factors. Pirc et al. (2016) argue that threat forecasting is a mixture of science, mathematics and technology. Therefore, the successful application of epidemiology to cybersecurity will be underpinned by mathematical rigour such as this, and a strong understanding of technology.

Epidemiology has been applied to WiFi Routers for spread analysis. Hu et al. (2009, p. 1318) found that "in densely populated urban areas WiFi routers form a tightly interconnected proximity network that can be exploited as a substrate for the spreading of malware able to launch massive fraudulent attacks". Using concepts of epidemiology, the authors simulated a 'contagion' event of malware spread and estimated that tens of thousands of routers would be 'infected' within 24-48 hours to two weeks. (Hu *et al.*, 2009). Mitigations were also modelled, using an epidemiological framework, to determine "possible containment and prevention measures" as well as "computational estimates for the rate of encrypted routers that would stop the spreading of the epidemics by placing the system below the percolation threshold".

In an analysis of cybersecurity simulations, using cognitive modelling, Veksler et al. (2018) cite that "cognitive modelling can address multi-disciplinary cybersecurity challenges requiring cross-cutting approaches over the human and computational sciences", through "dynamic simulations involving attacker, defender, and user models to enhance studies of cyber epidemiology and cyber hygiene" (Veksler *et al.*, 2018). The authors highlight the importance of simulation in both health care and cybersecurity, noting that "just as simulations in healthcare predict how an epidemic can spread and how it can be contained, such simulations may be used in the field of cyber-security as a means of progress in the study of cyber-epidemiology" (Veksler *et al.*, 2018). The authors contend that epidemiology usefully offers simulation to inform prediction models based on existing behavioural data of threats. These “process models” assist in simulating “realistic synthetic users for full-scale training/wargame scenarios”, and also “enable much-needed research in cybersecurity and cyber-epidemiology” (Veksler *et al.*, 2018).

4. Applications – Public Health

At the macro level, epidemiology highlights the pertinence of applying the concept of shared responsibility and accountability for cyber health. Cybersecurity should be treated as a public health issue. Given the “all-pervasive nature of the ever interconnected economy and the exponential nature of cybercrime, there is a growing need for cybersecurity to be treated holistically, by both the public and private sector, in a similar fashion to public health issues” (University of Queensland, 2019). In their recent response to Australia’s 2020 Cybersecurity Strategy, Oracle (2019) contended that the establishment of “shared situational awareness of the health and functioning of the Internet is critical to alerting to any deviations to the norm” and that this insight can inform Government and industry so that they “can take appropriate action to address malicious activities that might disrupt the normal function of the global internet”. Brantly (2019, p. 90) explores “big data analysis, public health modelling and disease prevention, tracking, and remediation to examine the current state of the field of cybersecurity and assess where new technologies and policies are leading in terms of solving significant national security issues presented by cyberspace.” In this article, Brantley explores how malware spread replicates biological pathogens, and therefore how the field of epidemiology can provide a backbone of data, dating back hundreds of years to support a contemporary analysis of cybersecurity. Brantley notes that “few analyses have fully leveraged the rigour of public health and epidemiologic study as a guide for cybersecurity”, despite it being a relatively comparable fit for cross-disciplinary analysis. Given the complexity of cybersecurity threats, “public health and epidemiological (EPI) modelling are useful for comparative purposes because they focus on not only the forensic biological, genetic, or molecular attributes of pathogens, but also on a combination of factors that include the human/societal behavioural attributes, environmental factors, and systemic modelling of incidents in the assessment, treatment, prevention, and mitigation of disease.”(Brantly, 2017).

Epidemiology supports the concept of a ‘health record’ for systems, and systems of systems. In their paper “Toward a framework for assessing the cyber-worthiness of complex mission critical systems”, Fowler and Sitnicova (2019) use aggregated cyber vulnerability risk modelling to a generic threat set to assess and then sustain through-life and in presence of threat evolution the cyber-worthiness of a system-of-systems. This form of assessment capability could be furthered, using epidemiological approaches to define systematic health checks, and consolidate a cybersecurity health record. These health records could be maintained by the risk owner and utilised to assess and manage risk through quantitative metrics.

Epidemiology has also been applied to understand the spread and susceptibility of social engineering attacks such as phishing. Camp et al. (2019) draw on experience in the evaluation of multi-national phishing campaign evaluation, to define an epidemiology-based approach to computer security health, through an analysis of “cultural, logistic, and regulatory challenges” to the “proposed public health approach to global computer assault resilience”. The authors contend that behavioural controls can be identified and analysed, to determine the likelihood and consequence of complex cyber-security attacks, once the general population is understood sufficiently to determine risk factors. Using epidemiology as a framework could identify a “holistic, ecologically valid approach to engendering resilience and understanding location-specific vulnerability to social engineering attacks” (Camp *et al.*, 2019).

5. Challenges for Public Health-based Cybersecurity

World health reporting is mandatory for centralised monitoring and coordination of joint response to threats. While there are numerous barriers including privacy regulations, communications exchange, multiple jurisdictions (i.e. civilian and military), authority and liability issues; mandatory cybersecurity reporting would establish a centralised unit, to enforce shared responsibility and resourcing efficiencies. A centralised, public health-based approach to cybersecurity will require a high degree of standardisation. International standards would provide “widely vetted, consensus-based frameworks for defining and implementing effective approaches to cyber security”, while also “facilitating common approaches to common challenges, thus enabling collaboration and interoperability” (IBM, 2019). A public health approach to cybersecurity resilience would require strong international cooperation, to most effectively prioritise resources to protect against cyber threats. For this approach to be implemented successfully, a large amount of current and historical data must be shared. From this, patterns and relationships can be analysed for a comprehensive and systematic analysis of cybersecurity risk.

Within the field of epidemiology, new concepts of ‘digital epidemiology’, or ‘health surveillance’ are emerging as a result of the prevalence of online data sources. Technical or digital epidemiology is a newly coined term, referring to a “new field that has been growing rapidly in the past few years, fuelled by the increasing availability of data and computing power, as well as by breakthroughs in data analytics methods” (Salathé, 2018). This surveillance intends to “detect health or security threats worldwide, in real-time, rooted in the mining of online data, including personal data from social media and even information on health behaviours and health attitudes” (Samerski, 2018). In their article “Digital disease detection: A systematic review of event-based internet biosurveillance systems”, O’Shea (2017) explores how internet connectivity has modified the way people report disease and contends that “event based Internet biosurveillance should act as an extension of traditional systems, to be utilised as an additional, supplemental data source to have a more comprehensive estimate of disease burden”. Essentially, digital epidemiology is a modern approach to epidemiology focussed on digital data that is always generated with the primary purpose of being utilised for epidemiology. This data is generally sourced from outside the health system, including internet trends, and social media. An example of this is the analysis of “symptomatic search queries for the purpose of syndromic tracking of influenza-like illnesses (Ginsberg et al., 2009)”. Issues with the ownership of this information however has blocked the development of an effective algorithm for tracking these events to correlate meaningful assertions (Salathé, 2018). The challenges faced by this approach have similarities to challenges that would face the concept of a centralised cybersecurity repository, due to similar issues with data sensitivities.

6. Conclusion and future work

This paper has discussed the approach of epidemiology to cybersecurity and highlighted the importance of research that combines both macro and micro-level approaches to provide a definitive evaluation of cybersecurity risk. This paper explores how epidemiology frameworks support a systematic model for the analysis of likelihood, consequence, management and prevention measures, by using cyber features as risk factors. While current research exists on the analysis of individual cybersecurity risk factors, there is a significant research gap on the collective interaction of these risk factors and their impact on the risk of cybersecurity compromise. Effective cybersecurity risk management requires the estimation of the probability of infection, based on a comprehensive range of historical and environmental factors, including system or network configurations and characteristics.

This paper has demonstrated that epidemiology offers two novel approaches for increasing the efficiency and potency of cybersecurity; through the comprehensive analysis of all cybersecurity risk factors, not just specific network vulnerabilities or uses, and the benefits of centralised reporting, monitoring and data collection for cybersecurity incidents to inform cybersecurity risk analysis and collective community preparedness and response for the mitigation of cybersecurity risks.

Our research explores the comprehensive application of these epidemiological concepts to cybersecurity, including analysis of the relationship between all cyber-relevant features as ‘risk factors’, to model the likelihood and consequence of cyber security compromise or ‘infection’, and the utilisation of an aggregate dataset to inform quantitative risk assessment and IPS/IDS methods.

References

Brantly, A. F. (2017) ‘Public health and epidemiological approaches to national cybersecurity: A baseline comparison’, *US National Cybersecurity: International Politics, Concepts and Organization*, pp. 91–109. doi: 10.4324/9781315225623.

Camp, L. J. *et al.* (2019) ‘Measuring Human Resilience in the Face of the Global Epidemiology of Cyber Attacks’, in *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. doi: 10.24251/hicss.2019.574.

Cohen, F. (1987) ‘Computer viruses. Theory and experiments’, *Computers and Security*, 6(1), pp. 22–35. doi: 10.1016/0167-4048(87)90122-2.

Ebrahim, M., Khan, S. and Bin Khalid, U. (2014) ‘Security risk analysis in peer 2 peer systems: an approach towards surmounting security challenges’, *Asian Journal of Engineering, Sciences and Technology*, 2(2), pp. 94–101. Available at: http://arxiv.org/abs/1404.5123%0Ahttps://arxiv.org/ftp/arxiv/papers/1404/1404.5123.pdf.

Ernst and Young (2019) *EY's view to Australia's Cyber Security Strategy 2020*. Available at: https://www.homeaffairs.gov.au/reports-and-pubs/files/cyber-strategy-2020/submission-173.pdf.

Feng, C. *et al.* (2008) 'Propagation Model of Active Worms in P2P Networks'. doi: 10.1109/ICYCS.2008.237.

Fowler, S. and Sitnikova, E. (2019) 'Toward a framework for assessing the cyber-worthiness of complex mission critical systems', in *2019 Military Communications and Information Systems Conference, MilCIS 2019 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/MilCIS.2019.8930800.

Frigault, M. *et al.* (2008) *Measuring Network Security Using Dynamic Bayesian Network*.

Gil, S., Kott, A. and Barabási, A. L. (2014) 'A genetic epidemiology approach to cyber-security', *Scientific Reports*. Nature Publishing Group, 4. doi: 10.1038/srep05659.

Gil, S. and Scientific, S. (2016) 'Network services as risk factors: A genetic epidemiology approach to cyber security', *Dynamic Networks and Cyber-Security*, 1, pp. 89–109. doi: 10.1142/9781786340757_0004.

Goel, S. and Bush, S. F. (no date) *Biological Models of Security for Virus Propagation in Computer Networks*.

Gordis, L. (2013) *Epidemiology*. 5th edn. Philadelphia: Elsevier.

Hu, H. *et al.* (2009) 'WiFi networks and malware epidemiology', *Proceedings of the National Academy of Sciences of the United States of America*, 106(5), pp. 1318–1323. doi: 10.1073/pnas.0811973106.

IBM (2019) *IBM Submission to: Australia's Cyber Security Strategy 2020*. Available at: https://www.homeaffairs.gov.au/reports-and-pubs/files/cyber-strategy-2020/submission-96.pdf.

Jerkins, J. A. and Stupiansky, J. (2018) 'MItigating IoT insecurity with inoculation epidemics', in *Proceedings of the ACMSE 2018 Conference*. Association for Computing Machinery, Inc. doi: 10.1145/3190645.3190678.

LaMorte, W. (2017) *The Evolution of Epidemiologic Thinkingt*, *Boston University School of Public Health*. Available at: http://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_history/EP713_History9.html (Accessed: 28 October 2019).

Newman, M. E. J., Forrest, S. and Balthrop, J. (2002) 'Email networks and the spread of computer viruses', *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 66(3). doi: 10.1103/PhysRevE.66.035101.

Novick, L. F. (2005) 'Epidemiologic approaches to disasters: Reducing our vulnerability', *American Journal of Epidemiology*, pp. 1–2. doi: 10.1093/aje/kwi164.

O'shea, J. (2017) 'Digital disease detection: A systematic review of event-based internet biosurveillance systems', *International Journal of Medical Informatics*, 101, pp. 15–22. doi: 10.1016/j.ijmedinf.2017.01.019.

Oracle (2019) 'Oracle's Comments on Australia's 2020 Cyber Security Strategy'. Available at: https://www.homeaffairs.gov.au/reports-and-pubs/files/cyber-strategy-2020/submission-104.pdf.

Ramachandran, K. and Sikdar, B. (2006) 'Modeling malware propagation in Gnutella type peer-to-peer networks', in *20th International Parallel and Distributed Processing Symposium, IPDPS 2006*. IEEE Computer Society. doi: 10.1109/IPDPS.2006.1639704.

Rothman, K. (2012) *Epidemiology: An Introduction*. New York: Oxford University Press.

Salathé, M. (2018) 'Digital epidemiology: what is it, and where is it going?', *Life Sciences, Society and Policy*. SpringerOpen, 14(1). doi: 10.1186/s40504-017-0065-7.

Samerski, S. (2018) 'Individuals on alert: digital epidemiology and the individualization of surveillance', *Life Sciences, Society and Policy*. SpringerOpen, 14(1). doi: 10.1186/s40504-018-0076-z.

Scarfone, K. and Mell, P. (2008) 'Vulnerability scoring for security configuration settings', in *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 3–7. doi: 10.1145/1456362.1456365.

Stegehuis, C., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2016) 'Epidemic spreading on complex networks with community structures', *Scientific Reports*, 6(1), p. 29748. doi: 10.1038/srep29748.

Thommes, R. W. and Coates, M. J. (no date) *Modeling Virus Propagation in Peer-to-Peer Networks*.

United States Patent (19) Arnold et al. 54 AUTOMATIC IMMUNE SYSTEM FOR (1993).

University of Queensland (2019) *The University of Queensland's submission to Australia's Cyber Security Strategy 2020*. Brisbane.

Uri, D. and Jaladhanki, S. (2017) 'Epidemiology of Browser-Based Malware'.

Veksler, V. D. *et al.* (2018) 'Simulations in cyber-security: A review of cognitive modeling of network attackers, defenders, and users', *Frontiers in Psychology*, 9(MAY), pp. 1–12. doi: 10.3389/fpsyg.2018.00691.

White, T., Pagurek, B. and Bieszczad, A. (1999) 'Network modeling for management applications using intelligent mobile agents', *Journal of Network and Systems*, 7(3), p. 295321.

Yang, S. J., Holsopple, J. and Sudit, M. (2007) 'Evaluating threat assessment for multi-stage cyber attacks', in *Proceedings - IEEE Military Communications Conference MILCOM*. doi: 10.1109/MILCOM.2006.302216.

Yang, X. *et al.* (2019) 'A genetic epidemiology approach to cyber-security', *Scientific Reports*. IEEE, 4(2), pp. 1–7. doi: 10.1038/srep05659.

Yang, X. and Yang, L.-X. (2012) 'Towards the Epidemiological Modeling of Computer Viruses', *Discrete Dynamics in Nature and Society*. Hindawi Publishing Corporation, 2012, p. 11. doi: 10.1155/2012/259671.

Yang, X. and Yang, L. X. (2012) 'Towards the epidemiological modeling of computer viruses', *Discrete Dynamics in Nature and Society*, 2012. doi: 10.1155/2012/259671.

Zhou, L. *et al.* (no date) *A First Look at Peer-to-Peer Worms: Threats and Defenses*.

Computational Propaganda: Targeted Advertising and the Perception of Truth

Julie Murphy, Anthony Keane and Aurelia Power
School of Informatics and Engineering, Technological University Dublin, Ireland
Julie.murphy@mytudublin.ie
Anthony.keane@tudublin.ie
Aurelia.power@tudublin.ie

DOI: 10.34190/EWS.20.503

Abstract: Social media has become an effective medium for the execution of cyberpsychological threats by adopting language to influence perceptions based on personal interests and behaviours. Targeted messages can be refined for maximum effect and have been implicated in changing the outcome of democratic elections and the decreasing uptake of vaccinations. However, computational propaganda and cyberpsychological threats are not well understood within the cybersecurity community. To address this, we adopt the theoretical model of *the illusory truth effect* to posit that how information is presented online, may solidify views in an 'undecided' group with 'some' knowledge of an argument. We test this hypothesis by employing an explanatory sequential design. We first analyse a dataset containing adverts related to Brexit to determine influential terms using the corpus linguistics method. Analysing term frequencies, collocational and concordance information, the results of our quantitative analysis indicate that function words such as the personal pronouns 'we' or the definite article 'the' play a significant role in the construction of computational propaganda language. We then conduct a qualitative analysis of a Facebook ad related to Brexit to further understand how the 'who' and the 'what' elements are realised in computational propaganda language, that is, who is targeted and what is the underlying message. We found that understanding these, one can gain insights into a threat actor's motivation, opportunity and capability and, thus, allows a defensive response to be put into place. In turn, how an audience responds, may provide insight on the impact of the threat.

Keywords: computational propaganda, illusion of truth, cyberpsychology, critical discourse analysis, threat intelligence

1. Introduction

Influence comes under many guises, with terms such as propaganda or advertising used interchangeably to describe what is determined by the motivation of the influencer. Their enduring prevalence verifies the effectiveness of these techniques. This study considers the significance of influential techniques for the cybersecurity community by investigating why these tactics are effective and their amplified risk when propagated online. A sociocognitive critical discourse study is applied to determine what intelligence can be derived from these indicators as part of a broader framework.

Targeted advertising can create a direct communication with an audience predisposed to a certain view. Social media platform providers are perfectly situated to facilitate such communication based on pre-determined information surrounding interests and behaviours. Therefore if the message is designed for a specific target, it is feasible to reverse engineer the communication to determine what the underlying message is, who is being targeted and why. If traction and perception surrounding the message are also monitored; a threat actors capability, opportunity and motivation may be determined.

This study addresses the construct and delivery of a message text as an example for the cybersecurity community to highlight the importance of cyberpsychological skills to defend against influence threats. The objective is to determine why language usage and communication methods are relevant from a security perspective. Computational propaganda and cyberpsychological threats are fundamentally transdisciplinary, which can lead to translation issues for researchers crossing social and computer sciences. The challenge is how can cyberpsychological attacks be categorised for analytical research; how can digital communications compound the effect of cyberpsychological threats and how can the construct language of digital communications contribute to cyberpsychological threats?

2. Literature Review

2.1 The new social contract

Social contracts are a fundamental part of functioning society as historically debated by philosophers such as Thomas Hobbes and John Locke whereby freedoms are surrendered to authority for protection. Arguably, modern society has entered into a contract with online platform providers surrendering personal data for convenience and communication services. Attention economics is of increasing interest to providers whereby attention is a commodity in short supply therefore platform providers are refining techniques and information to retain attention longer to acquire more data, thus continuing the cycle. Similarly Bernays (1928) surmises that as the public become more aware of influential tactics, leaders present their objectives more intelligently. Subsequently vast digital footprints are accumulated with highly personal traits that support the propensity of targeted influence attacks. Significantly, attacks of this nature may have debatable legal implications, supported by relative anonymity rendering them highly desirable to a malicious actor adapting historically proven methods to contemporary mediums. Computational propaganda is now a global area of research with increasing interest in light of controversies ranging from vaccine debates to election tampering.

2.2 Computational propaganda

Howard and Wooley (2016) describe computational propaganda as “the use of algorithms, automation, and human curation to purposefully manage and distribute misleading information over social media networks”. Their extensive multi-year study considers research from all over the world to assert how computational propaganda is promoted and enacted by different political regimes and actors and the consequences for democracy. Nimmo (2019) proposes a computational method to identify the extent of traffic manipulation which benefits the cybersecurity communities in attack analysis. Techniques employed to render manipulation campaigns effective have been identified by Bradshaw and Howard (2017) that formalise propaganda techniques as applied online. These approaches consider the impact of computational propaganda in the context of democracy and how it is effected online, building on existing research to understand why the online environment increases potential future risk. Accordingly cyberpsychologists endeavour to determine how engagement with technology impacts behaviours and psychological states. Barton (2016) describes models of persuasion and their importance within an online environment detailing how psychological approaches such as compliance, obedience, conformity and persuasion are applied alongside methods such as captology to great effect. This study examines concepts surrounding the ‘illusory truth’ effect to determine why computational propaganda and cyberpsychological threats are of particular importance to the cybersecurity community on the basis that “If people are told something often enough, they’ll believe it” (Hasher, Goldstein and Toppino, 1977).

2.3 The Illusory Truth Effect

Hasher, Goldstein and Toppino (1977) conducted a study to assess whether participants believe a statement to be true if repeatedly exposed to it over a period of time. Using statements from a broad range of knowledge areas that were either true or false, they concluded that people were inclined to believe a repeated plausible statement, commonly referred to as the ‘illusory truth effect’. Similarly, Bacon (1979) posited that ‘recognition’ of statements aligned to pre-existing views are considered more valid then opposing statements. Moreover people believe what they know to be true therefore, if new information confirms this knowledge it is more readily accepted. Arkes, Hackett and Boehm (1989) expanded on these theories by testing statements of ‘opinions’ instead of ‘facts’, concluding that the validity of the repetition effect is diminished when the audience is not well versed in the subject matter. Similarly Ozubko and Fugelsang (2010) posited that memory retrieval, or increased familiarity with a statement enhances the persuasive effect. The online environment is primed to build familiarity with statements and concepts as online concepts often translate to more traditional mediums such as national press, television and radio.

Becoming familiar and ‘knowledgeable’ of facts and statements by persistent exposure however, re-enforced by repetition and recognition of opinions and statements online compounds the perception of truth. Accepting these theories, consider the vast digital footprint of people, available to platform providers that is accurately customised for newsfeeds, information and advertisements based on distinctive personality traits and behaviours. Is it not conceivable that a malicious actor may employ targeted, often legal strategies, to influence the beliefs and opinions of a particular demographic?

If people had to weigh up the risks and consequences, or read the terms and conditions surrounding their online activity, their time online would increase exponentially in parallel with a dramatic decrease in productivity. Thus when faced with thousands of decisions a day, people trust their instincts and those around them to make assumptions that the tools and platforms they use are safe. It is perceived that influence tactics are recognisable and defensible when educated accordingly however, cognitive ability does not defend against the illusory truth effect as demonstrated by De keersmaecker et al. (2019). In their study, aspects pertinent to influence were assessed: cognitive ability, the need for cognitive closure, and cognitive style. Their conclusions demonstrated that people are predisposed to believe repeated statements regardless of their cognitive ability. Therefore highly personal traits are available for sustained targeted influence campaigns to those who wish to express a perspective to an audience known to be 'susceptible' to the view. For the campaign to be successful, the statement or opinion needs to be framed in a manner that the targeted audience is likely to accept. Danesi (2015) describes the craft of advertising as "*a way of presenting something in a socially appropriate way…in the same way that effective orators attempt to convince audiences to accept their messages and act upon them*". Moreover, the characteristics of discourse are synonymous with keywords and cognitive styles that are familiar to people based on a 'shared knowledge system' or 'common ground'. Further, effective advertising builds an association of 'personal amelioration'. Importantly, Danesi asserts that although the delivery of advertising has changed with technological advances, the fundamental persuasive tactics have not.

Contagious ideas generally originate by the few. Consider fashion, whereby exceptional people lead to disproportionate influence and propagation of a trend. Propagation is dependent on certain conditions described by Gladwell (2006) as connectors, mavens and salesmen. Connectors have the network of relevant people, mavens are experts on the cutting edge espousing new information, and salesmen have the power to persuade people to make decisions they may not necessarily have taken. The combination or connection between these three elements results in powerful carriers of concepts, ideas and changes, often based on simple changes. There is limited research considering these perspectives through a cybersecurity lens.

In summary, a threat actor need only target those that are 'undecided' to affect the success or failure of any topic, product or service. What facilitates an ad, or topic's importance aligns with what is important to 'me' and is incorporated then into general conversation. Computational propaganda informs the beliefs that 'I' am susceptible to, therefore we suggest that a critical discourse study may highlight why the construct and delivery of a message impacts the perceptions of a susceptible audience. The cybersecurity community must understand 'how' a message can impact the perception of an undecided audience. There is intelligence is to be gathered by understanding why a particular audience is targeted and who would gain by influencing them. Incorporating this into the lifecycle of an attack would help support interventive deflection measures.

3. Methodology

The underlying problem is that computational propaganda and cyberpsychological threats are not well understood or addressed within the cybersecurity community. To address the interdisciplinary nature of this problem, we employ van Dijk's sociocognitive approach (2016). Because "*...social interaction, social situations and social structures can only influence text and talk through people's interpretations of such social environments ..[and] discourse can only influence social interaction models, knowledge, attitudes and ideologies*" (van Dijk 2016) we focus on critically analysing the online discourse surrounding computational propaganda. According to Wodack (2012) "*Critical discourse studies (CDS) are therefore not interested in investigating a linguistic unit per se but in analysing, understanding and explaining social phenomena that are necessarily complex and thus require a multidisciplinary and multi-methodical approach*"

3.1 Explanatory Sequential Design

The premise of this study is based on the hypothesis that: how information is presented online may solidify views in an 'undecided' group with 'some' knowledge of an argument. To carry out our investigation we employ an explanatory sequential design. We first analyse our dataset from a quantitative perspective, followed by a qualitative study to explain the results in greater depth (Creswell, 2015). The objective is to ascertain whether there are identifiable patterns in the way language is used in computational propaganda discourse that may be built upon for future defences against cyberpsychological threats. The focus is not on the ideation of the content but the 'construct' and delivery of the message

3.2 Quantitative Data Collection and Analysis

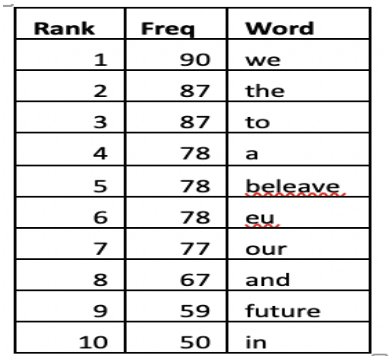
Using political adverts relating to Brexit, we analyse content from the BeLeave campaign to determine patterns surrounding the delivery and construct of online targeted ads. We use the corpus linguistics approach of analysing textual content to obtain statistical evidence of word frequencies, context and collocations. For example, at first glance, much of the language used appears relatively neutral, however when evaluated based on concordance and collocation evidence for example, other patterns may become apparent.

3.2.1 Dataset

Verified targeted ads were used as the basis to build a control dataset for future studies. 120 ads were selected from the BeLeave campaign ran in June 2016 (House of Commons, 2019). These ads achieved an impression[1] rate between 46,650,000 and 104,836,880 by the British electorate in the days prior to the Brexit referendum. 18% of these ads were slight variations of the same text. Duplicate ads was removed resulting in 43 ad instances. Analysis is based on text only. The AntConc tool is used for initial analysis to determine high level patterns.

3.2.2 Results and Discussion

Table 1: Most frequently used words



Rank	Freq	Word
1	90	we
2	87	the
3	87	to
4	78	a
5	78	beleave
6	78	eu
7	77	our
8	67	and
9	59	future
10	50	in

Table 1 shows the ten most frequently used words, ranked from 1 to 10. Grammatically, the majority of these words are function words, such as pronouns (we, our), articles (the, a) or prepositions (in), words typically discarded from analysis, being considered stop words that carry out little or no semantic value. However, we believe that they are essential in understanding the discourse of computational propaganda. For instance, there are 90 instances of the personal pronoun 'we' which may be explained by the fact that 'we' aligns the group ideology with the target's identity, adding to the sense of familiarity between the sender and the recipient. In addition, when one considers the context in which 'we' occurs (Table 2), it can be observed that the concordance of 'we' covers both positive and negative connotations. When 'EU' is presented prior to 'we' there is observable negative semantic prosody (Louw, 1993), resulting in an ideological polarisation effect with the ingroup 'Britain' which is positively represented, representations that can be further explained by van Dijk's (2016) 'idealogical square' which represents the 'self' as positive and the 'other' as negative. The polarisation of the contextual information associated with the pronoun 'we' is consistent throughout the dataset as evidenced by the fact that 'we' negatively relates to emotional terms such as 'powerless' and 'out of control' when co-occurs in the same context as 'EU', whereas 'we' positively relates to positive actions such as 'we can' when used to refer to the to the ingroup 'Britain' since they reflect empowerment. The prevalence and emphasis of the positive self-description 'we' exacerbates the division between the in-group British, and the negative other-description referring to the EU. Similarly, frequent use of the pronouns 'ours', 'they' and 'theirs' are synonymous with political groups and can achieve a polarising effect for example "Are *they* focused on *their* priorities or *ours".* Repeated instances of the activities that 'we' have to do, is consistent with the identification and alignment of ideological groups. Further the objective ideological 'norms and values' are expressed explicitly and implicitly with verbs such as *can* have, *can take back*.

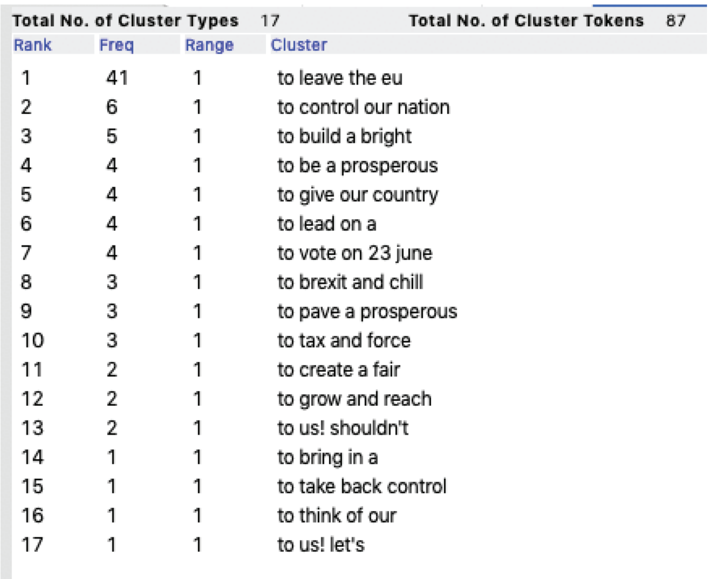
[1] "Impressions is a common metric used by the online marketing industry. Impressions measure how often your ads were on screen for your target audience." (Facebook, 2019)

Table 2: Negative and Positive use of 'we'

Negative – 'we'		Positive – 'we'	
Inside the EU,	we are powerless	So	we can bring
Under EU laws	we are unable	Under our control	we can bring
Under EU regulations	we are unable	On 23 June so	we can chart
Shouldn't	we be in control	On 23 June so	we can have
Isn't it time	we became an independent	Today, 23 June, so	we can have
It is time	we break free	On 23 June so	we can make
		Today, 23 June, so	we can make
		23 June so that	we can take back
		23 June so that	we can take back
		Isn't it time	we chart our own
		Opportunities	we could have outside
		Is so important.	We have an opportunity
		The opportunities	we know will make
		Why do	we let them do
		It is time	we take a stand
		It is time	we unite to give
		Of the EU	we will become

The definite article 'the' appears 87 times, and a collocational analysis links the use of 'the' to 'eu' on 62 occasions. This further enhances the polarisation effect, indicating that 'the' is used to underline the definite and real threat of an entity that opposes the referred to by the pronoun 'we'. Similar to the word 'the', there are 87 instances of the word 'to'. When clustered with three subsequent words, 'to' seems to almost exclusively serve as verb particle indicating that the target audience should take action '*to* leave the eu' or '*to* build a bright future'. By analysing these clusters further, we can also determine a link between the eu, British nationals, and an action to 'leave the eu' to assume the underlying message. Observe the use of positive self-descriptions such as 'our nation', and prepositions '*to* build', '*to* be a prosperous', which align the audience with the sender.

Table 3: Clusters of 'to'

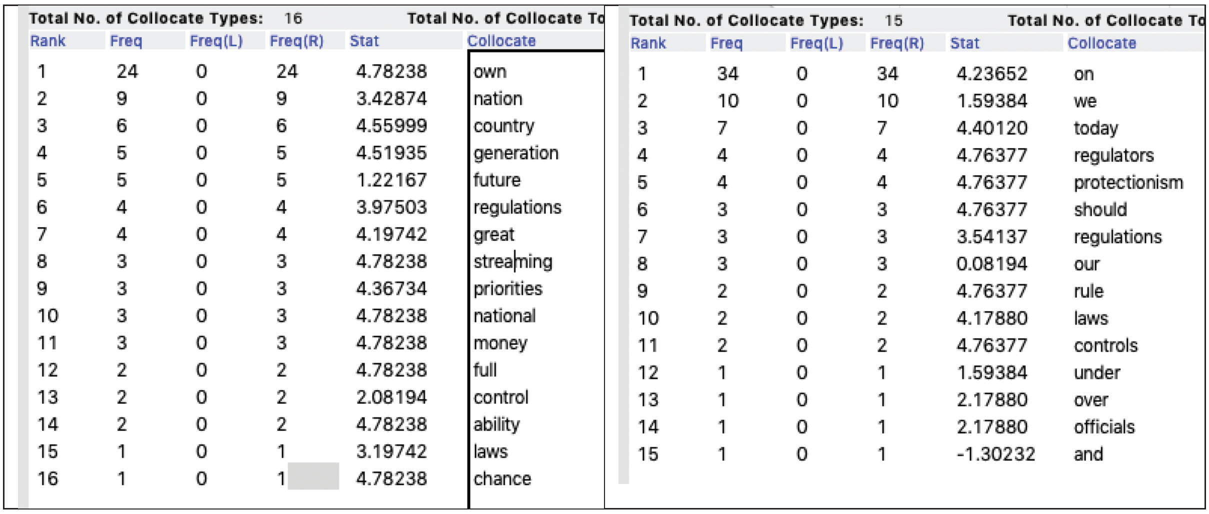


Total No. of Cluster Types 17 Total No. of Cluster Tokens 87

Rank	Freq	Range	Cluster
1	41	1	to leave the eu
2	6	1	to control our nation
3	5	1	to build a bright
4	4	1	to be a prosperous
5	4	1	to give our country
6	4	1	to lead on a
7	4	1	to vote on 23 june
8	3	1	to brexit and chill
9	3	1	to pave a prosperous
10	3	1	to tax and force
11	2	1	to create a fair
12	2	1	to grow and reach
13	2	1	to us! shouldn't
14	1	1	to bring in a
15	1	1	to take back control
16	1	1	to think of our
17	1	1	to us! let's

There are 77 instances of the pronoun 'our' and 78 instances of the noun 'eu'. By assessing the collocates of these two words (Table 4) provides an insight into the tone of the text body. 'Our' relates to the Britain ingroup 'nation', 'country', 'national' with positive connotations such as 'great' and 'control', whereas 'EU' relates to the EU outgroup with negative connotations such as 'regulators' and 'protectionism'.

Table 4: Collocates of ‘our and ‘eu’



Total No. of Collocate Types: 16 Total No. of Collocate To

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	24	0	24	4.78238	own
2	9	0	9	3.42874	nation
3	6	0	6	4.55999	country
4	5	0	5	4.51935	generation
5	5	0	5	1.22167	future
6	4	0	4	3.97503	regulations
7	4	0	4	4.19742	great
8	3	0	3	4.78238	streaming
9	3	0	3	4.36734	priorities
10	3	0	3	4.78238	national
11	3	0	3	4.78238	money
12	2	0	2	4.78238	full
13	2	0	2	2.08194	control
14	2	0	2	4.78238	ability
15	1	0	1	3.19742	laws
16	1	0	1	4.78238	chance

Total No. of Collocate Types: 15 Total No. of Collocate To

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	34	0	34	4.23652	on
2	10	0	10	1.59384	we
3	7	0	7	4.40120	today
4	4	0	4	4.76377	regulators
5	4	0	4	4.76377	protectionism
6	3	0	3	4.76377	should
7	3	0	3	3.54137	regulations
8	3	0	3	0.08194	our
9	2	0	2	4.76377	rule
10	2	0	2	4.17880	laws
11	2	0	2	4.76377	controls
12	1	0	1	1.59384	under
13	1	0	1	2.17880	over
14	1	0	1	2.17880	officials
15	1	0	1	-1.30232	and

There is a significant use of imperatives which are frequently applied in advertising to essentially tell someone to do something by making commands or recommendations for example ‘Let’s stop’, Let’s spend’ and ‘Let’s vote’ with the latter represented in 43 instances. Presupposition (or implied) meaning is present by ‘Let’s stop EU regulators from controlling…” whereby a reader infers that EU regulators already ‘are’ controlling. Similarly the definite ‘the’ is used in the context ‘the freedom to be’ and ‘chase the opportunities’ presuppose that ‘a’ freedom has been taken and ‘an’ opportunity has been taken. Advertising techniques such as slogans and taglines are present with 78 instances of ‘BeLeave’. Similarly rhetorical questions introduce familiarity by engaging the audience to participate in conversation. For example: ‘Are they focused on their priorities or ours?’, ‘Did you know that the EU controls over 60% of our regulations? Why do we let them do this?’, ‘Shouldn't we be in control of our future?’, ‘Isn't it time we chart our own destiny and chase the opportunities we know will make Britain successful?’, ‘ Let's vote to leave the EU on 23 June so we can have a clear and prosperous future?’

Another discourse factor that impacts the effectiveness of computational propaganda in the present dataset is modality which is represented by 44 concordances relating to ‘can’. ‘Can’ reflects the audiences ability to effect change . Modal claims combined with verbs invoke a sense of possibility for a positive outcome: ‘Can bring in talent’, ‘Can chart our own destiny’, ‘Can have a clear and prosperous future’, ‘Can make our own decisions’ and ‘Can take back control’. Predicational strategies (Reisigil and Wodak, 2001) describe the linguistic assignment of qualities to induce judgement by the audience as demonstrated by the use of the word ‘unelected’: ‘unelected EU officials’, ‘unelected foreign officials’, ‘unelected officials that have no idea’.

‘The Choice’ presented to the audience is synonymous with pro-war propaganda and is designed to end debate on the topic (Richardson, 2007) whereby the audience perceives only two options. In the context of this study the audience perceives that a) a ‘remain’ vote equates to ‘doing nothing’ and ‘doing nothing’ equates to ‘losing control / jobs’, or b) a ‘leave’ vote equates to ‘doing something’ and ‘doing something’ equates to a ‘brighter future’ with ‘prosperity’.

From the initial analysis we have determined that the principles surrounding the illusory truth effect are relevant to computational propaganda insofar as how the message is framed, interpreted and repeated may provide insight to the cybersecurity community. Consequently we can conduct a brief qualitative analysis to extrapolate what audience may be influenced by this message and why. This brief secondary study is applied to give context to the overall nature of the study.

3.3 Qualitative Data Collection and Analysis

In addition to the conclusions drawn , an individual ad from BeLeave’s Facebook page is briefly assessed using elements of van Dijk’s discourse-cognition-society triangle to link these complex theories. This assessment is not exhaustive as a detailed analysis is outside the scope of this paper, moreover it is applied to add context to the initial study. van Dijk asserts that although critical discourse studies surround society and discourse, *“a sociocognitive approach claims that such relations are cognitively mediated. Discourse structures and social* *structures are of a different nature, and can only be related through the mental representations of language users as individuals and social members"* (van Dijk, 2016).

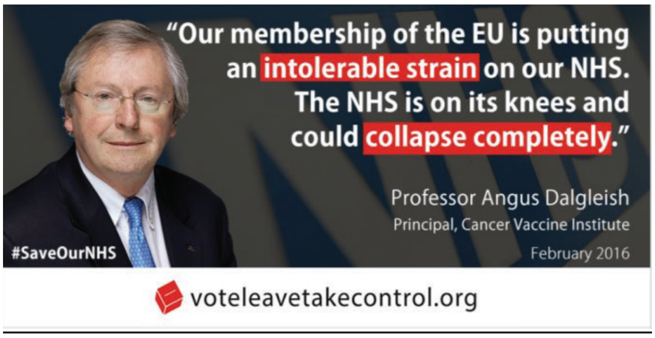

"Our membership of the EU is putting an intolerable strain on our NHS. The NHS is on its knees and could collapse completely."
Professor Angus Dalgleish
Principal, Cancer Vaccine Institute
February 2016
#SaveOurNHS
voteleavetakecontrol.org

Figure 1: BeLeave Facebook ad (BeLeave, 2019)

Firstly we consider a discursive and semiotic analysis to study the implied and inferred meaning of the ad in figure 1. It is assumed that there is a shared sociocultural awareness of the topic which is facilitated by online advertising. Numerous cognitive structures are required to interpret the message. By using the semiotic hashtag #SaveOurNHS, the reader is commanded to commit to the concept supported by this campaign. The possessive pronoun 'our' contextually refers to Britain with the presupposition being that the EU is responsible for the problems within the NHS. Social identity of an authoritative figure is semiotically denoted by the title 'Professor' and represented by formal appearance suggests authority and leads to greater likelihood that a reader will submit to the theory presented by a knowledgeable authoritative figure. People may not be familiar with the intricacies of the argument but assume an authoritative figure can be trusted. Observe also the use of the hashtag also supports analysis and measurement of sentiment surrounding the topic for success or failure of the approach. The inferred message is that the health of the British public is in imminent danger. There is a sense of urgency implied by the use of red highlighting 'intolerable strain' and 'collapse completely' against the backdrop of the NHS logo. The urgency implied prompts the reader to an emotional, urgent response. Essentially EU equates to 'intolerable' and 'collapse' of a fundamentally British institution resulting in a polarising effect between Britain and the EU.

In consideration of the cognitive structures we observe that sociocultural knowledge is assumed surrounding the struggling health service that people depend on and are likely familiar with either directly or indirectly which gives a personal context for doubt in the EU, that may result in an emotional response. There are implications for online focus as it facilitates targeted areas and demographics who are more readily impacted by these events to increase the polarising effect in the context of the ideological square. There is a suggestion that the readers way of life is at risk by presupposing that the EU has taken control implying that control has already being lost, which places the reader on the defensive. BeLeave, the advertiser, is communicating interactively with 'you' the British public.

4. Inferences Drawn

This foundational study supports the theory that adopting a sociocognitive perspective to discourse can contribute to the 'intelligence' derived from computational propaganda strategies. Through content analysis, cybersecurity professionals may extrapolate who is being targeted with a particular message allowing them to extrapolate which threat actors have the motivation, opportunity and capability to conduct such campaigns. The inferences drawn are based on the principle that content analysis is in place within an organisation, and there is an established baseline. Traffic patterns can then be analysed using the quantitative and qualitative methods discussed to identify who is being targeted with what message, followed by analysis for more detailed intelligence. When collocational profiles become synonymous with social issues there can be a psychological association for individuals as determined by Richardson's 'choice'.

A target audience is selected based on sociocultural knowledge that is available from social analytics. In consideration of recognised psychological assumptions, a simple easily recognisable message is constructed and repetitively presented and delivered under different guises until target takes a perspective and aligns with one view or another. The cycle is then repeated to reinforce the perspective in the targets view until target adopts perspective as their own opinion. Analytics furnish an attacker with feedback on the success or failure of a campaign allowing the approach to be refined.

Table 5: Adapted Context Model

Context parameters	Discourse Structures	Intelligence
Spatiotemporal dimension of online communication	Date: Leadup to Brexit referendum. Location: UK. Present tense as of February 2016	Opportunity: Event and duration of campaign
Advertiser: British citizen in the role of concerned citizen or campaigner	Ads: deictic expression: 'our'	Capability: Extent, expense and sophistication of campaign, and ancillary campaigns.
Engaged in asserting perceptions of societal injustices in the form of advertising which is an element of social media targeted advertising	Opinion expressed: health of the British public is at risk 'intolerable strain', 'on its knees', collapse completely'.	Motivation: To achieve polarizing effect between the EU and British electorate in support of votes. Who gains if objective is achieved?
Objective is to influence opinion of undecided voter	'voteleavetakecontrol.org'	Target: undecided groups. Induce heightened sense of uncertainty. Emotional response

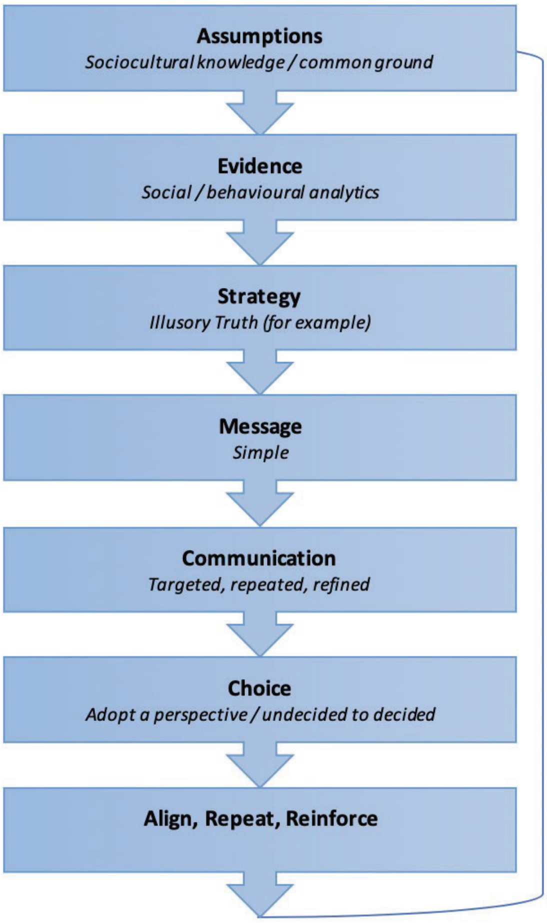

Assumptions
Sociocultural knowledge / common ground
Evidence
Social / behavioural analytics
Strategy
Illusory Truth (for example)
Message
Simple
Communication
Targeted, repeated, refined
Choice
Adopt a perspective / undecided to decided
Align, Repeat, Reinforce

Figure 2: Attack Lifecycle

Results align with principles of illusory truth effect and support the value of computational propaganda to an attacker. Social media supports the use of personalised targeting resulting in the ability to communicate a medium to an audience susceptible to the view. When messages become familiar to an audience they are more readily willing to accept the statements. How the message is interpreted is based on many factors as evidenced by van Dijk's sociocognitive approach. For example van Dijk's ideational square determines the ingroup and outgroup, Richardson's 'choice' solidifies a view and social media platform providers provide the means and basis to analyse response for improved results. To conclude, cyberpsychology is a key requirement within the cybersecurity community to respond to non-traditional cyber threats.

5. Criticisms of Approach

There are many criticisms of critical discourse studies that pertain to the broad and imprecise nature of the approach which is in contrast to the analytical approach generally adopted to technical threats, however as identified in this study, the prevalence of analytical tools for online communications supports the use of sociocognitive discourse studies to identify at a high level that a cyberpscychological attack may be present or emerging. Considering threats of this nature, allows the cybersecurity community to prepare and build appropriate responses for future threats.

6. Future

The objective of this study was to highlight why psychological attack methods are critical to the cybersecurity community. Future studies may broaden the psychological theories and analytical methods employed and overlay the results with other technical methods for greater results.

7. Conclusion

It is readily accepted that people are susceptible to influence campaigns of all descriptions however it is assumed that we are protected by educational strategies, or cognitive defences. This paper considered why psychological strategies when applied effectively are powerful when combined with the global nature of social media. On the premise that computational propaganda and cyberpsychological threats are fundamentally transdisciplinary, we considered how cyberpsychological attacks can be categorised for analytical research by proposing an attack lifecycle to be expanded on in future studies. We conclude that analytical tools available by social media providers ensure that targeted campaigns can be refined for greater effect. Impression rates, combined with content and sentiment analysis determine, and compound the effect of cyberpsychological threats. We determined that the construct language and delivery of digital communications can contribute to cyberpsychological threats when communicated and delivered to an audience susceptible to a view. To conclude, the symbiotic relationship between technology and people demands symbiotic consideration from the cybersecurity community to defend against emerging threats.

References

Arkes, H.R., Hackett, C. and Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of behavioural decision making.* 2. 81-94.

Bacon, F.T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of experimental psychology. 5(3). 241-252.*

Barton, H. (2016). *Persuasion and compliance in cyberspace*. In Connolly, I. et al (eds) An introduction to cyberpsychology. Oxen/New York: Routledge

BeLeave (2019) 20 May. BeLeave. Available at: https://www.facebook.com/voteleave/posts/each-week-we-send-350-million-to-the-eu-enough-to-build-a-new-fully-staffed-nhs-/552289318281330/ [Accessed 20 May 2019]

Bernays, E. (1928). *Propaganda.* New York. H. Liveright.

Bradshaw, S. and Howard (2017) 'Troops, trolls and troublemakers: A global inventory of organized social media manipulation'. The Computational Propaganda Project. University of Oxford.

Creswell, J.W. (2015). *A concise introduction to mixed methods research*. US. Sage.

Danesi, M. (2015) *Advertising discourse.* In Tracy, K., Ilie, C. and Sandel, T. (eds) The international encyclopedia of language and social interaction. John Wiley & Sons.

De keersmacker, J., Dunning, D.A., Pennycook, G., Rand, D.G., Sanchez, C., and Roets, A. (2019). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style [online]. Available at: https://psyarxiv.com/n7bze/ [Accessed 23 July 2019]

Facebook. (2019). Impressions. Available at: https://www.facebook.com/business/help/675615482516035 [Accessed 8 October 2019]

Gladwell, M. (2006) *The tipping point: how little things can make a big difference*. Available at: http://www.innovationlabs.com/tipping_point.pdf (Accessed 1 May 2018).

Hasher, L. and Goldstein, D. (1977). Frequency and the conference of referential validity. *Journal of verbal learning.* 16. 107-112.

House of Commons (2019) Fake news evidence. Available at: http://www.parliament.uk/documents/commons-committees/culture-media-and-sport/Fake_news_evidence/Letter-from-Rebecca-Stimson-Facebook-to-Chair-re-question-29-19-July-2018.pdf [Accessed: 10 May 2019]

Louw, Bill (1993) *Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies*. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) Text and Technology: In Honour of John Sinclair. Philadelphia/Amsterdam: John Benjamins

Nimmo, B. (2019). *Measuring Traffic Manipulation on Twitter.* Working Paper 2019.1. Oxford: Project on Computational Propaganda. Available at: https://comprop.oii.ox.ac.uk/research/working-papers/twitter-traffic-manipulation/

Ozubko, J.D. (2010). Remembering makes evidence compelling: Retrieval from memory can give rise to the illusion of truth. *Journal of experimental psychology: Learning, memory, and cognition.* 1. 1-7

Reisigl, M. and Wodak, R. (2001) *Discourse and Discrimination. Rhetorics of racism and antisemitism*. London: Routledge

Richardson, J. (2007) *Analysing Newspapers: An approach from critical discourse* analysis. Hampshire / New York: Palgrave

van Dijk, T. (2013) *CDA is not a method of critical discourse analysis*. In: EDISO Debate – Asociacion de Estudios Sobre Discurso y Sociedad. www.edisoportal.org/debate/115-cda-not-method-critical-discourse-analysis, accessed [12 July 2019]

van Dijk, T. (2016*) Critical discourse studies: A sociocognitive approach*. In Wodak, R. and Meyer, M. Methods of Critical Discourse Studies, 3rd Edition London: Sage

Wodak, R. (2012) *Editor's Introduction: Critical discourse analysis – challenges and perspectives*. In: R. Wodak (ed.) Critical Discourse Analysis. London: Sage.

Wodak, R. and Meyer, M. (2016) *Methods of Critical Discourse Studies*, 3rd. Edition. London: Sage

Woolley, S. C. and Howard, P.N. (2016). Automation, algorithms and politics: Political communication, computational propaganda, and autonomous agents. *International Journal of Communication.* 10(0) 9.

Comparing Personality Traits Between Penetration Tester, Information Security, and its Professionals from two Cohorts

Olav Opedal
Capella University, Minneapolis, Minnesota, USA
olav@opedalconsulting.com
DOI: 10.34190/EWS.20.081

Researchers from several disciplines have studied hacking, the effects of hacking on individuals and society, and how organizations use penetration testers to try to mitigate hacking. Penetration testers use the same techniques as criminal hackers to obtain access to computer systems. The difference is that organizations employ penetration testers to test the security of their computer systems and find any holes before criminal hackers can gain access. The focus of this study was to identify the Big Six traits associated with penetration testers. There was a gap in the literature about the personality traits of those who are employed to hack. There was also a gap in the literature on the personality traits for the Big Six (Sibley et al., 2011) traits for millennials and Generation Xers, and three groups of computer professionals: IT professionals, information security professionals, and professional penetration testers.

1. Review of the Literature

Researchers agree about distinguishable traits that are stable over time and cultures (Ashton, Lee, & De Vries, 2014; Costa & McCrae, 2008; Fleeson & Jayawickreme, 2015; McCrae, & Costa, 1997). There are differing opinions about how many traits there are. The first broad consensus about traits emerged with the development of the Big Five traits in the 90s, including openness, conscientiousness, extroversion, agreeableness, and neuroticism (Ashton, Lee, & De Vries, 2014; Costa & McCrae, 2008; Fleeson & Jayawickreme, 2015; McCrae, & Costa, 1997). Starting with the HEXACO model developed by Lee and Ashton (2004), a sixth personality trait emerged as a better answer for the dark triad traits of narcissism, Machiavellianism, and psychopathy. The HEXACO model was integrated with the Big Five by Sibley et al. (2011), which has since been tested and evaluated on large representative population samples by Milojev, Osborne, Greaves, Barlow and Sibley (2013) and Sibley and Pirie (2013). The model explains traits cross-culturally, and they are stable over time. The sixth trait is a continuum on the marker of honesty/humility (Sibley et al., 2011). The theme that emerged from information security research is that it has become multidisciplinary (Weems, Ahmed, Golden, Russell, & Neill, 2018). Some of the cyber-crime research from the field of computer science dealt only with algorithmic approaches. No studies were found using the extended Big Six model for measuring employment performance and personality traits among IT employees. The review led to the conclusion that the personality traits and agentic engagement of the three groups of professionals were unknown (Ledingham & Mills, 2015; Macdonald et al., 2008; Marcum & Higgins, 2014; Whalen & Gates, 2007).

2. Research Questions

Research question 1. Are there statistically significant differences in the combined linear variate scores for extroversion, agreeableness, conscientiousness, neuroticism, openness, and honesty/humility, between millennial or Generation X individuals who either self-identified as penetration testers, information security professionals, or IT professionals?

Research question 2. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) for scores on extroversion?

Research question 3. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) for scores on agreeableness?

Research question 4. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) scores on conscientiousness?

Research question 5. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) scores on neuroticism?
Research question 6. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) scores on openness?

Research question 7. Is there a significant main effect of self-identified professional group membership as a penetration tester, information security professional, or IT professional, and is there a significant main effect by age group (millennials versus Generation X) scores on honesty/humility?

3. Research Design

The study used a quantitative, nonexperimental research design with an online survey for data collection. The survey consisted of items from the Mini-IPIP6 and a short demographic survey collecting the age of the respondents. The research questions were answered using a two-way factorial MANOVA, and a post hoc using Sidak. The post hoc was conducted in SPSS with the Sidak correction. The Sidak was chosen because the method adjusts *p* values while controlling the family wise error rate (Valdes, Rhees, & Erlich, 2004). Using the Sidak correction addressed the problem of multiple comparisons (Valdes et al., 2004).

To ensure balanced data, post stratification random sampling was used (Levy & Lemeshow, 2013). The two-way MANOVA analysis investigated a binary factor, cohort, a categorial factor consisting of three levels, and six dependent variables. The binary factor was cohort membership in either the millennial group, or the Generation X group. The categorial factors were three types of computer professionals: IT professionals, information security professionals, and penetration testers. The six dependent variables were the mean scores from the Mini-IPIP6 instrument for extroversion, agreeableness, conscientiousness, neuroticism, openness, and honesty/humility.

4. Description of the Sample

Participants were recruited from online community platforms including LinkedIn, Reddit, and Amazon MTurk Internet sites. A total of 1099 people responded to the invitation to participate. Seven hundred and thirty-three respondents were included for the initial analysis. Table 1 shows the distribution of the respondents by profession and cohort.

Table 1: Distribution of Respondents by Profession and Cohort

Cohort	InfoSec	IT professional	Pen-tester	Total
Millennial	73	285	75	433
Generation X	57	171	72	300
Total	130	456	147	733

The dataset was unbalanced with more IT professionals responding compared to the other professional groups, and more millennials responded than Generation Xers. The imbalance was corrected by a post collection random stratification process (Levy & Lemeshow, 2013). Of the 1099 respondents, 733 completed the full survey. These 733 were then resampled, resulting in 427 responses for the final analysis.

5. Data Manipulation and Analysis Process

Statistical software, SPSS, was used to conduct a two-way MANOVA with a post hoc Sidak. Using Sidak correction addressed the problem of multiple comparisons (Valdes et al., 2004). The categorical independent variables were profession and cohort, and the dependent variables were openness, conscientiousness, extroversion, agreeableness, neuroticism, and honesty/humility. There were two independent variables, one dichotomous variable, and one categorical variable with three levels. The dichotomous variable cohorts were *millennials* or *Generation X*. The independent categorical variable *professional group membership* consisted of three levels: IT professionals, information security professionals, and penetration testers. There were 26 questions in the instrument, which resulted in six dependent variables and two independent variables. Two of the variables contained cohort and professional affiliation data, and six dependent variables measured personality traits based on the Big Six. The six dependent variables were the mean scores from the four questions for each of the six traits from the Mini-IPIP6 consisting of the means from each of the four questions for each of the six dependent variables.

6. Reverse-Keying

Seven hundred and thirty-three valid responses were analyzed for reliability and validity. The next step in the data transformation process was to reverse key the responses required for rekeying. Each personality trait dimension had two negatively keyed and two positively keyed entries in the Mini-IPIP6, and it was unlikely that responses with high or low scores were caused by low variances in answers or random answers. The responses with congruent high or low scores after reverse keying the negative answers were more likely to represent the population. The remaining responses were, however, kept as they were probable answers because each dimension had both negative keyed and positive keyed entries. The results of high or low scores were unlikely to be due to low variance in answers or random answers and more likely to represent the population. The means for each cohort showed differences between professions and cohorts, which indicated differences between cohorts on the honesty/humility dimension.

The MANOVA required that specific assumptions were met, which was achieved by using post-stratification (Levy & Lemeshow, 2013). The initial data collection resulted in large differences in group sizes, causing very uneven cell sizes; therefore, the study used a post stratification sample for the two groups that caused most of the imbalance (Levy & Lemeshow, 2013). Bartlett's test of homogeneity of variances of the rebalanced data was not significant for any of the dependent variables. Both the measures of kurtosis and skew were between -1 and 1, which indicated normal distribution. A two-way multiple analyses of variance was conducted with the dependent variables (a) mean extroversion scores, (b) mean agreeableness scores, (c) mean conscientiousness scores, (d) mean neuroticism scores, (e) mean openness scores, and (f) mean honesty/humility scores, and the independent variables: profession, and cohort. The analyses evaluated whether personality scores for the Big Six personality traits differed based on cohort membership (i.e., millennials, Generation X), and professional affiliation (i.e., IT professional, information security professional, penetration tester).

The MANOVA identified statistically significant differences between the type of computer professions and the dependent variables, $p = 0.003$, and between cohorts, $p < .0005$, without any significant interaction between the independent variables. The tests of between-subjects differences identified statistically significant differences in mean extroversion scores between types of computer professionals ($p = .035$), neuroticism ($p = .029$), openness ($p < .0005$), and honesty/humility ($p = .009$).

The tests of between-subjects differences identified statistically significant differences between cohorts in extroversion ($p < .0005$), agreeableness ($p = .022$), conscientiousness ($p = .002$), and neuroticism ($p < .0005$). Therefore, the null hypotheses were rejected for (a) extroversion, (b) agreeableness, (c) conscientiousness, and (d) neuroticism.

The post hoc analysis used the Sidak method for multiple comparisons, which identified statistically significant differences in extroversion between information security professionals and penetration testers ($p = .022$). Information security professionals are more extroverted compared to penetration testers. There were significant differences in neuroticism between information security professionals and IT professionals ($p = .01$). Information security professionals mean scores were higher than IT professionals. There were significant differences in openness between information security professionals and penetration testers ($p = .002$). Penetration testers scored highest in openness, followed by information security professionals, and IT professionals scored lowest. Penetration testers scored higher in openness compared to IT professionals ($p = .001$). Information security professionals scored higher than IT professionals ($p = .002$). There was no statistically significant difference between information security professionals and penetration testers in openness. There were statistically significant differences in means between information security professionals and IT professionals in honesty/humility ($p = .028$), with IT professionals scoring lower than information security professionals. There was no statistically significant difference between information security professionals and penetration testers. Penetration testers also scored higher in honesty/humility compared to IT professionals, $p = .024$ (see Table 2).

Table 2: Sidak Post Hoc Test of Mean Mini-IPIP6 Scores

Dependent Variable	(I) comp_prof	(J) comp_prof	Mean Difference (I-J)	Std. Error	Sig.
Extroversion	info_sec_pro	it_pro	.2244	.11787	.163
		pen	.3187	.11843	.022*
	it_pro	info_sec_pro	-.2244	.11787	.163
		pen	.0944	.11416	.793
	pen	info_sec_pro	-.3187	.11843	.022*
		it_pro	-.0944	.11416	.793
Agreeableness	info_sec_pro	it_pro	.1958	.11753	.263
		pen	.0891	.11809	.834
	it_pro	info_sec_pro	-.1958	.11753	.263
		pen	-.1067	.11383	.724
	pen	info_sec_pro	-.0891	.11809	.834
		it_pro	.1067	.11383	.724
Conscientiousness	info_sec_pro	it_pro	.1167	.09303	.508
		pen	.0446	.09347	.951
	it_pro	info_sec_pro	-.1167	.09303	.508
		pen	-.0721	.09010	.809
	pen	info_sec_pro	-.0446	.09347	.951
		it_pro	.0721	.09010	.809
Neuroticism	info_sec_pro	it_pro	.3319	.11259	.010*
		pen	.1538	.11313	.438
	it_pro	info_sec_pro	-.3319	.11259	.010*
		pen	-.1781	.10905	.279
	pen	info_sec_pro	-.1538	.11313	.438
		it_pro	.1781	.10905	.279
Openness	info_sec_pro	it_pro	.5006*	.14657	.002**
		pen	-.0084	.14726	1.000
	it_pro	info_sec_pro	-.5006*	.14657	.002**
		pen	-.5090*	.14195	.001**
	pen	info_sec_pro	.0084	.14726	1.000
		it_pro	.5090*	.14195	.001**
Honesty/humility	info_sec_pro	it_pro	.3950	.15137	.028*
		pen	.0058	.15208	1.000
		info_sec_pro	-.3950	.15137	.028*
	it_pro	pen	-.3892	.14660	.024*
		info_sec_pro	-.0058	.15208	1.000
	pen	it_pro	.3892	.14660	.024*

The analyses were conducted using two-way MANOVA with a post hoc using the Sidak procedure on post stratified sampling. The two-way MANOVA was statistically significant for the independent variables, type of computer professional (*p* = .003) and cohort (*p* < .0005), and the six dependent variables (a) extroversion, (b) agreeableness, (c), conscientiousness, (d) neuroticism, (e) openness, and (f) honesty/humility.

The between-subjects test identified statistically significant differences between types of computer professionals in mean extroversion scores (*p* = .035), mean neuroticism scores (*p* = .029), mean openness scores (*p* < .0005), and mean honesty/humility scores (*p* = .009). Differences in mean scores in agreeableness and conscientiousness were not statistically significant. The between-subjects test identified statistically significant differences between cohorts in extroversion (*p* < .0005), agreeableness (*p* = .022), conscientiousness (*p* = .002), and neuroticism (*p* < .0005). There were no statistically significant differences in openness and honesty/humility between cohorts.

The post hoc tests identified that information security professionals are more extroverted when compared to penetration testers (*p* = .022). Information security professionals mean scores in neuroticism were higher than

IT professionals (*p* = .01). Penetration testers scored higher in openness compared to IT professionals (*p* = .001). Information security professionals scored higher than IT professionals on the openness trait (*p* = .002). There was no statistically significant difference between information security professionals and penetration testers in openness. There were statistically significant differences in means between information security professionals and IT professionals in honesty/humility (*p* = .028) with IT professionals scoring lower than information security professionals. There was not a statistically significant difference between information security professionals and penetration testers. Penetration testers also scored higher in honesty/humility compared to IT professionals (*p* = .024).

7. Summary of the Results

The study investigated if there were personality trait differences between professional affiliation for IT professionals, information security professionals, and penetration testers. The study also evaluated differences in personality traits for millennials and Generation Xers from the professional clusters. The study measured extroversion, agreeableness, conscientiousness, neuroticism, openness, and honesty/humility using the Mini-IPIP6 instrument.

8. Discussion of the Results

The statistically significant differences between the type of computer professional in (a) extroversion, (b) neuroticism, (c) openness, and (d) honesty/humility and differences in cohort (millennials versus Generation X) in (e) extroversion, (f) agreeableness, (g) conscientiousness, and (h) neuroticism are important because they raise questions. The first question was whether there was a statistically significant difference due to maturation. A second question was whether there was a statistically significant difference in personality traits between millennial and Generation X computer professionals. The trait theorists Costa and McCrae (2008) posited that individuals have stable patterns of thoughts and feelings. If the differences in cohort stay stable over time, it would be an indication that those from the Generation X cohort who chose to become computer professionals differ in trait composition from the millennials who have chosen to become computer professionals. Parks-Leduc et al. (2015) noted that there is an interaction between values and cognitive traits, and it is possible for the cognitive traits to be influenced by values in adolescence and early adulthood. Traits can be influenced by the environment (Costa & McCrae, 2008). It is, therefore, also possible that the two cohorts had different experiences and values during the formative years; however, this does not explain the differences for neuroticism. The neuroticism trait does not have a corresponding value (Parks-Leduc et al., 2015). Prentice et al. (2019) noted that personality traits should be viewed as an individual's propensities to enact according to the trait to meet their goals. The cohort differences would either indicate that there are differences in goals between cohorts that remain as each cohort moves through time, or it means that goals change over time, which is what this study has uncovered. This study did not attempt to find the answer to questions about causality differences. Instead, the goal was to determine if there were differences between the two cohorts of computer professionals. There was, however, evidence from the literature and in the data here that shed light on the question of whether maturity could be the cause of differences in personality traits. Personality traits tend to be stable, but there is maturation over the lifespan (Damian, Spengler, Sutu, & Roberts, 2018). A definite answer about the causality of differences in cohort scores on personality traits is elusive; only longitudinal monozygotic twin research studies can fully answer the question of causality. A longitudinal study could possibly answer the question of causality by exploring the interactions between values, traits, environment, and maturation. Personality trait theory does predict maturation over the lifespan, and this study seems to confirm that the cohort difference is due to maturation.

The results identified that both penetration testers and information security professionals are humbler and more honest compared to IT professionals. The study also showed both penetration testers and information security professionals are more open to new ideas when compared to the IT professionals. Furthermore, information security professionals were more anxious compared to IT professionals. Finally, penetration testers were more introverted compared to information security professionals. The lack of a difference in agreeableness between the information security professionals and IT professionals was unexpected. So were the high scores in openness for information security professionals compared to IT professionals. Whalen and Gates (2007) found that information security professionals scored high in agreeableness and low in openness, which was different from the current study. Whalen and Gates (2007) compared information security professionals' personality traits with the general population. This study compared computer professional groups, so it is not a one-to-one comparison. Unfortunately, Whalen and Gates (2007) did not report mean scores for the traits, just how many individuals scored high, medium, or low on each trait and associated facets, which made any further comparison between the results of their study and this study impossible. Their study used the FFM, this study used the six-factor model of personality. Both studies used IPIP derived instruments. The lack of differences in agreeableness in this study may be due to participant selection. It may also be that those who become computer professionals are more agreeable compared to the general population. It may have been that computer professionals score higher in the agreeable trait compared to the general population, rather than just information security professionals. Another contrast was that Whalen and Gates (2007) solicited participants from a single information security conference and based their study on responses from 39 respondents. This study evaluated the responses from 130 information security professionals, 150 IT professionals, and 147 penetration testers collected data from multiple social platforms. Fleeson and Jayawickreme (2015) noted that personality traits describe identities with propensities for behaviors. In this study of differences between the computer professionals, there was no support for Whalen and Gates' (2007) finding that information security professionals are likely to be dutiful plodders. The chief difference in the instruments between Whalen and Gates' (2007) study and this one was the additional measurement to the IPIP scales of the honesty/humility construct that originated from the honesty/humility, emotionality, extroversion, agreeableness, conscientiousness, openness, HEXACO, constructs (Ashton, Lee, & De Vries, 2014). Menard et al. (2017) noted that the information security profession adopted the use of PMT that uses messages with fear appeal rather than motivational language. The choice of PMT over SDT could support the findings from Whalen and Gates (2007). Empirical research from Li et al. (2019). found that employees became more competent in managing cyber-security risks after they had been made aware of information security policies and procedures. This finding was supported by Torten et al. (2018) who found that awareness contributed to 61.2% of security behavior. Martens et al. (2019) supported that view and recommended to increase security awareness as the key countermeasure against cyber-crime. If information security professionals were open to new ideas, it could be argued that the choice of PMT is supporting evidence for low openness scores. However, the fact that there is an active field of study within information security that evaluated the effectiveness of PMT, SDT, and awareness, is an argument for higher scores in openness within information security professionals as they continue to question the status quo and current ideas in vogue. Bergmann et al. (2018) recommended further studies into online behaviors. Since traits are linked to behavior, this study supported their recommendation by elucidating the traits associated with computer professionals. The higher mean scores by penetration testers on honesty/humility compared to the other groups' test scores seems to support the notion that there are differences between those with lack of self-control who commit criminal hacking and white hat hackers (Marcum & Higgins, 2014).

The cohort differences in extroversion, agreeableness, conscientiousness, and neuroticism between the cohorts might be a possible tool for organizational change. Merhi and Ahluwalia (2019) reported that social normative and moral norms were effective to decrease employee resistance to protective cyber-security measures. It may be possible that Generation X cohort members might influence millennials if organizations work on instilling cultural norms that highlight the prosocial benefits of higher levels of agreeableness and conscientiousness. The study identified differences in both professional and cohort membership from the sample of computer professionals who self-identifying as either IT professionals, information security professionals, or penetration testers.

9. Limitations

This study was limited in scope. Greener (2018) noted that self-reported data cannot always be taken at face value and it can contain biases. The main limitations of the current study occurred because the respondents were selected from a convenience sample and because the study was conducted without any manipulation of the variables. Furthermore, the dataset was highly imbalanced and was rebalanced using a post-stratification random sampling of the responses from IT professionals (Levy & Lemeshow, 2013). The rebalancing is also a limitation with this study. According to Greener (2018), appropriate caution should, therefore, be used when viewing the results. Further limitations came from the choice of using of a small personality inventory that measured the Big Six traits using 24 questions. The Mini-IPIP does not measure any of the underlying facets that are part of the major traits. In comparison, the IPIP210 and IPIP300 instruments contain additional items that measure facets of personality, not just broad traits (Goldberg, 1999). The study assumed that the respondents were typical of their group of IT professionals, information security professionals, and penetration testers. There are also limitations with the model that informed the research. The six-factor model which is derived from the five-factor model, FFM with the addition of honesty/humility trait does not measure autonomy, and there is empirical evidence that autonomy differs from the FFM (Sheldon & Prentice, 2019).

10. Implications for Practice

The results showed generational differences between the cohorts of Generation X and millennials that might impact organizations hiring IT professionals, information security professionals, or penetration testers. The differences in personality traits between the Generation X and millennial cohorts could make it difficult for each generation to empathize with individuals from a different cohort, which could cause communication difficulties. The results, therefore, should inform human resource professionals in organizations that employ computer professionals about the possible generational differences between millennials and those in Generation X. Jones, Murray, and Tapp (2018) said that understanding the differences between generations in the workplace can improve recruitment and retention, could lead to the positive management of conflict, improve communications, and provide better planning for succession management. Practitioners of industrial-organizational psychology who support organizations that employ technical computer staff should note the differences in personality traits between the cohorts. Organizations seeking the services of penetration testers should understand the penetration testers in this study were generally more law-abiding than other IT professionals. This should alleviate some of the angst of hiring someone who has the skills necessary to breach the computer networks of the organization. It is better for the organization and society that an honest penetration tester, rather than a criminal hacker nation-state actor or unfaithful insider, have access to networks, systems, and applications.

11. Recommendations for Further Research

There are several areas for future research. Future research should investigate the causality of differences in personality traits between the cohorts to find if maturation causes the differences, or if differences in personality traits between millennials and those in Generation X are caused by generational differences in those who chose to become computer professionals. More research should be conducted on the personality traits of computer professionals using more comprehensive personality trait instruments that measure the Big Six personality traits to find sub-constructs for the Big Six traits. Such a study should also compare scores between computer professionals and the general population spanning the major cultures. This could elucidate if the differences found in this study generalize outside the United States. As argued by Parks-Leduc et al. (2015), such a study should also contain instruments that measure respondents' values. A follow-up study should be done to measure if there are differences in personality traits between those who self-identify as criminal hackers and those who self-identify as white-hat hackers. Future studies should also investigate the nonlinear relationships between professional affiliations and cohorts not studied here, which could be accomplished by using neural networks. Scarborough and Somers (2006) argued, for example, that artificial neural networks should be used in psychological analyses in organizational research. Future researchers should also determine who within the community of computer professionals remains law-abiding over time and compare them to those who choose to become criminal hackers.

12. Conclusions

The main conclusion was that the most important differences in personality traits were between cohorts. These were differences in agreeableness, conscientiousness, neuroticism, and the honesty/humility dimension. Caution should be exercised in the interpretation of the differences in conscientiousness between the cohorts due to the low Cohen's standardized alpha. There were unexpected significant differences between the professional groups for extroversion, neuroticism, openness, and honesty/humility. Expected differences between the groups on the openness trait and agreeableness trait did not materialize. Extroversion differences may contribute to the differences in career choices within the computer industry. This study identified that introversion is more common among penetration testers, compared to information security professionals who were the most extroverted of the groups. Of the three groups, penetration testers were the most honest, followed by information security professionals, with IT professionals scoring the lowest among the three groups for honesty.

There were important significant differences between cohorts in extroversions, agreeableness, conscientiousness, and neuroticism scores. The differences in the cognitive traits could be explained by differences in values between the cohorts, but that does not explain the difference in neuroticism. Parks-Leduc et al. (2015) noted that there was no relationship between emotionality (i.e., neuroticism) and any of Schwartz (2012) value constructs.

Personality traits and psychopathy have continued to capture the interest of psychologists and the population in general (Muris, Merckelbach, Otgaar, & Meijer, 2017), and the results of the current study identified a difference in the dark triad between types of computer professionals. The results showed penetration testers, in general, avoided manipulating others and (a) they did not behave in Machiavellian ways to obtain personal gain, (b) they had a propensity toward following rules, (c) they did not have a sense of entitlement, and (d) they did not seek social status. Information security professionals also scored significantly higher compared to IT professionals, yet that would be more in line with their chosen profession. The penetration testers also tended to be more introverted than their colleagues. Millennials scored lower in in extroversion, agreeableness, conscientiousness, and scored higher in neuroticism. Some of the results might be explained by maturation. Kandler (2012) and Reitz and Staudinger (2017) have noted that the stability of personality trait differences between individuals increases until middle age when it reaches a zenith and reduces in later life stages. The differences in personality traits for the professionals may translate into difficulties in understanding between the generations. Importantly, the results show differences between cohorts and professional affiliation in dark triad personality traits and neuroticism. These insights could spur collaboration between personality researchers, computer science researchers, political scientists, and IO professionals who share common interests in both normal and abnormal aspects of personality and self-reported dark triad traits among technical professionals. Based on the results of this study, anxiety reduction in the workplace for the computer industry professionals belonging to the millennial cohort seems to be warranted. The difference between cohorts should be investigated in a future study about cohort differences in personality traits among computer professionals.

The differences between the computer professionals for the constructs of agreeableness, conscientiousness, and neuroticism should inform educational and organizational policies to develop possible interventions for millennials in the workplace. The differences between the cohorts for honesty/humility warrants further study considering the negative impact the dark triad has on trust and cooperation in organizations. This study should also inform educators of the importance of teaching ethics for those majoring in computer science fields and computer engineering programs. The dark triad can have a negative impact on an individual's personal wellbeing and their interpersonal relationships (Muris et al., 2017). The literature review uncovered that changes in personality can be a function of age, biological function, and ecological influences (Kandler, 2012; Reitz & Staudinger, 2017). It is important to note that members from Generation X may be good mentors for the millennials for organizations that employ penetration testers, and the mentors should receive organizational support to model the behavior of agreeableness and conscientiousness. The results from this investigation show important differences in personality traits between millennial and Generation X computer professionals. The differences between generational cohorts could lead to both negative personal and organizational outcomes. If the differences are addressed, negative effects can be minimized. Educators, human resource professionals, and policymakers must understand the differences in personality traits between millennials and those in the Generation X cohort. Educators and human resource professionals should adjust to create policies with more emphasis on ethics in educational curricula.

References

Ashton, M. C., Lee, K., & De Vries, E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*, 139-152. doi:10.1177/1088868314523838

Bergmann, M. C., Dreißigacker, A., von Skarczinski, B., & Wollinger, G. R. (2018). Cyber-dependent crime victimization: The same risk for everyone? *Cyberpsychology, Behavior, and Social Networking, 21,* 84-90. doi:10.1089/cyber.2016.0727

Costa, P. T., & McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). In G. J. Boyle, G. Mathews, & D. H. Saklofske (Eds.), The *sage handbook of personality theory and assessment* (pp. 179-198). doi:10.4135/9781849200479.n9

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2018). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology.* Advance online publication. doi:10.1037/pspp0000210

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82-92. doi:10.1016/j.jrp.2014.10.009

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7*, 7-28. Retrieved from https://ipip.ori.org/A%20broad-bandwidth%20inventory.pdf

Greener, S. (2018). Research limitations: The need for honesty and common sense. *Interactive Learning Environments*, *26*, 567-568. doi:10.1080/10494820.2018.1486785

Jones, J. S., Murray, S. R., & Tapp, S. R. (2018). Generational differences in the workplace. *Journal of Business Diversity, 18*, 88-97. doi:10.33423/jbd.v18i2.528

Kandler, C. (2012). Nature and nurture in personality development: The case of neuroticism and extraversion. *Current Directions in Psychological Science, 21*, 290-296. doi:10.1177/0963721412452557

Ledingham, R., & Mills, R. (2015). A preliminary study of autism and cybercrime in the context of international law enforcement. *Advances in Autism, 1*, 2-11. doi:10.1108/AIA-05-2015-0003

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*, 329-358. doi:10.1207/s15327906mbr3902_8

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: methods and applications*. Hoboken, New Jersey: John Wiley & Sons.

Li, L., He, W., Xu, L., Ash, I., Anwar, M., & Yuan, X. (2019). Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. *International Journal of Information Management, 45*, 13-24. doi:10.1016/j.ijinfomgt.2018.10.017

Macdonald, C., Bore, M., & Munro, D. (2008). Values in action scale and the big 5: An empirical indication of structure. *Journal of Research in Personality, 42*, 787-799. doi: 10.1016/j.jrp.2007.10.003

Martens, M., De Wolf, R., & De Marez, L. (2019). Investigating and comparing the predictors of the intention towards taking security measures against malware, scams, and cybercrime in general. *Computers in Human Behavior*, *92*, 139-150. doi:10.1016/j.chb.2018.11.002

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*, 509. doi:10.1037/0003-066X.52.5.509

Menard, P., Bott, G. J., & Crossler, R. E. (2017). User motivations in protecting information security: Protection motivation theory versus self-determination theory. *Journal of Management Information Systems, 34*, 1203-1230. doi:10.1080/07421222.2017.1394083

Merhi, M. I., & Ahluwalia, P. (2019). Examining the impact of deterrence factors and norms on resistance to information systems security. *Computers in Human Behavior*, *92*, 37-46. doi:10.1016/j.chb.2018.10.031

Milojev, P., Osborne, D., Greaves, L. M., Barlow, F. K., & Sibley, C. G. (2013). The Mini-IPIP6: Tiny yet highly stable markers of Big Six personality. *Journal of Research in Personality, 47*, 936-944. doi.org/10.1016/j.jrp.2013.09.004

Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the dark triad (narcissism, machiavellianism, and psychopathy). *Perspect Psychol Sci, 12*, 183-204. doi:10.1177/1745691616666070

Parks-Leduc, L., Feldman, G., & Bardi, A. (2015). Personality traits and personal values: A meta-analysis. *Personality and Social Psychology Review*, *19*, 3-29. doi:10.1177/1088868314538548

Prentice, M., Jayawickreme, E., & Fleeson, W. (2019). Integrating whole trait theory and self-determination theory. *Journal of Personality*, *87*, 56-69. doi:10.1111/jopy.12417

Reitz, A. K., & Staudinger, U. M. (2017). Getting older, getting better? Toward understanding positive personality development across adulthood. In J. Specht (Ed.), *Personality development across the lifespan* (pp. 219-241). doi:10.1016/B978-0-12-804674-6.00014-4

Scarborough, D., & Somers, M. J. (2006). *Neural networks in organizational research: Applying pattern recognition to the analysis of organizational behavior.* Washington, DC: American Psychological Association. doi:10.1037/11465-000

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, *2*, 11. doi:10.9707/2307-0919.1116

Sheldon, K. M., & Prentice, M. (2019). Self-determination theory as a foundation for personality researchers. *Journal of Personality*, *87*, 5-14. doi: 10.1111/jopy.12360

Sibley, C. G., Luyten, N., Purnomo, M., Mobberley, A., Wootton, L. W., Hammond, M. D., . . . Wilson, M. S. (2011). The Mini-IPIP6: Validation and extension of a short measure of the big-six factors of personality in New Zealand. *New Zealand Journal of Psychology, 40,* 142-159. Retrieved from https://www.psychology.org.nz/wp-content/uploads/SibleyIPIP.pdf

Sibley, C. G., & Pirie, D. J. (2013). Personality in New Zealand: Scale Norms and Demographic Differences in the Mini-IPIP6. *New Zealand Journal of Psychology*, *42*(1). Retrieved from https://psycnet.apa.org/record/2013-12381-002

Torten, R., Reaiche, C., & Boyle, S. (2018). The impact of security awarness [sic] on information technology professionals' behavior. *Computers & Security*, *79*, 68-79. doi:10.1016/j.cose.2018.08.007

Valdes, A. M., Rhees, B., & Erlich, H. (2004). Reply to Kraft. *American Journal of Human Genetics*, *74*, 584. Retrieved from https://www.psychology.org.nz/wp-content/uploads/SibleyIPIP.pdf

Weems, C. F., Ahmed, I., Golden, G. G. R., III. Russell, J. D., & Neill, E. L. (2018). Susceptibility and resilience to cyber threat: Findings from a scenario decision program to measure secure and insecure computing behavior. *PloS One, 13*, e0207408.

doi/10.1371/journal.pone.0207408

Whalen, T., & Gates, C. (2007). A psychological profile of defender personality traits. *Journal of Computers*, *2*, 84-93. doi:10.1.1.64.2342

A Framework for a Foundational Cyber Counterintelligence Awareness and Skills Training Programme

Thenjiwe Sithole, Jaco du Toit, Victor Jaquire and Sebastian von Solms
Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa
thenjiwes@icloud.com
jacodt@uj.ac.za
jaquire@gmail.com
basievs@uj.ac.za

DOI: 10.34190/EWS.20.036

Abstract: Organisations from all sectors are confronted with cyber threats and cyberattacks of increasing sophistication and intensity. Now, even more than before, cyber attackers' modus operandi includes intelligence collection, espionage and influencing. Unsurprisingly then, conventional defence-only cybersecurity is no longer effective in protecting the confidentiality, integrity and availability (CIA) of information and information systems from these rapidly evolving cyber threats. Consequently, organisations have to evolve the protection of their information and information systems' CIA by implementing a proactive, approach that incorporates cyber counterintelligence (CCI) at its core. CCI includes both defensive and offensive activities. This approach is not only the purview state actors and their security apparatus, but (with certain qualifications) can significantly benefit the private sector and other non-state actors. Organisations' successful adoption of CCI depends on the implementation of an effective CCI Awareness and Skills Training (CCI AST) programme. Such a programme's utility and importance range from providing employees with the knowledge to proactively counter cyber threats; to CCI career pathing that involves the acquisition of more advanced CCI skillset. Despite its importance, there is limited published academic and industry literature on CCI awareness, education and training. This paper proposes a CCI AST framework that could assist organisations - irrespective of nature of business and size - to implement a foundational CCI awareness and skills programme. The CCI AST framework identifies and concisely describes the critical elements (fields of knowledge and skills) indispensable to a foundational CCI awareness and skills programme. The paper then proceeds with outlining the pitch and approach to be followed in transferring knowledge and skills in respect of each of the identified elements. The paper concludes with a summary of findings.

Keywords: Cyber counterintelligence, awareness, skills programme, CCI critical elements.

1. Introduction

As organisations from all sectors are increasingly reaping the benefits of digital advancement, so does the number of distinct cyber threats and cyberattacks increasing exponentially in sophistication and intensity. In addition to what is regarded as traditional, that is, attacks aimed directly at organisations, states and non-state actors now make use of social media to collect data and spread disinformation. According to the documentary, "The Great Hack", data is now regarded as a more valuable asset than oil (The Great Hack, 2019). Therefore, data can be collected through social media user data points in order to use the data to influence decision-making, cause reputational damage to organisations, destabilise countries, and influence masses through disinformation from fake news, deep fakes, satire news and memes (The Great Hack, 2019; Chesney & Citron, 2019). The increasingly pervasive data breaches (Check Point, 2020) and cyber propaganda (Bradshaw & Howard, 2019) that are felt in organisations of all sectors across the globe serve as an indication that organisations need to be pro-active to protect their critical assets.

The digital advancement is inevitable, and although it has significant challenges and cyber risks, these challenges and cyber risks can be mitigated. Conventional defence-only cybersecurity solutions are no longer effective and well-suited to deal with these rapidly evolving cyber risks. Consequently, organisations must evolve the protection of their critical information and information systems by implementing a proactive approach with cyber counterintelligence (CCI) at its core.

CCI can be defined as that subset of counterintelligence (CI) "aimed at detecting, deterring, preventing, degrading and exploiting and neutralisation adversarial attempts to collect, alter or in any other way breach the C-I-A of valued information assets, through cyber means and degrading adversaries' intelligence capabilities" (Duvenage, et al., 2019).

The effectiveness of the CCI's implementation requires a skilled CCI workforce as well as CCI-aware employees (Sithole, et al., 2019). This necessitates the implementation of an effective CCI Awareness and Skills Training (CCI AST) programme. This programme will increase awareness and knowledge to proactively counter cyber risks; enlighten organisations and people interested in the CCI field and its career pathing; and assist with the understanding of the knowledge, skills and attitude required for CCI workforce. CCI AST will play a central role in strengthening the CCI and cybersecurity posture of the organisation.

CCI is an emerging field that incorporates two distinct fields of expertise: counterintelligence and cyber (Black, 2014). Therefore, the design, development and implementation of the CCI AST come with challenges and complexity, such as:

1. Insufficient open-source information and research on CCI AST because CCI is an emerging multi-disciplinary field with limited published academic and industry research (Duvenage, et al., 2017).
2. CCI is multidisciplinary and requires a combination of several skillsets. Therefore, the design and development of CCI AST needs to take into consideration that, for example, a skilled cyber workforce or CI workforce does not necessarily have knowledge of counterintelligence or cybersecurity.
3. Design and/or content of CCI AST is not a "one-size-fits-all" but requires to be in line with each organisational training and competence needs, nature of business and strategic objectives (Sithole, et al., 2019).
4. Provide sufficient foundational information in specialised content, i.e. technical cyber or CI, without overwhelming the specific targeted groups with irrelevant information thereby obstructing the importance and relevance of CCI to the organisation.

A CCI AST framework can assist organisations, irrespective of nature of business and size, to implement a foundational CCI AST programme. As suggested by its title, the CCI AST comprises two main pillars, namely (i) CCI Awareness and (ii) CCI Skills Training. This paper's primary aim is to advance a proposition on CCI AST framework that incorporates both.

To achieve this aim, the rest of the paper is structured as follows: Section 2 identifies and concisely describes the critical elements (fields of knowledge and skills) indispensable to a foundational CCI awareness and skills programme. To this end, section 2: (i) points out the facets that are fundamental to the credible identification of critical elements (subsection 2.1) and the configuration of an integrated CCI AST programme (section 3); (ii) describe the cycle or the identification and refinement of CCI AST critical elements (subsection 2.2); (iii) identifies the critical elements of a CCI Awareness Programme (subsection 2.3); and (iv) identifies the critical elements of a CCI Skills Training Programme (subsection 2.4). Section 3 presents a structural outline of the integrated CCI AST framework for configuring the pitch and approach to be followed in transferring knowledge and skills in respect of each of the identified critical elements (in Section 2). This structural outline of the integrated CCI AST framework is a calibration instrument that allows for the configuration of the CCI AST in accordance with the particular needs of an organisation. Section 4 examine the effective transfer of knowledge and skills. The paper concludes with a summary of findings.

2. Critical elements for a CCI awareness and skills training programme

As was noted above, this section focuses on the identification of critical elements of which a CCI AST should comprise. We firstly describe the fundamental concepts and then describe the process to be followed for such identification. Subsequently, the critical elements (fields of knowledge and skills) indispensable to the CCI AST are identified.

2.1 Facets fundamental to the identification of CCI AST critical elements

The identification of critical elements of a CCI awareness and skills training programme should duly consider the imperatives posed by the fundamental concepts of knowledge, skills and attitude (KSA). The integration of knowledge, skills and attitude provides the requisite competencies to implement an effective CCI and CCI AST (Baartman & de Bruijn, 2011). Therefore, for the CCI AST to be effective, it is necessary for employees to attain acceptable levels in all three facets, namely knowledge, skills and attitude (Hassanzadeh, et al., 2014). Cyber risks are evolving at a rapid rate, requiring organisations to ensure that their employees persistently develop knowledge and skills and improve attitude. These three facets can concisely be described as follow:

1. **Knowledge:** "I know" – is what (facts, information and concepts) the employees need to know and understand in order to accomplish their tasks effectively.

2. **Skills:** "I can do" – is what employees need to do or perform (hands-on) to accomplish their tasks effectively.
3. **Attitude:** "I believe or I think I can" is the way of thinking, believe or feeling that enables the employee to use the knowledge and skills gained to accomplish their task. Attitude *can influences commitment, confidence, motivation, responsibility, performance and adaptability.*

The fundamental concepts of knowledge, skills and attitude not only impose imperative on the identification of critical elements of a CCI AST but are also used in Section 3 for the calibration of the integrated CCI AST programme.

2.2 Cycle for the identification and refinement of CCI AST critical elements

To determine the critical elements and have effective training, a systematic approach to the training cycle is required. The training cycle assists in determining if there is a lack of knowledge, skills and attitude. It also determines if there is performance gap in CCI and cybersecurity. It further sets training objectives, designs and develops a CCI awareness and skills training programme and lastly identifies training and assessments methods to use. Figure 1 graphically depicts our proposition on such a CCI AST programme training cycle:

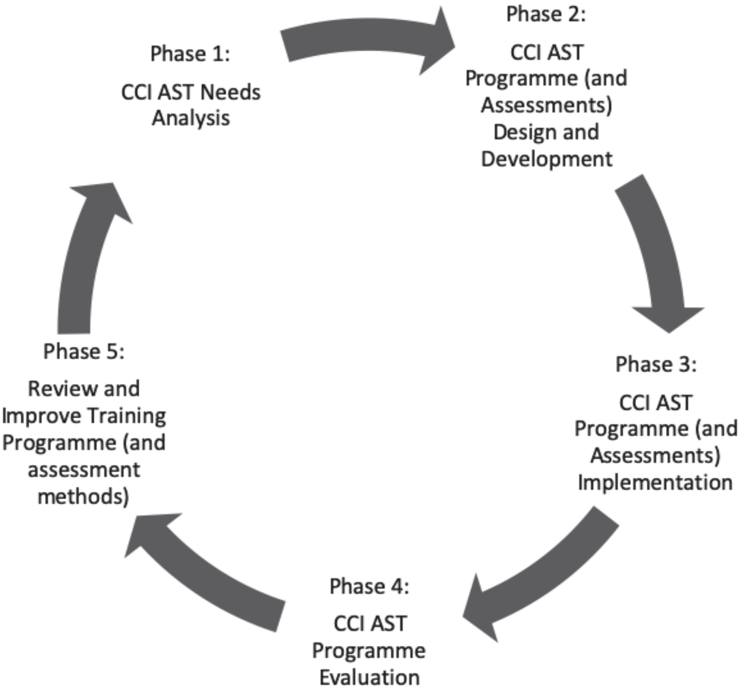

Phase 1:
CCI AST Needs Analysis
Phase 2:
CCI AST Programme (and Assessments) Design and Development
Phase 3:
CCI AST Programme (and Assessments) Implementation
Phase 4:
CCI AST Programme Evaluation
Phase 5:
Review and Improve Training Programme (and assessment methods)

Figure 1: CCI AST Programme Training Cycle (phases 1-4 adapted from MacCauvlei Learning Academy, 2016, and phase 5 included by the authors).

1. **Phase 1:** The training needs analysis phase determines the awareness and skills training required to fill the competence gap. The needs analysis seeks to determine the gap between the employee's current knowledge and skills status and the desired status required to improve employee competence and organisational capability (Ludwikowska, 2018; Zafar, 2016). The needs analysis looks at the knowledge and skills (and attitude/behavioural change) required by the organisation, job or task and individual. In this phase, an organisational cybersecurity and CCI culture in terms of awareness part of CCI AST can be looked at as well as determining the skills and capacity that the organisation has in terms of skills training part of CCI AST for re-skilling and up-skilling.
2. **Phase 2:** The programme design and development phase establishes the learning and training objectives and constructs learning outcomes. The learning objectives and outcomes must be aligned with the organisational goals and specifically with on-the-job functions/tasks. Subsequently, the design of the programme and the development of the training material and assessments. This phase also determines content, the programme duration and the training and assessment method (Sithole, et al., 2019). This

should not only develop the training material but also develop a well-defined training strategy and training plan.

3. **Phase 3**: The programme implementation phase, is where training is actually delivered. Various training methods can be used for the implementation of awareness and skills training, such as classroom, online, practical, simulations, on-the-job training and so on.
4. **Phase 4:** The programme evaluation and follow-up phase evaluate the effectiveness of the programme regarding learned knowledge and skills and the improvement of attitude. The evaluation of the training programme follows the Kirkpatrick's model of four levels of evaluation: reaction, learning, behaviour and results (Kirkpatrick & Kirkpatrick, 2007). Follow-ups are done in the workplace after a certain period to measure the sustainability of knowledge, skills and attitude (Sithole, et al., 2019).
5. **Phase 5:** The last phase reviews and improves the training programme according to the feedback received from the evaluation phase. This ensures that the training programme is current and relevant by incorporating new insight as technology changes rapidly, is proper for the target group, training duration and training delivery methods are correct. This phase also reviews, improves and ensure that the assessment methods are current.

This subsection outlined the process to be followed in the identification, review and refinement of CCI AST programme in general, and the identification of critical elements in particular. In the next two subsections, these critical elements are identified in respect of the CCI AST's two pillars, namely Awareness (sub-section 2.3) and Skills Training (subsection 2.4).

2.3 CCI AST- Awareness Programme: identification of critical elements

Awareness is the first line of defence for any organisation. Awareness aims to bring to attention security concerns and possible solutions to those concerns. The National Institute of Standards and Technology (NIST) Special Publication 800-16 defines awareness as "a learning process that sets the stage for training by changing individual and organisational attitudes to realise the importance of security and the adverse consequences of its failure" (de Zafra, et al., 1998). An awareness aims to encourage employees to practice safety precautions and apply countermeasures.

The objective of the CCI awareness programme is to introduce all employees to the fundamental concept of CI and CCI. This blended CI and cybersecurity awareness programme advises on the threats and vulnerabilities of using the Information and Communications Technology systems and social media, on how to identify and mitigate the threats (including insider threats) and vulnerabilities as well as suspicious activities, identify organisational critical assets. The awareness programme also ensures that everyone understands their roles and responsibilities.

CCI is regarded as the subset of a multi-disciplinary CI (Duvenage, et al., 2015; Black, 2014). CCI involves the usage of counterintelligence principles in the cyberspace (Jaquire, 2018). Therefore, the critical elements are identified from CI functions (defensive) (Prunckun, 2019). The CCI AST's awareness pillar will focus on CI activities because it relates to the cybersecurity domain. Considering the fundamentsls discussed in Section 2.1 and following the process explained in Section 2.2, the critical elements we propose for the CCI AST's awareness pillar are as follows:

1. Information Technology (IT) Fundamentals.
   An introduction to basic IT knowledge and skills to develop an understanding of IT concepts: IT terminology, hardware, networks, software, cloud concept, IT risk and security, IT roles and responsibilities.
2. Introduction to CI, Cybersecurity and CCI
   An introduction to the concept of counterintelligence, cybersecurity and cyber counterintelligence. Putting CI in CCI.

3. CCI: Cyber risks
   An introduction to the concept of risk management and the components of risk management framework. It also provides learners with understanding of how to identify, assess and mitigate cyber risks to the organisational critical assets to safeguard the confidentiality , integrity and availability of the critical assets.
4. CCI: Personal Security

5. Introduces learners to cyber threats and vulnerabilities to personal security, the importance of safeguarding personal identifiable information (PII) and the application of personal cybersecurity defences.
6. CCI: Personnel (Targeting HR, finance, or any structure dealing with confidential PII)
   It introduces the importance of handling and classifying documents that have information regarded as sensitive PII, data privacy and data security.
7. CCI: Computer and Internet Security
   Introduction to the threats and vulnerabilities to computers and internet as well as the best practices for computer and internet security.
8. CCI Mobile Security
   Introduction to the threats and vulnerabilities to mobile networks (e.g. Wi-Fi and Bluetooth) and devices. Understanding of defences for mobile network, including home Wi-Fi network, and security for mobile devices.
9. CCI: Social Medial Security
   An introduction to social media risks and security measures to address social media risks for personal and organisation.
10. CCI: Communications Security
    An introduction to the vulnerability of communication systems to cyber-attacks or information warfare and the security measures for voice data and video communications over the internet.

2.4 CCI AST Skills Training Programme: identification of critical elements

Organisations and employees need to be made aware of the CCI field. Therefore, the skills training programme draws attention to the skills required and/or career opportunities within the CCI field. This programme introduces different disciplines within the CCI field, of which each discipline will require its training. This provides organisations with an opportunity for re-skilling and up-skilling. The skills programme focus is entirely on the basic concept of CCI and the disciplines as well as knowledge building.

CCI is regarded as the subset of a multi-disciplinary CI (Duvenage, et al., 2015; Black, 2014). CCI involves the usage of counterintelligence principles in the cyberspace (Jaquire, 2018). Therefore, the critical elements are identified from blending CI functions (offensive and defensive dimensions) (Prunckun, 2019) and cybersecurity. This paper also looked at the suggested CCI roles (Jaquire, 2018; Black, 2014). Considering the fundamentsls discussed in Section 2.1, following the process explained in Section 2.2 and the critical elements for CCI AST's awareness pillar identified in Section 2.3, the critical elements we propose for the CCI AST's skills training pillar are as follows:

1. Introduction to Cyber Intelligence(understanding digital intelligence) and cyber counterintelligence
2. Fundamentals of CCI Investigation
3. Fundamentals of CCI Collection
4. Fundamentals of CCI Analysis
5. Digital Forensics Fundamentals
6. Introduction to Malicious Software and Ethical Hacking
7. Critical Information Infrastructure Protection
8. Cryptography Fundamentals

This section examined the fundamentals and processes to be followed in the identification of critical elements. We then proceeded with identifying the critical elements in respect of both the CCI AST's main thrusts, namely Awareness and Skills Training. The next section presents the CCI AST's structural outline.

3. A structural outline of the CCI AST framework

The structural outline of the CCI AST framework provides an instrument for configuring the pitch and approach to be followed in transferring knowledge and skills. The structural outline therefore features all the elements identified in the preceding section. To aid the CCI AST's configuration to the needs of a specific organisation, the structural outline adds a matrix for calibrating 'knowledge', 'skills', and 'attitude' (Section 2) on the basic, intermediate and advanced levels. In tabulated format, the structural outline of the CCI AST framework is as follows:

Table 1: A structural outline of the CCI AST framework

Critical Elements	Level of content delivery (Pitch)								
	Knowledge			Skills			Attitude		
	B*	I*	A*	B*	I*	A*	B*	I*	A*
CCI Awareness Programme									
CCI: Cyber risks									
CCI: Personal Security									
CCI: Personnel (Targeting HR, Finance)									
CCI Workplace Security									
CCI: Computer and Internet Security									
CCI Mobile Security									
CCI: Social Medial Security									
CCI: Communications Security									
CCI: Cyber risks									
CCI: Personal Security									
CCI Skills Programme									
Introduction to Cyber Intelligence (understanding digital intelligence) and cyber counterintelligence									
Fundamentals of CCI Investigation									
Fundamentals of CCI Collection									
Fundamentals of CCI Analysis									
Digital Forensics Fundamentals									
Introduction to Malicious Software and Ethical Hacking									
Critical Information Infrastructure Protection									
Cryptography Fundamentals									

*B = Basic; I = Intermediate; A = Advanced.

The level of the content that the CCI AST programme is pitched at varies according to the target group and their background. The level of pitching will be determined after a pre-assessment is conducted before the delivery of the programme. How well-structured and designed, the success of CCI AST ultimately depends on the optimal transferring of skills and knowledge. This is thus the focus of the next section.

4. Framework for transferring of knowledge and skills

Organisations leverage training as a way of increasing the required knowledge, skills and attitude to achieve organisational goals and on-the-job functions/tasks. However, the effectiveness of training will be accomplished when knowledge and skills learned is applied to the real execution of the work. Transfer of knowledge and skills allow employees to apply what is learned in awareness and training into the workplace. Despite organisations investing in training, some research shows that transfer of training into the workplace is not efficient as no more than 15-20% of knowledge and skills learned in training are used in the workplace (Hughes, et al., 2018; Faizal, et al., 2017; Martin, 2010). Effective and efficient transfer of knowledge and skills in the workplace should be directed by the training strategy. Additionally, properly following the training cycle ensures that the transfer of knowledge and skills is efficient. The following are the primary factors identified that influence the transfer of critical elements (field of knowledge, skills and attitude), aligned against the training cycle:

4.1 Needs Analysis

Buy-in and commitment must be obtained from top management, not only for needs analysis and performance gap assessment, but for the approval of the training programme outcomes (knowledge, skills, and attitude) and material (from phase 2 of the training cycle). Functional management, subject matter experts and employees must be properly consulted so that the design and development of the training programme is the representation of the knowledge, skills and attitude required to effectively accomplish the tasks. Individual analysis, job analysis and organisational analysis must be conducted.

4.2 Programme design and development

Appropriate design and development of the training programme are critical as it is a major contributor to the knowledge and skills transfer (Martin, 2010). Learning objectives and outcomes must meet the needs of the learner (employee), the structure concerned and the organisation. The programme must have a well-defined training strategy to outline and promote training transfer (Martin, 2010; Chidananda & Udayachandra, 2018).

4.3 Programme Implementation

The use of real-life scenarios and relevant examples is critical and must be encouraged. The training programme must have appropriate tools (technology, policies, etc.) that are used in the organisation and also make use of effective training techniques and training g aids. Inclusion of practical exercises, if possible, is highly recommended, more especially if it involves technology, e.g. IT Fundamental training. The training programme must demonstrate to the learners the relevance of training to their job.

4.4 Programme evaluation and Follow-up

This factor is a critical step as it provides an opportunity to receive feedback from both the learners and their management. It further provides an opportunity to determine if the objectives of the training were met and that employees are or will be competent to do their functions effectively and efficiently after the training.

- The reaction of the learners to the training provides with the understanding of (i) learners' satisfactory to the training, (ii) learners' achievability of training objectives and outcomes, and (iii) learners' recommendations on improving the training.
- Understanding if the learners improve their knowledge and skills and if the attitude has changed.
- Understanding if learners are transferring the knowledge, skills and attitude acquired when they are back at the job.
- If overall performance or targeted outcomes are achieved as a result of training.
- If management ensures what is learned from training is applied on the job, performance of learners is assessed and feedback is provided to the training structure to improve the training programme.

This section described the primary factors that influence the transfer of knowledge and skills learned from training to the workplace.

5. Conclusion

Cyber counterintelligence is a proactive approach solution to mitigate the evolving cyber threats and attacks. Effective CCI implementation depends on CCI-aware employees and properly skilled CCI workforce. This paper proposes a CCI AST framework that could assist organisations, regardless of the nature of business and size, in implementing a foundational CCI awareness and skills training programme. The CCI AST framework consists of critical elements (fields of knowledge, skills and attitude) indispensable to a foundational programme. It encompasses technical and non-technical sub-disciplines and focuses on both CCI dimensions – offensive and defensive. Our proposition on an integrated CCI AST framework allows for the pitch and approach to be configured in line with the organisation's needs. The integrated CCI AST framework is thus a theoretical construct with practical application.

This research benefitted, in part, from support from the Faculty of Science at the University of Johannesburg.

References

Baartman, L. K. & de Bruijn, E., 2011. Integrating knowledge, skills and attitudes: Conceptualising learning processes towards vocational competence. *Educational Research Review,* 6(2), pp. 125-134.

Bradshaw, S. & Howard, P. N., 2019. *The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation.* [Online] Available at: https://comprop.oii.ox.ac.uk/research/cybertroops2019/ [Accessed 12 December 2019].

Black, J. M., 2014. *The Complexity of Cyber Counterintelligence Training, Master of Science in Cybersecurity, Utica College. ,* s.l.: Unpublished.

Check Point, 2020. *2020 Cyber Security Report.* [Online] Available at: https://www.checkpoint.com/downloads/resources/cyber-security-report-2020.pdf?mkt_tok=eyJpIjoiTm1FNU5qRTNPV1F4TmpoaSIsInQiOiJlOGNlalQ2VGdodzJTWFFtejZZcUV2SXFHb3pncERwUzFwSisybWdYdXBjbllzb3VUZlVyK3NPTlwvcVlYWXhMU05VZ2xoWXlmVGVIdWF3c1dOVDdjOWtDMGd5NEk3NExOV [Accessed 16 January 2020].

Chesney, B. & Citron, D., 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review,* Volume 107, pp. 1753-1820.

Chidananda, H. L. & Udayachandra, P. N., 2018. Relationship of Training Design and use of Transfer Strategy with Transfer of Training. *International Journal of Management Studies,* 4(4), pp. 32-37.

Duvenage, P.C., Jaquire, V. J. & von Solms, S.H. (2019) 'A cyber counterintelligence matrix for outsmarting your adversaries' in *Published Proceedings of the 18th European Conference on Cyber Warfare and Security*, Coimbra, Portugal.

Duvenage, P., Sithole, T. & von Solms, S., 2017. *A Conceptual Framework for Cyber Counterintelligence – Theory That Really Matters.* University College Dublin, Ireland, 16th European Conference on Cyber Warefare and Security.

Duvenage, P., von Solms, S. & Corregedor, M., 2015. *The Cyber Counterintelligence Process - a Conceptual Overview and Theoretical Proposition.* Hatfield, UK, 14th European Conference on Cyber Warfare & Security.

Faizal, A. N. Y. et al., 2017. Impact of Work Environment on Learning Transfer of Skills. *Social Sciences & Humanities ,* 25 (S), pp. 33-40.

Hassanzadeh, M., Jahangiri, N. & Brewster, B., 2014. A Conceptual Framework for Information Security Awareness, Assessment, and Training. In: B. Akhgar & H. R. Arabnia, eds. *Emerging Trends in ICT Security.* Waltham: Morgan Kaufmann, pp. 99-110.

Hughes, A. M., Zajac, S., Spencer, J. M. & Salas, E., 2018. A checklist for facilitating training transfer in organizations. *Journal of Training and Development ,* 22(4), pp. 334-345.

Jaquire, V. J., 2018. *A Framework for a Cyber Counterintelligence Maturity Model. ,* s.l.: [Unpuplished]: University of Johannesburg. Retrieved from: https://ujcontent.uj.ac.za/vital/access/manager/Repository/uj:28690?site_name=Research+Output&exact=sm_creator%3A%22Jaquire%2C+Victor+John%22 (Accessed: 12 November 2019).

Kirkpatrick, D. L. & Kirkpatrick, J. D., 2007. *Implementing the Four Levels: A Practical Guide for Effective Evaluation of Training Programs.* San Francisco: Berrett-Koehler Publishers, Inc..

Ludwikowska, K., 2018. *The effectiveness of training needs analysis and its relation to employee efficiency,* s.l.: Zeszyty Naukowe Politechniki Poznańskiej. Organizacja i Zarządzanie.

MacCauvlei Learning Academy, 2016. *Higher Certificate in Occupational Directed Education, Training and Development Practices,* Pretoria: unpublished.

Martin, H. J., 2010. Improving Training Impact Through Effective Follow-Up: Techniques and Their Application. *of Management Development,* 29(6), pp. 520-534.

Prunckun, H., 2019. *Counterintelligence Theory and Practice.* 2nd Edition ed. Maryland: Rowman & Littlefield Publishing Group, Inc..

Sithole, T., Duvenage, P., Jaquire, V. & von Solms, S., 2019. *Eating the Elephant - A Structural Outline of Cyber Counterintelligence Awareness and Training.* Stellenbosch, South Africa, ICCWS 2019 14th International Conference on Cyber Warfare and Security: ICCWS 2019.

The Great Hack. 2019. [Film] Directed by Karim Amer, Jehane Noujaim. United States of America: Netflix.

Zafar, S. S., 2016. *Training Needs Analysis (TNA) in the Organization.* [Online] Available at: https://www.academia.edu/30241722/Training_Needs_Analysis_TNA_in_the_Organization [Accessed 16 January 2017].

The Three "Cyber Logics" and Their Implications to Cyber Conflicts: Germany, France, and the United Kingdom as Case Studies

Bruna Toso de Alcântara
Federal University of Rio Grande do Sul, Porto Alegre, Brazil
Alexander von Humboldt Institute for Internet and Society, Berlin, Germany
bruna.toso@ufrgs.br

DOI: 10.34190/EWS.20.063

Abstract: The crescent use of cyberspace attached to daily activities and critical infrastructures, makes states worry about cyber-attacks and develop certain cyber operations to reactively or actively counter them. The picture gets more complicated when a variety of paradoxes appears in front of the state's conventional actions related to sovereignty, offensive, defensive, and liability issues. To diminish these paradoxes, while countering cyber-attacks, states ended creating power relations not only among themselves but also with other international actors, especially companies. These power relations ultimately shape the degree of stability of cyberspace, which means influencing the openness and freedom of it, mainly in the application layer. Thus, one may wonder how these power relations can develop to make what now has been considered the fifth domain of war more or less prone to conflict escalation. This paper makes the hypothesis that if the cyberspace dynamics allows hierarchy power relations among states the way this hierarchy will be seen by them can be shaped with either a cooperative (friendship), competitive or enmity (conflict) logic, causing the international and digital system to have a interweaving along the lines of what Wendt (1999) spells out with the three logics of anarchy of the International System. Thus, if cyberspace is a manmade space and perceptions of actors can shape it, what in states' perception could result in a more cooperative (Kantian), competitive (Lockean), or enmity (Hobbesian) action pattern? Using qualitative document analysis from official cyber documents, within some European states (i.e., Germany, France, United Kingdom) this paper, aims to track the features encompassing states logics when positioning toward cyberspace ("cyber logics") and determines which one is predominant. The structure of the paper will start with a theoretical review of the anarchy logics, passing to the analysis of the logic in the cases chosen and finally comparing them, defining which one seems to be dominant and its consequences for conflict escalation in cyberspace.

Keywords: Cyberspace, Power, Anarchy, Conflict, Europe

1. Introduction

The crescent dependence on technology attached to the cyber realm has been changing societal relations. These changes encompass economic, political, and military spheres creating paradoxes between the physical and digital world. These paradoxes can be summarized as follows: (a) between the regularization of cyberspace and its original proposal for freedom/openness; (b) between a physical infrastructure (e.g., wired means) that passes through sovereign territories and the virtual part of cyberspace, which does not consider borders, involving many countries and, therefore, conventional legislation when it comes to actions in this field;(c) between connectivity and degree of vulnerability, because the more connected the country, the more vulnerable are the threats of cyberspace since one cannot speak of absolute security in the virtual environment. Thus, issues of vulnerability, freedom, and sovereignty in the face of cyber-attacks pose challenges on how to possess, store, and internationally project power in the arena of cyberspace.

The scenario gets, even more, complicate when one realizes that to think about power relations in cyberspace is to accept not only power diffusion since states are no longer the sole or the most relevant security providers - as 90% of government communications including some military traffic travels via private network- but also accept that technology is difficult to understand and control - as it changes faster than policymaker and analysts are used to - and that there is an expanding spectrum of cyber events lying between the definitions of war and peace (Kello,2017). A concrete example of this imbroglio is the relationship between Telegram, a popular online messaging service, and countries such as Iran and Russia. Both countries demanded access to the platform's encrypted content and, upon its refusal, took measures to ban it. These intervention attempts were justified on a security logic basis, once the platform was accused of acting as a facilitator of terrorist activities by offering anonymity and data protection to its users, also generating, as Iran stated, moral decadence. Akbari and Gabdulhakov (2019, p. 230) explain the Telegram case "not only presents clashes between non-democratic states and users who struggle to access free flows of information but also highlights important issues about platform independence, alternative commercial models of platform development, and the future of platform surveillance across various political contexts."

This type of cyber events reflect the development of cyber capabilities that according to Perkovich and Levite (2017 p. 250) "are uniquely protean as instruments of intelligence gathering and coercion," being able to be used in a variety of ways to a wide range of objectives, including "intelligence collection, political- psychological warfare, deterrence signaling, discrete sabotage, combined- arms military attacks, and campaigns of mass disruption." The spectrum of cyber events is even more concerning as it is fomenting the weaponization of software and hardware triggering a cyber arms race between states and state and non-state actors, in a yet not fully legislated domain, and leading collective imagination into the cantered in the idea of cyberwar or cyber warfare (Prunckun, 2018).

The insecurity around the cyber domain is increased when one realizes that it remains in some sense anarchic (Riordan, 2019), and that as different sectors of society are impacted by high-level cyber incidents (i.e., individuals, business, sates) states are called upon to managed security issues, bind by a social contract that gives them the monopoly of force. As Riordan (2019, p.13-1144) explains "the cybersecurity agenda ensures the continued relevance and importance of the state," once only states, or state backed-groups, have the technical and financial resources to (1) penetrated well-protected targets and (2) defend critical infrastructure against cyberattacks and only them (3) have the legitimacy to penetrate foreign computer systems as to ascertain capabilities and motivations. Indeed, considering the Internet (i.e., the application layer of cyberspace) an institution, Drezner (2004, p. 482) proposes that states, "will naturally prefer that global regulations mirror their own national standards" since most social issues originated as domestic problems before globalization made them international matters. State dynamics would encompass not only economical and preference variables but also the bargaining process generated by the preferences of lesser power or peripherical states (as an intervening variable).

Following this reasoning, one may highlight some recurrent ideas involving cyberspace: anarchy, power, and interstate relations. These features encompass the core concepts of the study field of International Relations (IR). Thus, in IR theory, there may be answers to explain or even predict state behavior in cyberspace.

Indeed, realists, as the mainstream theory of IR, already developed some thoughts on the issue. For these scholars, cyberspace is following (cyber)security dilemma logic "whereby defensive actions in cyberspace by a defender can be perceived as offensive by an attacker, leading it, in turn, to take defensive action the original defender views as aggressive" (Grigsby, 2017). As a matter of fact, "the core of the cybersecurity dilemma is about fear and escalation: the fear that causes the dilemma to arise and the escalation that the dilemma potentially brings about. This pattern of fear and escalation shapes policy-makers' decisions" (Buchanan, 2016, p. 193). Furthermore, realists explain power in cyberspace as another tool of national power, seeking to define ways in which cyberspace (considered a 'new domain' of war) could best serve as an extension or complement to it, in all war domains (land, sea, air, space, cyberspace) and within the aspects of P / DIME (diplomatic/political, informational, military and economic fields) (Calvety, 2018a). They transfer the military/strategic thinking into interstate cyber relations when they propose that states employ cyber strategies (i.e., disruption, espionage, and degradation) to gain a position of advantage relative to their rivals" and affirm that coercion in cyberspace rarely produces concessions (Valeriano, Jensen, Maness, 2018).

However, despite the relevance of the state in cyberspace and, more specifically, in cybersecurity, realists fail to pay attention to the digital domain *per se*. After all, cyberspace is humanmade, thus has a clear and substantial intersubjective component in it. Cyberspace itself encompasses a sociological approach since it can be viewed as operating in an inverted four pyramidal layer structure. Being the layers that form cyberspace: hardware (physical network), logic (how internet functions and how information is distributed through it), data (information flowing around the Internet), and social (where humans interact with the Internet) (Riordan, 2019). The view of cyberspace as socially constructed, and subjected to systemic analysis, is not new. It can be found in the so-called Social Construction of Technology (SCoT), present in the philosophy of technology. For which technology is considered not only technological artifacts (hardware) but also their use. For SCoT, technology is neither a neutral tool for problem-solving (i.e., instrumentalist view of technology) nor a value-laden force that threatens human autonomy (deterministic view of technology). For SCoT, technology would be shaped and influenced by society and should be considered as an expression of norms and expectations within society (Carr, 2016). Thus, "the starting point for SCoT perspectives in International Relations is the recognition that technology must be conceptualized as part of the international political system" (Mccarthy 2015, p.33).

Moreover, International Relations theory has a branch that deals with immaterial elements and the co-constitution between the international system and actors: constructivism. This approach looks at cyberspace with social lens, typically focusing on non-state interactions and emphasizing symbolic politics (Devi; Rather, 2019). However, one may notice a change in the constructivist research agenda on cybersecurity issues. From 2000's until 2010, this IR theory focused on information and the discourse itself, encompassing analysis on the securitization process of cyberspace and concepts more linked to information warfare and cyber terrorism, and highlighting identity boundaries (Choucri, Reardon; 2012, Der Derrian 2000, Eriksson; Giacomelli, 2006, Hansen; Nissenbaum, 2009). But, since 2010 till the present date, besides identity formation, there seems to be an increasing focus by constructivists on better explaining the role of norms in cyberspace and states and non-states interest formation and strategic motivations (Calvety, 2018b; Ciolan,2014; Schulze, 2018; Cohen, 2018). Thus, there are gaps that this new research agenda seeks to cover, leaving room for applying constructivism perspectives into different cyberspace dynamics, including its structure.

In sum, Riordan (2019, p. 105) is right when he states that "in many ways, the nature and structure of the problems in cyberspace mirror those in physical space," and that one of the main difficulties of constructing some order and stability in cyberspace reinforces the requirements of the transition to a multipolar physical space. Scholars, like Kello (2017), are right when highlighting the political and technological components of cyberspace, describing that the growing multipolarity in cyberspace has two faces: (1) general realignment of power in the system and (2) the unbeaten proliferation of cyber arms. Notwithstanding, if cyberspace is a human construction that at the same time mirrors International System dynamics (especially its anarchical and hierarchical dimensions) and invokes new and unique perceptions among its actors, this means that actors can change its dynamics intentionally. Thus, the idea of reality construction from the constructivist IR approach seems to me the best option to analyze what states are making of cyberspace. This paper will draw on this idea, seeking to fill the gap between constructivist theory and state interest's generation toward cyberspace. In particular, this paper proposes to look at cyberspace will tackle the anarchic structure of cyberspace through Wendt's (1999) theory and propose an analysis that goes beyond a deterministic Hobbesian view. Thus, the present paper asks what in states' perception could result in a more cooperative (Kantian), competitive (Lockean) or enmity (Hobbesian) action pattern?

To answer this question, the paper goes specifically into Europe, as a region of cyber relevance, since according to the Global Cybersecurity Index 2018 (ITU, 2019) it scored the highest ranks in comparison with other world regions in all five pillars analyzed (e.g., legal, technical, organizational, capacity building, and cooperation). In particular, the paper will focus on three study cases: Germany, France and the United Kingdom, as they are among the countries seen as "leaders" on cybersecurity issues within the European Union (Denninson et al., 2018) and as traditional regional powers in a shifting EU, they may be good indicators of different cyber threat perceptions, which may materialize into different influence directions within the region. Concretely, using qualitative document analysis from official cyber documents, the paper aims to identify in three study cases the features encompassing states logics when positioning toward cyberspace ("cyber logics") determines which one is predominant and its consequences for conflict escalation in cyberspace.

2. Wendt and the logics of Anarchy

According to Wendt (1999, p.366), "The social process consists of interlocking actions seeking to satisfy identities and interests by adjusting behavior to changing incentives in the environment. The constructivist model assumes that agents themselves are in process when they interact." Thus, social construction shapes the environment within which international actors interact, and the International System instead of static becomes a dynamic environment. Relevant here is that identities are shaped through social acts - that are, processes of signaling, interpreting, and responding to the Other, creating shared knowledge and social learning. According to Wendt (1992a, p. 397), "Actors acquire identities-relatively stable, role-specific understandings and expectations about self-by participating in such collective meanings."

Thus, although Wendt (1999) considers culture as a self-fulfilling prophecy (i.e., culture reproduces itself, creating stability and being stronger in the degree it is internalized), he stresses that identity is relational, being the base of state interests (which in turn will be defining situations) and able to change accordingly to the shared culture that exists. This features are relevant, because a "world in which identities and interests are learned and sustained by intersubjectively grounded practice, by what states think and do, is one in which "anarchy is what states make of it'" (Wendt, 1992b, p.183) Wendt argues that anarchy is a social construct, having no previous logic but instead different shared cultures (i.e., cultures of anarchy). He considers anarchy "an empty vessel," which "only acquire logic as a function of the structure of what we put inside them" (Wendt 1999, p.249). As Guzzini and Leander (2006, p.77) explain, "once states are understood as sharing a culture, their very identity is open to conceptualization, and not only their utility calculus." What matters then in the balance of power is what policymakers actually take to be existing (Kazmi, 2012) and "distribution of power may always affect states' calculations, but how it does so depends on the intersubjective understandings and expectations, on the "distribution of knowledge," that constitute their conceptions of self and other" (Wendt 2011, p. 180)

Therefore, for Wendt (1999), there are three different collective identities in the international system which provide the "broad scope conditions for understanding the agency role-identities (from enemy to friend), and hence the behavior of states" (Guzzini, Leander, 2006). These identities come from structural cultures, and they are based on the roles that dominate the international system: enemy, rival, and friend. The Hobbesian culture is attached to a competitive security system in which "states identify negatively with each other's security, so that ego's gain is seen as alter's loss" (Wendt 2011, p. 181). In this culture, enmity prevails between states leading to the will of utilizing unlimited violence due to the perception of lack of existential autonomy [sovereignty] of the other (Wendt 1999, p.260). There is an egoistic element behind the behavior, as enemies' intentions are perceived as being unlimited, tending to favor security dilemmas "in which the efforts of actors to enhance their security unilaterally threatens the security of the others, perpetuating distrust and alienation" (Wendt 2011, p.183). This logic represents the true "self-help" of realists, where what matters for survival is military power. Four main macro-level tendencies appear: (1) war may be unlimited,(2) the "unfit" actors are eliminated, (3) balances of power may exist but will be challenging to sustain and (4) system will tend to suck all of its members into the fray, making non- alignment or neutrality very difficult (Wendt 1999, p. 265-266)

The Lockean culture is entrenched in an individualistic security system (Wendt, 1992a; 2011) in which states are indifferent to the relationship between their own and others' security. Rivalry prevails between countries, making them still egoistic but on a lesser level. Thus in both enemy and rival systems "power politics within such systems will necessarily consist of efforts to manipulate others to satisfy self-regarding interests"(Wendt 2011, p.181) However, "rivals expect each other to act as if they recognize their sovereignty, their ``life, and liberty,'' as a right, and therefore not to try to conquer or dominate them" (Wendt 1999, p. 279).

Additionally, when sovereignty recognized as a right, it becomes not only property but an institution. Thus "Whether out of self-interest or the perceived legitimacy of its norms, the members of a well-functioning society must also restrain themselves" (Ibid, p.280). In this system culture Wendt emphasizes that rivals seek a shallow revisionism (i.e., they recognize the Self 's right to life and liberty, and therefore seeks to revise only its behavior or property) and thus the rivals' intentions are limited (Ibid, p.261) Additionally, he highlights that rivalry is the collective representation of "rival" as a "preexisting position in a stock of shared knowledge that supervenes on the ideas of individual states." Thus, when rivalry acquires this collective status, "states will make attributions about each other's ``minds'' based more on what they know about the structure than what they know about each other, and the system will acquire a logic of its own" (Ibid, p.283). That is why other four macro-level tendencies appear: (1) configurate wars (i.e., the parties accept the units, who are fighting over territory and strategic advantage) (2) warfare is limited (3) the balance of power can exist and are more stable (as it does not form itself out of survival necessity) and (4) neutrality or non-alignment becomes a recognized status (Wendt, 1999, p.283-5).

The Kantian culture is entrenched in a cooperative security system (Wendt 1992a; 2011) in which states identify positively with one another so that the security of each is perceived as the responsibility of all. Friendship prevails in this system generating "role structure within which states expect each other to observe two simple rules: "(1) disputes will be settled without war or the threat of war (the rule of non-violence), and (2) they will fight as a team if the security of anyone is threatened by a third party (the rule of mutual aid)"(Wendt 1999, p. 298-9). Moreover, this system is characterized by "pluralistic security communities" (disputes within a group,) and "collective security" (disputes between a group and outsiders) in which the norm is based in reciprocity, generating at the same time a tendency of multilateralism or other-help concerning national security (Ibid, p. 299-300) and, depending on how well developed the collective self is, making collective action less dependent on the presence of active threats and less prone to free-riding (Wendt, 2011). As Senyuk (2018) summarizes, states in this system are aware of the external restrictions imposed on them by the others and are capable of imposing self-control on the use of violence. The efforts thus advance one's objectives, or "power politics," in terms of shared norms rather than relative power (Wendt 2011, p.181).

Guzzini and Leander (2006) explain that Wendt, in his encompassing view, "having based the theory on the relentless (re)construction of collective identities," needed a process component. Thus, Wendt (1999) seeks to deliver this component through a 3X3 matrix, crossing the anarchy cultures with the degrees of rule internalization, as accordingly to him, the different degrees of rule internalization generates pathways by which anarchic structures can be produced. These degrees would be in crescent order: (a) force (agents are forced to follow the rules), (b) price (agents follow the rules out of self-interest), and (c) legitimacy (agents want to follow the rules). However, Wendt (1999, p.250) explains that "is only with the third degree of internalization that actors are really ``constructed'' by culture; up to that point, culture is affecting just their behavior or beliefs about the environment, not who they are or what they want." Thus, each logic of anarchy becomes "multiple realizable," which means the same effect can be reached through different causes, the pathway lying on empiricism.

In sum, since Wendt (1999) proposes that the culture of an international system is based on a structure of roles (enemy, rival and friend), which will influence the tendencies and structure of anarchy, besides highlighting that states will be under corresponding pressure to internalize that role in their identities and interests. He provides a model that can be imported to the anarchic structure of cyberspace, as it resembles the interactions of the international system itself. Therefore, the next section takes Wendt's reasoning, and trough the analysis the study cases seek to identify which one of the three cyber- logics is shaping the "cyber anarchic culture."

3. European case studies

Germany, France, and the United Kingdom were chosen as study cases as they represent dominant narratives at both regional and international levels, especially regarding cyberspace.

3.1 France

The first document to focus directly on the nation's cyber threats as a critical risk to the country's security and sovereignty was the 2008 White Paper on National Defence and Security. From there to the present time its is possible to track a centralization of cyber policies with the development of a civil-centered institutional framework, concentrated in the agency created in 2009 titled ANSSI (i.e., *Agence Nationale de la sécurité des systèmes d'information*), which is responsible for (1) assisting the state's institutions on issues of cybersecurity, (2) organizing cybersecurity standards for industries and critical infrastructures, and (3) organizing awareness campaigns and education for civilians (Brangetto, 2015; Baezner, 2018).

This civil-centered structure does not mean that France is not concerned with military developments in cyberspace. Indeed, its first national cybersecurity strategy (i.e., Information Systems Defence and Security: France's Strategy) stated the will for France to become a "world power in cyber defense," and strengthened the authorities of ANSSI. The White Paper (2013) declared the development of cyber offensive capabilities. In 2014 the Defence Minister confirmed in an interview that cyber weapons could be used as support for conventional forces or in response to traditional attack (Laudrain, 2019) an idea reinforced in 2019, by Florence Parmy (2019), the current Armed Forces Minister, when she stated that France is "not afraid" of using cyberweapons and by the release of the country's first doctrine for offensive cyber operations (Ministère des Armées, 2018). However, and so far, the idea transmitted in official documents seems entrenched in a defensive framework, nearest to deterrence.

Trying to balance civil and military roles, France has been developing two parallel structure systems one focused on cyber defense (under the management of the Defence Minister) and another focused-on cybersecurity (under the supervision of ANSSI). However, the line between these structures systems seems to be sometimes blurred when it comes to multilateral positioning, as "economic and industrial cybersecurity standards are favored to be discussed within the EU framework" (Brangetto, 2015, p.9). Notwithstanding, all French strategies touch upon international cooperation, especially within the EU and NATO structures. Besides, initiatives as the Paris Call (2018) and the Cyber Norm Initiative (2019) demonstrate a concern not only with sovereignty but with the establishment of norms in cyberspace.

Thus, France seems to seek for a Kantian anarchic logic within cyberspace, but (1) its self-perception as a great cyber power (Cabirol,2015) generating a hierarchical grading within cyberspace, (2) its attachment with sovereignty as an institution, and thus a self-interest in norms promotion and (3) its will of controlled use of violence, make it falls as an agent of a Lockean culture.

3.2 Germany

Germany started to change from a technical to a broader approach encompassing political-economic and societal aspects in 2011 when it launched its first National Cyber Strategy (Hathaway et al., 2016). This document listed ten strategic areas, being one of them the "Effective coordinated action to ensure cybersecurity in Europe and worldwide" Indeed the Strategy stated not only the necessity of a multilateral approach but also the establishment of "a code for state conduct in cyberspace (cyber code) […], which is signed by as many countries as possible and includes confidence-building security measures"(BMI, 2011, p.11). An idea that would persist in the development of a German "cyber identity."

Conventionally known as a "civilian- power" Germany also developed two system structures: one dealing with cybersecurity (civil one) having the Ministry of Interior in charge, and another focused on cyber defense under the leadership of the Defence Ministry. Moreover, since the Snowden doxing, a more proactive foreign policy toward cyber started to develop, and the 2016 Cyber Strategy declared the wiliness of Germany to actively shape cybersecurity policy within the European Union and internationally (Robin, 2018; BMI 2016).

However, since 2016 a shift in the official documentation occurred, and not only a process of centralization of policies began but also the military started to receive a more prominent focus. The 2016 White Paper on German Security Policy and the Future of the Bundeswehr (BMVg, 2016) emphasize that the country is ready to assume a leadership role and admits that cyberattacks can lead to conflict escalation. Moreover, the establishment of a Cyber Command and some government actors calling for the ability of Germany "hack back," lead the debate on offensive and defensive capabilities (Conrad; Werkhäuser, 2019).

Despite this aggressive tone, Germany still has a robust civilian position regarding cyberspace and sees itself as a reliable actor who "can also use its existing skills to support partner states and regions in the future through cyber-empowerment. These capabilities include developing cybersecurity strategies, legislation, institutions, certification, research, education and training, and regional initiatives" (BMI 2016, p. 42). These capabilities have a defensive design. Thus, any offensive moves could be seen as deterrence attempts. Moreover, the country's attachment to its structural economic power contributes to its more pragmatic decision making.

In sum, Germany seems to lead to a Kantian logic to cyberspace, as it searches for (1) norm respect, (2) multilateral, and (3) collective security engagements. As the 2016 cyber strategy declared, "a clear legal framework, confidence-building and greater resilience in Europe and the world also mean more protection for Germany" (BMI 2016, p.30).

3.3 The United Kingdom

The first Cyber Security Strategy of the UK was developed in 2009 after the threat of cyber-attacks mentioned in the National Security Strategy of 2008. Two years later, an update was made in the Cyber Security Strategy (UK Cabinet Office, 2011), and in 2016, the country launched its third Cyber Strategy, framed to encompass 2016-2021. Besides that, several additional documents were developed toward cyberspace, and the institutional structure of the UK has evolved into a centralized and hierarchic one.

The Cabinet Office is the pre-eminent institution for cybersecurity policy and strategy, where it hosts the Office of Cyber Security. Under its leadership is the Government Communication Headquarters, which hosts the National Cybersecurity Center and the Ministry of Defence leading the Cyber Security Operations Centre. This arrangement indicates a robust civilian approach to the cyber realm (Dewar, 2018). However, there are no distinct documents regarding cybersecurity and cyber defense in the UK's cyber document portfolio. The National Cyber Security is the primary policy document for both cybersecurity *and* cyberdefense. Thus, it makes sense that in the Cyber Strategy, there is a clear mention development of "offensive cyber capabilities" (UK Government, 2016, p. 41).

Although offensive cyber capabilities are mentioned, further explanations on their use falls under defensive approaches, within deterrence (UK Government, 2015) and the 2015 Cyber Strategy emphasizes that the cyber realm is seen more broadly, for its social and economic status (UK Government 2016, p.61) Indeed, the UK's approach in cybersecurity is to maintain its "world leader position" and to "safeguard the long-term future of a free, open, peaceful and secure cyberspace, driving economic growth and underpinning the UK's national security" (Ibid.,p.63) Moreover, it pursues a multilateral agenda not only within EU and NATO but also within the Five Eyes Alliance and BRICS, to consolidate its self-perception as an international actor. Thus, as Dewar (2018) explains, the cyber realm seems a microcosmos of the UK's national security posture.

Regarding these points is it is clear that the UK follows France in its tendencies toward a Lockean logic to cyberspace. It is concerned with (1) sovereignty issues (2) how international law (i.e., norms) applies to cyberspace to control the use of violence, as mentioned in the Attorney General's Office, Jeremy Wright's (2018) speech "Cyberspace is not – and must never be – a lawless world" and (3) it's national security mainly and its position in a hierarchical "leadership scale."

4. Final Remarks

The present paper confirms the hypothesis that cyberspace dynamics can be shaped with either a cooperative (friendship), competitive or enmity (conflict) logic, causing the international and digital system to have an interweaving along the lines of what Wendt (1999) considers to be the three logics of anarchy of the International System. The data shows that it is possible to think about cyberspace outside a Hobbesian logic. Indeed, the cases showed there are tendencies toward Lockean and Kantian logic within European countries, with a predominance of the first one, present on two of the three cases analyzed. Thus, the evidence is that there is room to manage cyber conflict escalation.

Moreover, the cases expose that states are developing different agenda-setting followed by their own identities and traditional interests, reflected in cyberspace. In other words, Germany identifies itself as a civilian and economic power. Thus, it seeks to maintain its status quo with a holistic approach toward cyberspace, while France and the United Kingdom, identifying themselves as "normal powers," are concerned with their national security and borders (i.e., sovereignty) in the digital realm. Besides, all three countries seek defensive designs (focus on deterrence), and their link to legality as the venue for conflict escalation control can be considered an indicator of the relevance of norms to cyberspace, especially as ways of socialization, mixing on and offline interstate interactions. However, the practice and the internalization of norms, and consequently, a dominant and unique culture among the possible cyber logics lies within states and their interactions with other relevant cyber actors.

5. Conclusion

The present paper sought to fill a gap between constructivism and states interest's generation toward cyberspace, by tackling the anarchic structure of cyberspace and its correlation to Wendt's (1999) anarchy cultures thesis. Asking what in states' perception could result in a more cooperative (Kantian), competitive (Lockean) or enmity (Hobbesian) action pattern, the paper showed the relation between physical and digital world interests and agenda-setting, elucidating that deterrence and norm-based approaches may be an emerging way of socializing interactions in cyberspace, prompting less aggressive logics for the cyber realm. The future of cyberspace dynamics, in last instance, will rely upon cyber actors' interactions, involving states and non-state actors.

References

Akbari, A; Gabdulhakov. R. (2019) "Platform Surveillance and Resistance in Iran and Russia: The Case of Telegram." Surveillance & Society vol 17 no.1/2, pp. 223-231.

Baezner, M (2018). "France" In Dewar, R.S *National Cybersecurity and Cyberdefense Policy Snapshots*, Cyber Defense Project (CDP), Center for Security Studies (CSS), ETH Zürich.

Brangetto, P. (2015). National Cyber Security Organisation: France, National Cyber Security Organisation. NATO Cooperative Cyber Defence Centre of Excellence, Tallinn.

Buchanan, B. (2017) *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*, Oxford University Press, Oxford.

Bundesministerium der Verteidigung(BMVG). White Paper: on German security policy and the future of the Bundeswehr. Berlin

Bundesministerium des Innern (BMI). Cyber-Sicherheitsstrategie für Deutschland 2016. Berlin.

Cabirol, M. (2015). "La lutte informatique offensive n'est pas un tabou" (Jean-Yves Le Drian) [online]. RP Def. URL http://rpdefense.over-blog.com/2015/09/la-lutte-informatique-offensive-n-est-pas-un-tabou-jean-yves-le-drian.html (Accessed 22.01.2020).

Calvety, M. D (2018a). "Europe's cyber-power" *European Politics, and Society,* Volume 19, Issue 3, pp. 304-320

Calvety, M.D.(2018b) Thomas Rid, Cyber War Will Not Take Place, ERIS, 1-2018, pp. 131-134. https://doi.org/10.3224/eris.v5i1.17

Carr, M. US Power and the Internet in International Relations: The Irony of the Information Age. Hampshire: Palgrave, 2016

Choucri, N.; Reardon, R (2012) The Role of Cyberspace in International Relations: A View of the Literature In 2012 ISA Annual Convention San Diego, California. April 1, 2012, https://nchoucri.mit.edu/sites/default/files/documents/%5BReardon%2C%20Choucri%5D%202012%20The%20Role%20of%20Cyberspace%20in%20International%20Relations.pdf (Accessed 14.02.2020)

Ciolan, I.M (2014) "Defining Cybersecurity as the Security Issue of the Twenty First Century. a Constructivist Approach," *The Public Administration and Social Policies Review* VI, vol. 1 no.12. pp.120-136

Cohen, M.S. (2018) *Beyond theory: applying empirical evidence to cyberspace theories* Ph.D. Thesis. Northeastern University, Boston.

Conrad, N; Werkhäuse (2019) Germany debates stepping up active cyberoperations. Deutsche Welle [online] RP Def. URL https://www.dw.com/en/germany-debates-stepping-up-active-cyberoperations/a-49359866 (Accessed 22.01.2020).

Dennison, S; Franke, U. E; Zerka, P. (2018) *The nightmare of the dark: the security fears that keep Europeans awake at night*. London: European Council of Foreign Relations (ECFR). Available at https://www.ecfr.eu/page/-/SECURITY_SCORECARD_267.pdf (Accessed 13.02.2020)

Derian, D., 2000. "Virtuous War/Virtual Theory." *Royal Institute of International Affairs* vol. 76 no.4, pp. 771–788.

Dewar, R. S (2018) The United Kingdom in Dewar, R.S *National Cybersecurity and Cyberdefense Policy Snapshots*, Cyber Defense Project (CDP), Center for Security Studies (CSS), ETH Zürich

Drezner, D. W. (2004). "The Global Governance of the Internet: Bringing the State Back" *Political Science Quarterly,* vol.119 no.3., pp. 477–498.

Eriksson, J., and Giacomello, G. 2006. "The information revolution, security, and international relations."*International Political Science Review (SAGE) Publications* vol 27 no.3, pp.221–244.

Grigsby, A. (2017). "The End of Cyber Norms," *Survival*, vol 59 no. 6, pp. 109-122.

Guzzini, S; Leander A. (2006) "Wendt's Constructivism" In: Guzzini, S.; Leander, A. (ed) *Constructivism and International Relations: Alexander Wendt and his Critics*, Routledge, London, pp.73-91

Hansen, L; Helen N. (2009) "Digital Disaster, Cyber Security, and the Copenhagen School," *International Studies Quarterly* vol. 53 no.4 pp.1155-1575.

ITU (International Telecommunication Union) (2019) Global Cybersecurity Index 2018. Geneva: ITU Publications. Available at https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf (Accessed 13.02.2020)

Kazmi, Z. (2012) Polite Anarchy in International Relations Theory, Palgrave Macmillan, New York.

Kello, L (2017). "Cybersecurity: Gridlock and Innovation" In Hale, T., Held, D. *Beyond Gridlock*, Polity Press, Cambridge pp.205-228

Laudrain, A.P.B (2019). France's New Offensive Cyber Doctrine Lawfare [online] RP Def. URL https://www.lawfareblog.com/frances-new-offensive-cyber-doctrine(Accessed 22.01.2020).

Mccarthy, D. R. (2015) *Power, Information Technology, and International Relations Theory: The Power and Politics of US Foreign Policy and Internet*. Palgrave, Hampshire

Ministère de la Défense, 2013. French White Paper Defence and National Security 2013.

Ministère des Armées, 2018. Éléments Publics de doctrine militaire de lutte informatique offensive.

Parmy, F., 2019. Stratégie cyber des Armées 18 Janvier 2019, Paris

Perkovich, A.G.; Levite, A. E *Understanding Cyber Conflict: 14 Analogies*, Georgetown University Press, Washington DC.

Prunckun, H (2018). Weaponization of Computers In Prunknun, H. (ed). Cyber weaponry: Issues and Implication of Digital Arms, Springer International Publishing pp.1-12

Riordan, S. (2019). *Cyberdiplomacy: managing security and governance online*. Polity Press, Cambridge.

Robin, P. (2018) Germany In Dewar, R.S *National Cybersecurity and Cyberdefense Policy Snapshots*, Cyber Defense Project (CDP), Center for Security Studies (CSS), ETH Zürich

Schulze, M. (2018) *From Cyber-Utopia to Cyber-War. Normative Change in Cyberspace*. Ph.D. Thesis. Friedrich-Schiller-Universität Jena. Jena

Senyuk, Y (2018) "The Cultures of Anarchy of the International system and their influence on European Integration Process," *European Political and Law Discourse* vol 5 no.1 pp.55-59

UK Cabinet Office, 2011. *The UK Cyber Security Strategy*: Protecting and promoting the UK in a digital world, London

United Kingdom Government (2016) UK Government, 2016. *National Cyber Security Strategy 2016-2021*, London

Valeriano, B., Jensen, B., Maness, R. (2018) *Cyber Strategy: the evolving character of power and coercion*, Oxford University Press, Oxford.

Wendt, A. (1992a). "Anarchy is what states make of it: the social construction of power politics," *International Organization* vol 46 no.2, pp.391–425.

Wendt, A. (1992b). "Levels of analysis vs. agents and structures: part III," *Review of International Studies* Vol 18, no.2, pp.181–5.

Wendt, A. (1999). *Social Theory of International Politics*, Cambridge: Cambridge University Press

Wendt, A. (2011) "The Social Construction of Power Politics" In Hughes, C W; Meng, L Y. (ed) *Security Studies: a reader,* Routledge, London pp.179-186

Wright, J. (2018). *Cyber and International Law in the 21st* Century 23 May 2018, London.

Cyber Deterrence and Russia's Active Cyber Defense

Maija Turunen[1] and Martti J Kari[2]
[1]Finnish National Defence University, Helsinki, Finland
[2]Jyväskylä University, Finland
maija.turunen@ftia.fi
martti.j.kari@jyu.fi

DOI: 10.34190/EWS.20.038

Abstract: The Russian Military Doctrine (2014), the National Security Strategy (2015), and the Information Security Doctrine (2016) consider the information space as a domain of warfare, where the war against Russian digital sovereignty is waged. Russia's cyber defense includes the protection of critical information infrastructure and increasing digital sovereignty by improving the readiness and capabilities to isolate the Russian segment of the Internet from the global Internet. Yet a credible defense also requires credible deterrence. The traditional deterrence options are deterrence by punishment, deterrence by denial and deterrence by entanglement. Deterrence by punishment is based on a defender's capability to retaliate. Deterrence by denial means that a defender is capable to limit damages by denial attacker´s success. Deterrence by entanglement is based on nations' mutual interest and their political, economic, and strategic interdependence. An effective deterrence strategy has to be credible, based on capability and send the right message to a possible adversary. Credibility means that the state displays the willingness to counteract an attack, and capability means that the state has the tools to do so. Cyber warfare differs from warfare in other domains because actors in cyber conflicts are not always military actors. State actors, criminals, and terrorists attack state authorities, media, and critical infrastructure. The attribution of an attack to a specific party is often difficult or impossible to do. An attacker in cyberspace can act with deniability, impunity, and anonymity. Traditional deterrence is difficult or impossible to implement in cyberspace because denial can be too difficult and there is no possiblity to identify an attacker for retaliation. The strategic option of cyber deterrence is an active cyber defense, which is a combination of deterrence by denial and deterrence bypunishment. This paper discuss Russian active cyber defense, which combines defensive and offensive cyber capabilties. In its theoretical context, this paper is based on the theory of deterrence and the theory of strategic culture. The theory of deterrence is used to describe how Russia constructs its cyber deterrence. The theory of strategic culture seeks to explain Russia's motives for these choices.

Keywords: Russia, cyber deterrence, strategic culture, active cyber defense

1. Introduction

Most industrial countries are preparing for cyber warfare, while some countries are already waging these wars. In the Western debate, cyber warfare is used more as an independent concept (McCulloh & Johnson 2013 pp. 5-14), while in Russia it is seen as part of information warfare (Hathaway & Klimburg 2012, 17-18; Nye 2017, 49). States are seeking to increase both their defensive and offensive cyber capabilities as well as to create deterrence of adversaries. To this end, states are using strategic communication to deter potential attackers, and different official documents, such as strategies and doctrines, are an important part of this communication. These documents disclose how the state publicly understands its position in cyberspace and communicates its goals and capabilities there. Creating a cyber deterrent is thus part of the effort to influence an adversary's war character.

The concept of deterrence is traditionally derived from the concept of nuclear deterrence, but it is not functional in the cyber world. According Manjikian (2016, 16) there are the knowledge problems or the problem of attribution; the temporal problem, or the ways in which time functions in cyberspace as opposed to during nuclear attacks; the payoff or reward structure for both types of events.

Cyber deterrence can be seen to consist of the capabilities of the actor (state/adversary), the willingness to use them, and the image that this state manages to project to its potential adversaries. Adversaries' cultural, political, and ideological background and their perception of the prevailing character of war affect how credible this cyber deterrent is. According to Goodman (2010, 105), deterrence has eight elements: an interest, a deterrent declaration, denial measures, penalty measures, credibility, reassurance, fear, and a cost–benefit calculation.

The use of cyber methods is suitable for the implementation of the basic principles of Russian strategic psychological thinking, like surprise and confusion. Similarly, with the use of these methods of subtle nature and the fact that the methods are left to speculation, these methods are well suited as a means of influencing the so-called grey area, in the initial period of war, where some Russian war theorists (Chekinov & Bogdanov, 2012) claim that wars in the future will be won . These grey areas can be considered as the new normal in both Russian and Western military strategic and deterrence thinking.

2. Theories in the background

In its theoretical context, this paper is based on the theory of deterrence. In international politics, deterrence theory is the idea that a state, owning destructive power, could deter a more powerful adversary with this destructive power. Generally, deterrence means preventing an adversary from taking undesired action. The general theory of deterrence is defined as the use of means of decisive influence over an adversary's decision-making. Traditional deterrence is based on an adversaries' perception that a threat of retaliation exists, or the planned attack or other action cannot be successful or the costs of the attack outweigh the benefits. (Jasper, 2018, 161) Further, Goodman (2010, 108) highlights that "in addition to strong denial measures, classical deterrence theory demands that penalty measures be certain, severe, and immediate; however, cyber deterrence emphasizes certainty more so than severity or immediacy." Three traditional deterrence options are deterrence by punishment, deterrence by denial and deterrence by entanglement. In addition, one can also talk about deterrence by association, deterrence by norms and taboos.

Deterrence by punishment or by retaliation is based on a defender's capability to first identify the attacker and then to retaliate against the adversary's hostile activity. Deterrence by punishment means, in the kinetic world, the capability to create a credible threat to an adversary, even with a deadly response that can cause unexpected losses. In cyberspace, it is more challenging to define an adversary's hostile activity, which gives the right to retaliate than it is in other domains. Identifying the cyber threat is challenging due to the attribution of the attacker, the infinite number of constantly changing actors, the irrelevance of geographical distance between the target and the attacker, and the likelihood of an attacker's plausible deniability (Mandel, 2017, 5). According to Kello (2014, 144), the problem with deterrence by punishment is also: "The difficulty of attaining credible attribution of the identity and location of an assailant degrades the crucial psychological base of the logic of penalties. The difficulty inheres in the opaqueness of cyberspace, but it is also concerns the traditional satecentric procedures of international relations."

Deterrence by denial is a defensive strategy that tries to convince an adversary that an attack is futile, because it would not achieve successful and cost-effective results. Deterrence by denial means that a defender is capable of limiting damages and withholding benefits of malicious activity by denial of attacker's success. The denial of an attacker's success is implemented by ensuring the security of the target of the attack. In cyberspace, this denial is realized by increasing the security of information systems, information and telecommunication networks, and automatic control systems (Jasper, 2017, 15). One of the ways to increase the security of information systems is to use so-called defense-in-depth, consisting of multiple layered, overlapping and mutually supportive defense systems, which includes detective and preventive functionalities. In practice, this means increased capability to detect a cyber-attack and respond to an attack as well as to reduce and mitigate risk (Pescatore & Sager, 2013). A well-planned and implemented cyber defense can detect and repulse the majority of cyber-attacks. According to Kello (2014, 143-144) the deterrence by denial involves the following types of problems as if the impossibility of reducing the full effects of cyberattack impedes denial by means of redundancy and resilience and these indirect effects are largely unknowable before attack.

Deterrence by entanglement is based on nations' mutual interest and political, economic, commercial, and strategic interdependence in cyberspace as well as some degree of vulnerability (Jasper, 2018, 198, 234, 269). Mutual interest means a common reliance on the Internet that unites otherwise non-friendly, competing or even hostile states and actors. These actors refrain from attacking each other because they gain, for example, financially due to the Internet connectivity (Ryan, 2017). Mutual interest is at the core of the concept of deterrence by entanglement and it expands to include other methods of deterrence, such as encouraging responsible state behavior through norms and taboos. Nye (2017, 59) has stated, that: "Entanglement is sometimes called "self-deterrence" and treated as a case of misperception. The term "self-deterrence," however, should not lead one to dismiss the importance of entanglement, whether in a bilateral or a general sense. The perceptions that costs will exceed benefits may be accurate, and self-restraint may result from rational calculations of interest.”

According to Kari (2019, 25), the theory of strategic culture is a suitable theory for exploring and explaining the Russian vision of cyber conflicts, its enemies, cyber threats and preferences for responding to cyber threats. Johnston defines strategic culture as a set of historical patterns of how state leadership thinks about the use of force to achieve political goals. According to him, strategic culture consists of a central paradigm and a set of strategic preferences. The central paradigm describes the nature of the conflict and the perception of the enemy and threat as well as how to respond to that threat. Strategic preferences are assumptions about what options are the most effective against a particular threat (Johnston, 1995a). According to Johnston (1995b), historical factors have a predominant influence on the formulation and outcome of a state’s strategic culture. These historical factors are influenced by the political, cultural, and cognitive characteristics of the state. Technology, threat level, and organizational structures are of secondary importance. This paper applies Johnston’s definition of strategic culture and his methodological framework. The central paradigm of Russian strategic culture corresponds to the Russian threat perception and strategic preferences correspond to Russia’s deterrence by denial in cyberspace.

3. An active and effective cyber defence and deterrence

Denning and Strawser (2017, 194-195) defines an active and passive cyber defence as follows: **“**Active cyber defense is a direct defensive action taken to destroy, nullify, or reduce the effectiveness of cyber threats against friendly forces and assets, and passive cyber defense is all measures, other than active cyber defense, taken to minimize the effectiveness of cyber threats against friendly forces and assets. Put another way, active defenses are direct actions taken against specific threats, while passive defenses focus more on protecting cyber assets from a variety of possible threats.” Denning (2013, 3) has argued, that: ”Active cyber defenses can be characterized by four features: scope of effects, degree of cooperation, types of effects, and degree of automation. Together, they place active cyber defenses in a four-dimensional space and provide a framework for distinguishing different types of active cyber defenses and analyzing the ethical issues they raise.”

The strategy of active cyber defence is based upon the real-time detection and analysis of network security breaches seeks to create a comprehensive, automated and thereby unavoidable response to neutralize and reinforce the effects automatically through associated auto trigger of legal simultaneous countermeasures inside and beyond network and state territorial boundaries. The remedies for this are cyber threat intelligence, real-time detection of intrusion and analysis of network security breaches. The offensive part of active cyber defense includes the capability to give early warning of a possible cyber-attack, that is, to reveal and identify the cyber threat and respond to the incoming cyber-attack in the early stages of the cyber kill chain to withstand the attack. The exact attribution of the attacker is not always vital for the response. Rather, it is more important to stop the attack before damage occurs (Jasper, 2017, 207- 209; Mandel, 2017, 35).

Effective deterrence requires capability, credibility, and communication. Deterrence operates in a grey area. Escalation into war means that deterrence has failed. Strategic communication can compensate for the shortcomings of capability, but not the loss of credibility. Because deterrence rests on perceptions, its effectiveness depends on credibility. An effective deterrence strategy should be credible from the perspective of the actor that is to be deterred. Effective deterrence should be based on capability and it should transmit the correct and intended message to the desired audience. Credibility means that the state shows the willingness to employ proposed actions to counteract the attack while capability means that the state has the capability and tools to do so (JP 3-0, 2017). That means offensive capabilities, which may be possessed by the state itself or its allies. Passive defense alone will not work.

4. Russian version of an active cyber deterrence

According to the Russian view, the number and severity of threats to Russia have increased in cyberspace, and those threats are shifting to Russia’s internal sphere (PP-2796, 2014). Russia’s national interests can be threatened in or through cyberspace internally as well. Terrorists and extremists may direct cyberspace attacks at strategic targets to disrupt the management and decision-making system and to paralyze Russia’s strategic leadership. In addition, cybercriminals may threaten Russia’s national interests in or through cyberspace by penetrating the state information systems (see RBA, 2013; SBRF, 2013; MDRF, 2014). The Russian view about deterrence and its functions in warfare is not a mainstream view in West. For example the development of a multistakeholder approach in which the state's interests are not dominant, does not happen. From the Russian point of view, it is always a question of Russian government`s interest and national security. These values take precedence over everything else, and it is up to the stakeholders to implement them in a manner defined by the authorities.

Russia has tried to improve cyber deterrence by punishment, deterrence by denial, and by the combination of deterrence by association, deterrence by norms and taboos, and deterrence by entanglement. Russia is building an active cyber deterrent as part of its other deterrent systems by all available means. Indeed, Bērziņš (2019, 166–167) considers that the Russians have placed influence at the center of their operational planning. Examples of such influence include skillful internal communication, masking operations, psychological operations, and well-constructed external communication. Strategically, therefore, the operational nature of the new war cannot be regarded as a military campaign in the classic sense of the term, but as an opportunistic combination of different strategies. Bērziņš also emphasizes that the key to understanding Russia's strategy is to understand that it is eclectic and relies on whatever works in a given situation.

One important part of communication about effective deterrence are Russian military strategies and doctrines. In particular, they emphasize the various threats and recognize and oblige them to respond strongly to these threats in all domains. They also contain clear cyberspace management measures. The Russian Military Doctrine not only has the function to guide the planning, operation and legislation of the authorities, but it also serves as a tool for strategic communication. The Military Doctrine creates strategic deterrents against those who threaten the state, while also communicating to Russian citizens about threats to the state and justifying the means of defending against such threats. As in his speech in 2017, General Gerasimov (Komsomolskaya Pravda, 2017), the General Staff Commander of the Russian Army, seems to focus on the fight in the minds of the citizens, the sixth dimension of battlefields. Gerasimov stated that victory is not only achieved through material resources, but also through the nation's intellectual resources, its cohesion, and the unification of all forces to resist aggression.

The doctrine lists the neutralization of potential military threats by political, diplomatic, and other non-military means. However, the doctrine, in paragraph 21s), includes an important additional measure, even with offensive features: "creating conditions that reduce the risk of using information and communication technology to achieve military policy goals." This may be seen as referring primarily to political diplomatic means to influence, for example, international law, but also to empowering the armed forces to take action to eliminate targets that increase such risks.

In the Russian philosophy of warfare, concealment and deception play an important role. These elements are also central to cyber operations, along with the control of intangible agents and service providers by Russian security service authorities. The coordinated role of such service providers in close proximity to state actors is an essential and likely growing component of both Russian cyber capability and other operational warfare, while maintaining the traditional leading role and mission of traditional armed forces. Using service providers is not always an active operational choice, but allowing them to serve their own strategic and tactical goals in Russia may obscure the adversary's situation awareness and create uncertainty, which may open up new opportunities for state actors or contribute to action by state actors to deceive and conceal their own operations. One of Russia's most important tasks is to assess and forecast the development of the global and regional conflicts and military policy using modern technical tools and information technology in order to curb and prevent military conflicts, which can be considered a form of intelligence. In his speech at the Russian Academy of Sciences in March 2019, Gerasimov (Krasnaya Zvezda, 2019) again stressed the importance of considering modern warfare as consisting of military and non-military means of war, and the importance of achieving surprise. He also highlighted the importance of preventive measures, the identification of vulnerabilities, the importance of creating deterrence, and maintaining the ability to take strategic initiative.

According to Bērziņš (2014, 3), Russia's military strategy is based on three interrelated doctrinal levels: unilateralism, strong legitimacy, and the systematic denial of the use of military force. In this context, scholarly unilateralism refers to the idea that legitimacy could be derived from the successful use of force, while respect for legality refers to Moscow's attempts to base its actions on some "legal" grounds. The ban on the use of open military force can be better understood in the light of Russian diplomatic rhetoric in Crimea. Instead, in evaluating recent strategic communications by the Russian military leadership, the ban on open military force has somewhat receded. Given the events in the Kerch Strait, the credibility of communications has been strengthened through the use of legitimately justified military force.

The Russian Information Security Doctrine (2016, 3-6) states that the use of cyber methods and the need to protect against them are clear. The doctrine defines Russia's national interests in the information space to include maintaining continuous and smooth information operations on the information infrastructure, especially in the case of Russia's critical information infrastructure and the Russian integrated telecommunications network in the case of a direct aggressive threat and during the war. The doctrine also emphasizes geopolitical interests, the information infrastructure for military purposes, and the intelligence and psychological tools used by intelligence services to undermine internal political and social situations in different regions of the world. The Russian security services are obliged by all means to act in accordance with these goals also outside state borders.

The Russian military strategy is supposed to include the goal of creating an alternative reality to the adversary's operating environment, as well as the anti-access / area-denial zones for both its own and international information space. Russia is pursuing the leadership of its forces in a common information space, but is also preparing for threats to its information space by both defensive and offensive methods, while reinforcing its own national narrative with the legitimacy of its actions and the western belief in the actualization of created threats. Kari (2019, 17) has summarized this approach to cyber threats from the Russian doctrines as follows: "Russia's response to cyber threat means the implementation of mutually connected measures to predict, detect, suppress, prevent, and respond to cyber threats and mitigate their impact." The technical management of the cyber environment plays a major role here. Russia's technical cyber defense includes protection of critical information infrastructure and increasing digital sovereignty by improving the readiness and capabilities to isolate the Russian segment of the Internet from the global Internet. Russia's cyber deterrence is based on active cyber defense that combines cyber defense and offensive operations. Russia's offensive cyber operations include distributed denial of service (DDoS) attacks and advanced persistent threats (APT). Russian cyber deterrence is seen as credible, because the country has capability and it has sent the right message to adversaries, for example in 2007 to Estonia and in 2008 to Georgia (Kari, 2019, 89-92).

Yet only after 2015 has legislative action to promote information security progressed relatively quickly. The most important legal reforms for digital sovereignty and cyberspace management focused on the security of critical information infrastructure, the ban on network users' anonymity, the operators' obligations to store message and call content (the so-called Yarovaja package). These legislative packages empowered, among other things, the Russian Federal Technical and Export Control Authority to issue precise regulations on hardware, software, standards, and operating processes, as well as on operators considered to be subject to critical infrastructure regulations. In addition, unrestricted access to messages for Russian security services and other security authorities was authorized, along with access to encrypted communications that telecommunications operators and other communications service providers are required to record.

In April 2019, the Russian House of Commons also approved a bill on the RUNET that obliges Russian telecom operators to ensure that RUNET is available without connection to the rest of the Internet and that all communications are routed through nodes approved or controlled by Roskomnazor. The RUNET legislation aims to ensure national information security in Russia, manage network operators and acceptable content, and reduce the traffic and security risks related to transboundary traffic (Kukkola & Ristolainen, 2019, 74-78).

Permanent war against Russian digital sovereignty is waged every day (Sinovets, 2016). The USA has been conducting cyberspace exploitation in Russian power grid systems since 2012 and prepared cyber-attacks by installing malware in the Russian information infrastructure. One threat to Russia in cyberspace is the technological gap between the Russian Federation and the leading foreign states in information and communication technology (ICT). This gap limits Russia's capabilities to respond to cyber threats (UP-646, 2016) – to resort, in other words, to deterrence by punishment.

Russian cyber deterrence by denial can be explained by the factors and elements of Russian strategic culture. The factors influencing Russian strategic culture, such as a sense of vulnerability, fear of surprise attack and invasion, the narrative of Russia as a besieged fortress and the concept of permanent war, influence Russian strategic culture in cyberspace as well. One of the basic assumptions of Russian strategic culture is that the world, including cyberspace, is a dangerous and volatile battlefield, where Russia has to fight every day for its digital sovereignty (Sinovets, 2016). The National Security Strategy of the Russian Federation (UP-683, 2015) states that the use of force in international politics is increasing. Historical experiences have created a sense of vulnerability and the fear of invasion within Russian strategic culture (Cimbala, 2013). Geography, too, has an influence. The country's lack of natural borders has created a sense of vulnerability and the need for a buffer zone, political and military control of neighboring spaces, and territorial expansion to natural, easily defensible borders (Ermarth, 2006). Russia's technological inferiority and its backwardness in the development of high technology also has an influence on the country's strategic culture. This means that Russia must be prepared for a surprise attack by a technologically superior enemy. The best option is deterrence by denial.

5. Conclusions

Russia has created an active cyber deterrent through both psychological and technical methods. In its strategic documents Russia has created a comprehensive picture of cyber threats and the need to protect against them. At the same time, the Russian military leadership has been actively communicating the importance of asymmetric methods of warfare and the need to increase the state's capabilities in this area.

As regards the emphasis on the development and use of cyber capabilities, Russia has moved in a more operational direction in its strategic communications, which may indicate an improvement in cyber capabilities and require the armed forces to use them more widely and for a variety of purposes. The Russian Military Doctrine creates cyber deterrence by demonstrating a readiness to defend against cyber operations, using conventional force when needed, though the use of nuclear weapons is reserved as a defense mechanism only against conventional attacks.

Technically dominating a well-controlled information space such as the RUNET opens up many opportunities for Russia to create an active cyber deterrent. First, the RUNET provides a (theoretically) defined and controllable space to be attacked and defended. Second, the lockable cyber domain allows for a variety of offensive capabilities to be developed and tested more secretly than they would be in the global information space. Thirdly, the RUNET can also attract states or other actors close to Russia to use it and thereby strengthen the creation of a common entanglement deterrence. For Russia, the most difficult question in responding to cyber threats is that the country is lagging behind the leading foreign countries in the development of competitive information technology, including supercomputers, and this gap strengthens the Russian perception of its strategic vulnerability in cyberspace.

Russia's psychological and technical measures to create cyber deterrence clearly illustrate the impact of Russia's strategic culture and the accompanying need to reinforce the narrative of Russia as a besieged and threatened fortress. They also reflect Russia's approach to adopting various deterrent strategies and applying them flexibly as the case requires.

However, the existence of cyber deterrence and the desire to use cyber abilities are still largely a matter of faith for adversaries. States know that they and their critical functions are being attacked, but uncertainty about the attackers and their motives, as well as the Western interpretation of international law, act as a deterrent to responding to the attack other than by stable defensive means.

We have previously discussed how Russia creates cyber-deterrence. Russia's way of creating a cyber-deterrent is not commensurate with the West (NATO) because NATO is made up of different countries with different legal, political, geopolitical and religious views. Russia has only one view at a time.

References:

Bērziņš J. (2014) Russia's New Generation Warfare in Ukraine. From Implications for Latvian Defence Forces. National Defence Academy of Latvia Center for Security and Strategic Research, 2014, p.3-4. (https://sldinfo.com/wp-content/uploads/2014/05/New-Generation-Warfare.pdf)

Bērziņš J. (2019) Not 'Hybrid' but New Generation Warfare. From Russia`s military strategy and doctrine. Howard G. E. and Czekaj, M. (Eds.) (https://jamestown.org/wp-content/uploads/2019/02/Russias-Military-Strategy-and-Doctrine-web.pdf?x29008&x87069), pp. 157-184

Chekinov, S. G. and Bogdanov, S. A (2012) "Initial Periods of War and Their Impact on a country's Preparations for a Future War," Voennaya Mysl' (Military Thought), No. 11 2012, pp. 14-27

Cimbala, S. (2013) Russian Threat Perceptions and Security Policies: Soviet Shadows and Contemporary Challenges. The Journal of Power Institutions in Post-Soviet Societies. 14/15 (https://journals.openedition.org/pipss/4000)

Denning, D. E. (2013) Framework and Principles for Active Cyber Defense. http://hdl.handle.net/10945/59868

Denning D.E. and Strawser B.J. (2017) Active Cyber Defense: Applying Air Defense to the Cyber Domain. From Understanding Cyber Conflict: Fourteen Analogies Perkovich G. and Levite A.E. (Eds) Published by Georgetown University Press (https://carnegieendowment.org/files/GUP_Perkovich_Levite_UnderstandingCyberConflict_Ch12.pdf)

Ermarth, F. (2006) Russian Strategic Culture: Past, Present, and… in Transition? Defense Threat Reduction Agency Advanced Systems and Concepts Office. (https://fas.org/irp/agency/dod/dtra/russia.pdf)

Goodman, W. (2010) Cyber Deterrence Tougher in Theory than in Practice? Strategic Studies Quarterly -Fall (https://apps.dtic.mil/dtic/tr/fulltext/u2/a528033.pdf)

Hathaway M. E. and Klimburg A. (2012) Preliminary considerations: On national cyber security. From National cyber security framework manual. Edited by Klimburg, A. NATO Cooperative Cyber Defence Centre of Excellence. (https://ccdcoe.org/uploads/2018/10/NCSFM_0.pdf) pp.1-43Jasper, S. (2018) U.S. Strategic Cyber Deterrence Options. (http://centaur.reading.ac.uk/79976/1/22839264_Jasper_thesis.pdf)

Jasper, S. (2017) Strategic cyber deterrence. The active cyber defence option. Rowman & Littlefield Publishers. London.

JP 3-0 (2017) Joint Publication 3-0 Joint Operations 17 January 2017 Incorporating Change 1 22 October 2018 (https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_0ch1.pdf?ver=2018-11-27-160457-910)

Kari, M. J. (2019) Russian Strategic Culture in Cyberspace Theory of Strategic Culture – a tool to Explain Russia´s Cyber Threat Perception and Response to Cyber Threats. JYU DISSERTATIONS 122

Kello L. (2014) The Virtual Weapon: Dilemmas and Future Scenarios. in Politique étrangère 2014/4 (Winter Issue) Pages 139 – 150 (https://www.cairn-int.info/article.php?ID_ARTICLE=E_PE_144_0139)

Krasnaya Zvezda (4.3.2019) Векторы развития военной стратегии. (http://redstar.ru/vektory-razvitiya-voennoj-strategii/)

Komsomolskaya Pravda (26.12.2017) Начальник Генштаба Вооруженных сил России генерал армии Валерий Герасимов: «Мы переломили хребет ударным силам терроризма» (http://archive.redstar.ru/index.php/component/k2/item/35551-my-perelomili-khrebet-udarnym-silam-terrorizma)

Kukkola J. and Ristolainen M. (2019) Projected Territoriality: A Case Study of the Infrastructure of Russian Digital Borders. In: GAME PLAYER. Facing the structural transformation of cyberspace. Kukkola & all (Eds.). Puolustusvoimien tutkimuslaitos. Julkaisuja 11. Riihimäki 2019, pp.65-89

Manjikian, M. (2016) Deterring cybertrespass and securing cyberspace: Lessons from United States from border control strategies. (https://publications.armywarcollege.edu/pubs/2401.pdf)

Mandel, R. (2017) Optimizing Cyberdeterrence: A Comprehensive Strategy for Preventing Foreign Cyberattacks. Georgetown University Press.

McCulloh T. and Johnson R. (2013): Hybrid Warfare. https://jsou.socom.mil

MDRF. (2014) Военная доктрина Российской Федерации. [Military doctrine of the Russian Federation.] (http://www.scrf.gov.ru/documents/18/129.html)

Nye J. S. Jr. (2017): Deterrence and Dissuasion in Cyberspace. International Security, Vol. 41, No. 3 (Winter 2016/17), pp. 44–71 (www.mitpressjournals.org › pdf › ISEC_a_00266)

Pescatore, J. and Sager, T. (2013) Critical Security Controls Survey: Moving From Awareness to Action June 2013. SANS Whitepaper. (https://slidelegend.com/queue/sans-2013-critical-security-controls-survey-moving-sans-institute_59d017541723dd3afea35be0.html)

RBA. (2013) Agreement between Belarus and the Russian Federation on cooperation in the field of international information security. (http://www.pravo.by/main.aspx?guid=3871&p0=A01300055&p1=1)

Ryan, N J (2018) Five Kinds of Cyber Deterrence. Philosophy & Technology. September 2018, Volume 31, Issue 3, pp 331–338. (https://link.springer.com/article/10.1007/s13347-016-0251-1#Sec1)

SBRF. (2013) Основы государственной политики Российской Федерации в области международной информационной безопасности на период до 2020 года. [Basics of Russian Federation national policy on international information security to 2020.] (http://base.consultant.ru/cons/cgi/online.cgi?req=doc&base=LAW&n=1 78634&fld=134&dst=1000000001,0&rnd=0.5310172209117789)

Sinovets, P. (2016) 'From Stalin to Putin: Russian Strategic Culture in the XXI Century, Its Continuity, and Change', Philosophy Study Vol. 6, No. 7 (July 2016), 417-423 doi: 10.17265/2159-5313/2016.07.002

UP-683. (2015) Russian National Security Strategy. (http://www.ieee.es/Galerias/fichero/OtrasPublicaciones/Internacional/2016/Russian-National-Security-Strategy-31Dec2015.pdf)

The Doctrine of Information Security of the Russian Federation (2016) (http://www.mid.ru/en/foreign_policy/official_documents/asset_publisher/CptICkB6BZ29/content/id/2563163)

The Roots of Restraint in Cyber Conflict

Federico Variola
SIRPA, Fudan University, Shanghai, China.
Federico.variola92@gmail.com

DOI: 10.34190/EWS.20.037

Abstract: This research intends, first, to illustrate how conventionally powerful states are better equipped to exploit the unconventional nature of cyber conflict. Then, from the analysis of cyber threat reports and *The Dyadic Cyber Incident and Dispute Dataset*, the author will set a form game of sequential decisions under complete information to simulate cyberattack scenarios. Once established that conventionally powerful states possess significant advantages in the realm of cyberspace the author aims to explain, through a novel analysis of the role of attributional difficulties, why restraint rather than escalation has, for the most part, characterized states' interaction in cyberspace. In sum, this research advances several propositions: it challenges the notion that cyber weapons function best as asymmetric weaponry, it illustrates part of the reason why those same states that hold comparative advantages in cyberspace do not resort to increasingly severe cyberattacks to achieve political goals, and finally, it advances that attributional difficulties do not necessarily lead to the escalation of conflicts, neither in pace or intensity. Moreover, through a game theoretical approach the author aims to provide original interpretations of existing concepts: first, that the same features that make cyber weapons convenient instruments progressively lessen when offensive strategies are reiterated. Second, that attacks that are tactically successful in terms of immediate outcome do not necessarily reflect a beneficial grand strategy. As states' interactions in cyberspace become more common, causing the first-mover's advantages to reduce, sustained restraint interrupted by sporadic attacks of medium-to-high intensity is likely to remain the norm.

Keywords: Cyber Conflict, Game Theory, Restraint, Cyberwarfare, Offense-Defense, Cyberattack

1. Introduction

Despite bleak forecasts, severe cyberattacks between states occur rarely, as seen in *The Dyadic* Cyber Incident and Dispute Dataset (Maness, Valeriano, and Jensen, 2019). Slayton (2017) finds that the way states employ complex offensive attack strategies in cyberspace seems to be, at times, in contrast with analyses of costs and benefits in purely economic terms. Theoretical frameworks such as Offense/Defense Theory can therefore explain the high frequency of attacks of low intensity, but struggle to justify the manner severe attacks still occur between states.

In order to circumvent the theory's shortcomings, this research has developed a scale for the objectives of cyberattacks based not only on their economic value but also on their significance to the security of a state. This scale has then been implemented into a game theoretical model that simulates scenarios of cyberattack. Through said approach, the author was able to test cyberattacks of the desired intensity, and analyze in detail how agents would behave following cyber incidents.

Once shown that conventionally powerful states possess comparative advantages in cyberspace, this analysis will also demonstrate that, in regard to severe attacks, restraint tends to emerge as a consequence of the way states exploit one of the defining characteristics of cyber conflict: attributional difficulties. Although attributional difficulties allow the proliferation of attacks of low intensity, the opposite occurs for severe attacks. Offensive agents are incentivized to utilize attack strategies of higher intensity scarcely, in order to exploit them at their best. The results of the model have then been paralleled with several instances of cyberattacks of varying intensity, to corroborate the overarching hypotheses.

1.1 Definition of terms

Cyberattacks are still problematic to define. The author adopts the intersection of two noteworthy definitions: Hathaway et al.'s (2012) and Roscini's (2014). Hathaway et al. (2012) propose an effect-based definition that emphasizes the threats posed by offensive cyber operations through a combination of the operations' hard target and their goal:

> *"A cyber-attack consists of any action taken to undermine the functions of a computer network for a political or national security purpose."*

Hathaway et. al (2012) include kinetic attacks against computer networks, and further categorize network attacks as a subset of cyberattacks. Instead, Roscini (2014) argues that cyber operations present three types of effects: primary effects on the targeted system or network, secondary effects on the infrastructure operated by the targeted system, and tertiary effects on the persons or property affected by the destruction or incapacitation of the targeted system. Therefore, Roscini (2014) argues that physical damage can never be a primary effect of a cyberattack. An intersection of Hathaway et al.'s definition and Roscini's classification seems to best represent how states sense cyberattacks: substitutive or auxiliary strategies to kinetic strikes but different in nature.

1.2 Literature review

In regard to the current application of International Law to cyberspace, the natural starting point is the Tallinn Manual in its two editions. Assembled by CCDCOE, the Manual represents a noteworthy attempt by a group of experts to codify states' behaviors in cyberspace. Other notable works are: Roscini (2010), who sets forth that cyber force can be considered use of force under Article 2 Para 4 of the UN Charter, and describe the view of several commentators to affirm that cyberattacks can be considered armed attacks when producing effects comparable to a kinetic attack; Schmitt (2011), who specifies seven conditions that, when met by a cyberattack, will compel the victim state to consider the strike as an armed attack, but also recognizes the limits of present legal frameworks; Locatelli (2015), who offers a comprehensive analysis of legitimate self-defense practices against cyberattacks under International Law. This research also acknowledges the possibility that grey areas in current frameworks are not only tolerated but actively preserved by states.

A great deal of attention has been devoted, in the nascent analysis of cyber conflict, to how the relative costs of offense and defense allow one to forecast the behavior of states in cyberspace, a concept which closely relates to the body of theories of Offense-Defense, from the field of International Relations. In cyber conflict, offense is often said to be cheaper than defense, as stated by Libicki (2009). This has led some, such as Locatelli (2013), to estimate that escalation and proliferation of attacks will occur as long as the current Offense-Defense Balance in cyberspace persists. Similarly, other estimations such as Liff's (2012) see cyber weapons as the ultimate asymmetric armament.

On the other hand, others, such as Lindsay (2015) and Slayton (2017), believe engineering successful offense can be costlier than the damage an offensive cyber strategy can inflict, a position shared in this research. Significant doubts remain also on the validity of Offense-Defense theory's and Offense-Defense Balance's estimations. For instance, Huntington (1987) and Tang (2010) question whether offensive weapons, rather than posture, can be discerned.

Within the field of Computer Science, Game Theory has shown promise in modeling cyberattacks: Jormakka and Mölsä (2005), Attiah, Chatterjee, and Zou (2017), etc. Although fascinating and fruitful, these researches can only marginally contribute to the study of International Relations, their chief limit being that said studies do not aim to characterize the agents beyond the "red team/blue team" types. This research will instead specifically address how the characteristics and conventional capabilities of agents affect their interactions.

2. Theory

A theoretic framework that has been considered suitable to understand the role of military technology in state behavior is Jervis' (1978) Offense-Defense Theory. Jervis (1978) identifies the security dilemma as a process leading states to increase the insecurity of other states through defensive actions that should otherwise be perceived as non-threatening. The intensity of this dilemma depends, following Jervis' (1978) reasoning, on two variables: the discernibility of offensive and defensive military technology and whether the defense, or the offense, hold comparative advantages. Even if states' intentions were transparent, this second variable, the Offense-Defense Balance, can still exacerbate the security dilemma, as states will act offensively when offensive weapons present advantages.

Glaser and Kaufmann (1998) advance a formula for measuring the Offense-Defense Balance: a cost ratio of the forces needed to capture a territory versus the cost to hold said territory, finding evidence that satisfies the Offense-Defense model. These findings are also supported by Adams' (2004) analysis of armed conflict among Powers from 1800 to 1997. Another noteworthy contribution is van Evera's (1998), who describes the mechanisms which may explain aggression in an offense-dominant world. Nevertheless, the "technological" aspect of Offense-Defense theory has been criticized on several grounds. Huntington (1987) and Tang (2010) propose that the discernibility of weapons is simply not useful, and oftentimes outright impossible.

Tang (2010), and Lieber (2000), offer a supplementary categorization of the balance: objective and subjective. While the objective balance has an impact on the outcome of a single battle or war, the subjective balance weighs over the cause of war systematically. Regarding the objective balance, disagreements persist concerning the importance of pure technology. Lynn-Jones (1995) and Lieber (2000) favor the technology-only objective balance but others, such as Tang (2010), advocate that many of the empirical cases brought by supporters of Offense-Defense Theory fail to account for the impact of first-mover's advantage and innovation. Glaser and Kaufmann's (1998) interpretation of the Offense-Defense Balance offers a partial solution to these issues, but increases the complexity of measurement, while also making a dyadic variable of the objective balance. Although less suitable to understand the eruption of war, a limited dyadic approach can help forecast the outcome of specific conflicts, which is the objective of this research.

In the respect of Offense-Defense Balance and cyber weapons, many scholars argue that cyberspace favors the offense: Nye (2010) and Lieberthal and Singer (2012) argue that offense has the advantage in cyberspace because cyberspace was designed to share information. Therefore, it could be reasoned that cyberattacks are not particularly costly and have high payoffs, whereas defensive tactics are expensive and often ineffective, as defensive agents have to cover all their weaknesses.

A minority of scholars argues instead that an approach that focuses on perceived utility better explains the behavior of states in cyberspace. Slayton (2017) claims that technology, organizations, and skill are not separable variables, and that an approach that models the balance in terms of costs only may miss the mark regarding the calculations of utility maximization. Following this technology-plus reasoning, Slayton (2017) finds that attackers enjoy advantages only as long as their goals remain limited.

To embed Slayton's (2017) contribution, this research has developed a parameter (π) for strategic gains and losses that extends beyond economic costs: this step will be instrumental in correctly modeling cyber weapons not as simply "cheap" weapons, but rather as complex instruments for States to pursue objectives whilst safely operating below the threshold of retaliation.

2.1 Hypotheses

H1: Conventionally powerful states equipped with cyber-capabilities can exploit the benefits of offensive cyber-strategies at their best.

H2: When reiterated, offensive cyber-strategies offer diminishing returns to any offensive agent.

2.2 Methodology

Before diving into the description of the Game Theoretical Model a few caveats are in order. The aim of this research is not to model the sole cause of states' restraint in cyberspace, but rather to elucidate a process that, safe the eradication of attributional difficulties in cyberspace, will constantly impact states' actions in this domain. Therefore, the author will preface that concomitant explanations for states' restraint in cyberspace exist. A second caveat concerns the qualitative choices that have been necessary when structuring the model: not all parameters can be measured with the same degree of accuracy. Overall, given the goal of this research to explain restraint in cyberspace, the author has chosen to give qualitative edges to the offense, so that these would not skew the model's results towards the hypotheses.

2.3 The basic model

There exist two agents: both are nation states, with the offensive actor (Player 1) in possession of capabilities and expertise that allow him to successfully, and repeatedly, target the defensive agent (Player 2) at will. The author has chosen to model cyberattacks as certain for two reasons: first, generally, states victim of failed cyberattacks do not respond to them via any of the strategies pertinent to this model. Second, as the author wishes to compare the differences between consecutive iterations of a specific interaction, the model can be vastly clarified, while maintaining internal coherence, by removing offensive uncertainty on all iterations.

2.3.1 Data

τ, conventional capabilities:

Concerning the parameter τ, this research adopts the National Power Rankings of countries 2019, (Białoskórski, Kiczma, and Sułek, 2019) a ranking structured through the study of powermetrics. This methodology lists three types of state power. The economic (general) power (EP. Eq. 1) consists of economic outcome (GDP, gross domestic product), demographic factors (L, population) and spatial factors (a, territory area). The military power (MP, Eq. 2) consists of military and economic factors (MEX, military expenditures), demographic and military factors (S, number of active-duty soldiers) and spatial factors (a). All exponents are in accordance with the Sułek Model (2013). Conventional capabilities are represented by the geopolitical power (GP, Eq. 3), which is calculated as the arithmetic mean of economic (general) power and double military power (to indicate the significance of the military factor in shaping the current distribution of power).

$EP = GDP^{0.652} \times L^{0.217} \times a^{0.109}$

Equation 1

$MP = MEX^{0.652} \times S^{0.217} \times a^{0.109}$

Equation 2

$GP = \frac{EP + (2 \times MP)}{3} = \tau$

Equation 3

This approach further specifies three different types of militarization: among these, the demographic militarization index (md) will be instrumental when assessing a country's potential for mobilization and its credibility.

$md = S^{0.217} / L^{0.217}$

Equation 4

π_n, strategic gain:

Regarding the parameter π, the author has developed a multilayer scale based on The Dyadic Cyber Incident and Dispute Dataset and tracing the HIIK conflict barometer, shown in Tables 1, 2 and 3, that combines three dimensions (as shown in Eq. 5): target type (private entity, 0.5; public entity, 1; critical national infrastructure, 1.5), damage type (data breach, 0.5; reversible damage, 1; irreversible physical damage, 1.5; and irreversible physical damage leading to loss of life (P+L), 2), and intensity (contained: single or multiple targets, main operations not significantly affected, ×0.5; extended: single or multiple targets, main operations significantly affected, ×1; systematic: multiple targets, all operations interrupted, ×2). The Intensity multipliers' weights have been assigned so that the value of this research's core variable π does not differ significantly from the Severity Scale of The Dyadic Cyber Incident and Dispute Dataset's codebook, while also accounting for the absence of cyber-espionage in this research's approach. This research does not intend to assess the payoffs of single cyberattacks in terms of underlying strategy, but in terms of the incursions' immediate effectiveness, as a cyberattack can achieve all of its anticipated goals and still harm the handler's interests in the long run, as stated by Maness and Valeriano (2015).

$\pi_n = \textit{target type} + (\textit{damage type} \times \textit{intensity})$

Equation 5

Table 1: Contained

×0.5	**Data breach (0.5)**	**Reversible (1)**	**Physical (1.5)**
Private (0.5)	0.75	1	1.25
Public (1)	1.25	1.5	1.75
CNI (1.5)	1.75	2	2.25

Table 2: Extended

x1	Data breach (0.5)	Reversible (1)	Physical (1.5)	P + L (2)
Private (0.5)	1	1.5	2	2.5
Public (1)	1.5	2	2.5	3
CNI (1.5)	2	2.5	3	4.5

Table 3: Systematic

×2	Data breach (0.5)	Reversible (1)	Physical (1.5)	P + L (2)
Private (0.5)	1.5	2.5	3.5	4.5
Public (1)	2	3	4	5
CNI (1.5)	2.5	3.5	4.5	5.5

κ_n, ε_n and ϕ_n, reparations, first response's damage, and cost:

Derived from π: κ_n (reparations, Eq. 6), ε_n (first response's damage, Eq. 7), φ_n (first response's cost, Eq. 8), exist proportionally on the basis of (π_n). This reasoning follows data from *The Dyadic Cyber Incident and Dispute Dataset*. The costs of the first response grow exponentially with the complexity of the attack, as suggested by Slayton (2017), approaching the damage's worth at the highest values of ε_n.

$\kappa_n = \pi_n$
Equation 6

$\varepsilon_n = \pi_n$
Equation 7

$\varphi_n = \pi^{2.43}/10$
Equation 8

r_n, reputational costs:
Reputational costs r_n (Eq. 9) exist on the basis of π_n, p (only for the defensive agent), and *md*. Meaning that the size of suffered damage, the probability of the attacker being considered responsible and the capabilities gap all play a role in the size of reputational harm. Reputational costs tend to be larger for agents whose legitimacy lies in the country's international stature, which is why a militarization index (*md*) has been included in the calculations.

$r_n = (\pi_n) \times p \times (md)$
Equation 9

γ_n, conventionally inflicted escalation damage:
Lastly, γ_n (Eq. 10) exists on the basis of ($\pi^{1.4}$) and ΔMP, meaning that the consequences of escalation must be qualitatively more impactful than both the damage inflicted through cyberattacks and first response's damage. Δ values are calculated via dividing each agent's parameter by the sum of both agents' values.

$\gamma_n = \pi^{1.4}{}_n \times \Delta MP$
Equation 10

p(t), probability distribution over time representing attributional difficulties:
The probability distribution p captures the gist of attributional difficulties: agents struck by a cyberattack are forced to act under conditions of uncertainty, which can be modeled through a probability distribution representing the defensive agents' beliefs.

Following an attack, the defensive agent's beliefs regarding the offensive agent's culpability will shift over time, and will reach a peak as close to $p = 1$ as evidence and existing technologies allow. Early beliefs over a state's culpability are specific to the standing relations between the agents: countries involved in ongoing disputes are more likely to be suspicious of each other, thus increasing the initial value of p at t_0. From t_0 to the peak of p the function $p(t)$ will resemble a bell-shaped curve.

Past its peak, the function $p(t)$ will slowly decay, tapering differently from the values on the left side in an asymmetric fashion as relations normalize. Thus, it is possible to visualize the overall function $p(t)$ as a positively skewed Gaussian distribution, as shown in Figure 1. The slow decay of $p(t)$ is certainly one of the strongest explanations of restraint, one that links attributional difficulties with restraint in terms of conflicts' pace: to acquire a strategic gain equivalent to the highest π available again, the offensive agent is left with the only option to wait for $p(t)$'s value to reset.

It is important to note that the value of p will not be established quantitively, but rather qualitatively. Probabilistic range can be compared with a plausible estimation determined from the study of the interactions between two agents at a given time.

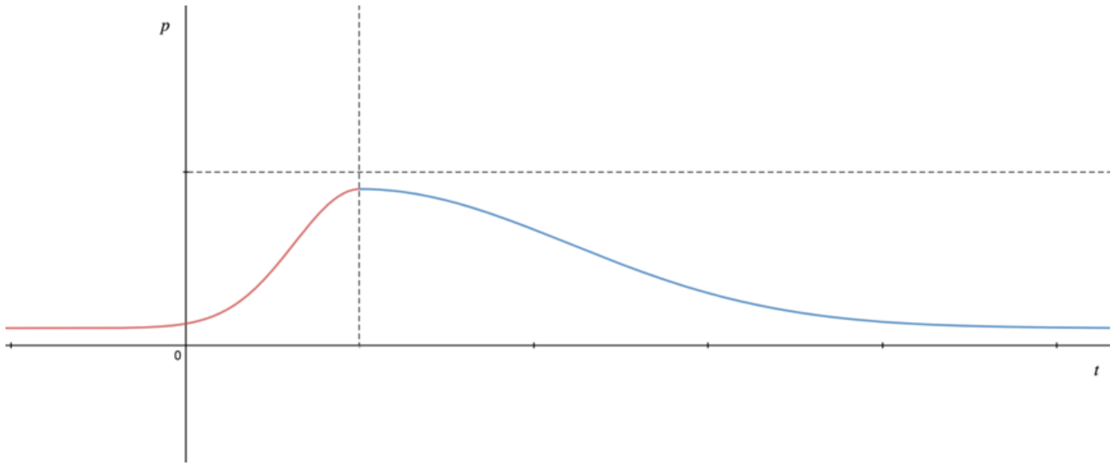


Figure 1: Probability distribution over time

2.3.2 Environment and timing

Consider an infinite horizon model in which two agents coexist. Analysis of their interactions begins at time *t* = 0, when the offensive agent can first strike. Both agents are concerned with the possession of a certain item *π*, held by the defensive actor at t_0 and acquirable through an offensive cyber strategy (for simplicity, *π* is a zero-sum element, meaning that offense and defense can trade it back and forth or acquire a reward of corresponding value in conflict).

The timing of events can be summarized as follow: in any iteration of the model (see figure 1) an offensive agent (P1) in possession of offensive cyber capabilities faces an initial choice, whether to use said capabilities or not: not acting preserves the status quo, which, for this particular game, entails the defensive agent (Player 2) holding a strategic item π. On the other hand, acting sets in motion a "lottery" outside of P1 control that, would point to P1 with probability *p*. Hence P1 can either be considered responsible for the attack (*p*), and potentially face consequences, or acquire the strategic gain (*1-p*) without repercussions.

While in the second instance the game ends successfully for P1, in the first scenario it is then up to P2 to choose between strategies, either to respond or not: not responding would terminate the game, leaving P2 to suffer both the loss of the strategic item and certain reputational costs for not acting when struck (*0–r(2)*), instead responding (thus inflicting damage, ε) would let P1 act again.

At this final node, P1 can choose to acquiesce, in which case the attacker would incur into the damage from the first response (ε), certain proportional reparations (κ) and reputational costs r(1), while the defender would benefit from reparations, although still carrying the operational costs (ϕ) of the first response; else P1 can choose to respond: in this scenario, both players can obtain the strategic item π with a certain probability, based on their conventional capabilities gap (Δτ), but both will suffer additional escalation damages (γ).

Repeated iterations show a significant difference: with each reiteration the defending actor will update his beliefs. In future iterations, therefore, the probabilities for the attacker to be considered responsible will increase by Bayesian logic.

Another possible scenario shows the two agents attacking alternatively: when the second attack is a consequence of the first, several parameters will carry over from one iteration to the next. These instances are more complex to model, especially regarding how the parameter π_1 will be affected by the second strike. Generally, if one believes that states employ offensive cyber operations to achieve a specific objective whilst avoiding the escalation of conflict, the parameter π_1 will not carry over to the second interaction in case of escalation (the payoff for the offensive agent at the last node, therefore, remains the same) unless the damage is significant and long-lasting. Previously incurred damage will be accounted for at any other node among the defensive agent's payoff, for instance at the previous node: (*π – r*(2)*)* instead of (*0 – r(*2*)*).

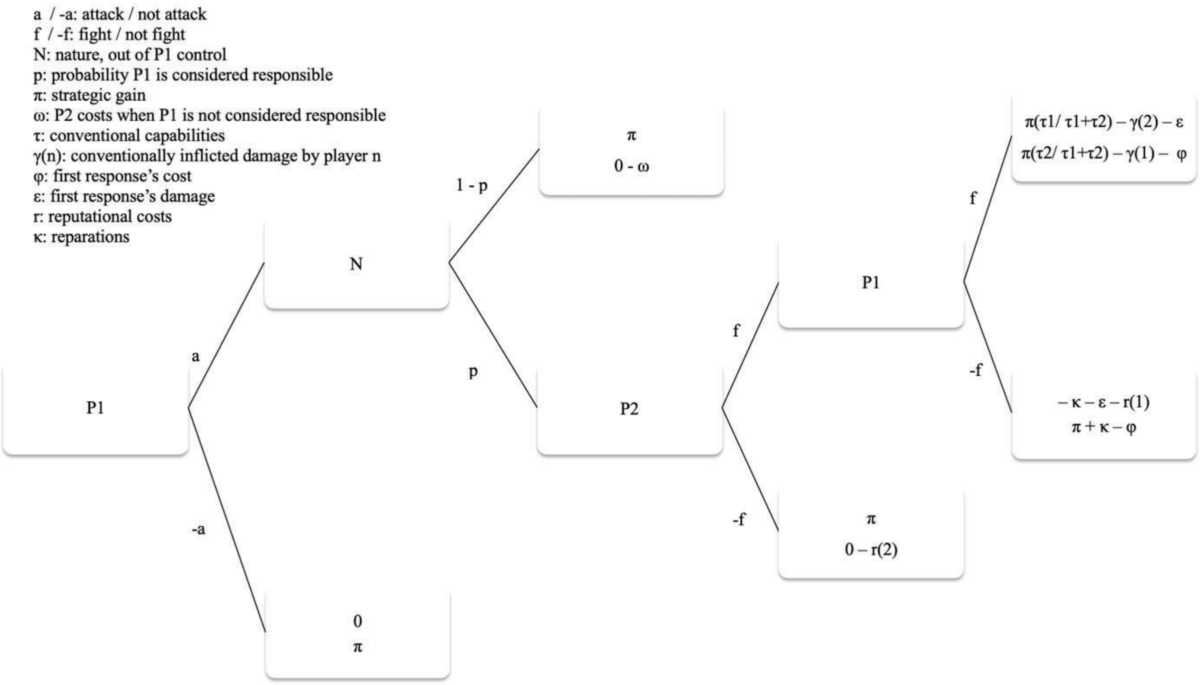

a / -a: attack / not attack
f / -f: fight / not fight
N: nature, out of P1 control
p: probability P1 is considered responsible
π: strategic gain
ω: P2 costs when P1 is not considered responsible
τ: conventional capabilities
γ(n): conventionally inflicted damage by player n
φ: first response's cost
ε: first response's damage
r: reputational costs
κ: reparations
P1
a
-a
N
0
π
1 - p
p
π
0 - ω
P2
f
-f
π
0 – r(2)
P1
f
-f
π(τ1/ τ1+τ2) – γ(2) – ε
π(τ2/ τ1+τ2) – γ(1) – φ
– κ – ε – r(1)
π + κ – φ

Figure 2: Standard model.

What certainly carries over from one iteration onto the next, is how the first interaction's defensive agent will be susceptible to escalation (Eq. 11). In the eventuality of escalation, the former defensive agent (now on the offensive) will increase the intensity of its response, adding previously incurred harm:

$\gamma_n = (\pi_n + \pi_{n+1})^{1.4} \times \Delta MP.$
Equation 11

2.4 Analysis

In this section the author will describe, via backward induction, the conditions that may prompt offensive agents to attack.

Assumption 1 claims that

$$\kappa - \phi < 0 - r(2),$$

Equation 12

the costs of initial countermeasures drawn by a defensive agent won't surpass the benefits of acquiring reparations, while 0 – r(2) assumes, with certainty, a negative value.

Assumption 2 claims that

$$p(t_n) \geq p(t_0).$$

Equation 13

By Bayesian logic *p(t_n)* can only assume a larger or equal value to *p(t₀)*. The offensive agent is given only two behavioral paths, aggression or passivity, which is a simplification, as there certainly exists behaviors that improve trust between States. This practical choice serves the purpose of highlighting the obstacles an aggressive agent would face in pursuing a repeatedly belligerent strategy.

Propositions for t_0:

If

$$\pi_0(\tau 1/ \tau 1+\tau 2) - \gamma_0(2) - \varepsilon_0 > - \kappa_0 - \varepsilon_0 - r_0(1)$$

Equation 14

and

$$0 - r_0(2) > \pi_0(\tau 2/\ \tau 1+\tau 2) - \gamma_0(1) - \phi_0,$$

Equation 15

P1 attacks for any $\pi > 0$. These very specific conditions would, in point of fact, induce conventionally powerful offensive agents to attack for any strategic gain, as stated in *Hypothesis 1*.

If

$$\pi_0(\tau 1/\ \tau 1+\tau 2) - \gamma_0(2) - \varepsilon_0 > - \kappa_0 - \varepsilon_0 - r_0(1)$$

Equation 16

and

$$0 - r_0(2) < \pi_0(\tau 2/\ \tau 1+\tau 2) - \gamma_0(1) - \varphi_0,$$

Equation 17

P1 attacks when

$$p_0 < \pi_0/(-\pi_0\ (\tau 2/\ \tau 1+\tau 2) + \gamma_0\ (2) + \varepsilon_0 + \pi_0).$$

Equation 18

In this type of equilibrium, both agents showcase the readiness to fight if necessary. Thus, an offensive player would be compelled to attack only under a favorable combination of probabilities and conventional power gap.

If

$$\pi_0(\tau 1/\ \tau 1+\tau 2) - \gamma_0(2) - \varepsilon_0 < - \kappa_0 - \varepsilon_0 - r_0(1)$$

Equation 19

and

$$\kappa_0 - \phi_0 > 0 - r_0(2),$$

Equation 20

P1 attacks when

$$p_0 < \pi_0/(\pi_0 + \kappa_0 + \varepsilon_0 + r_0\ (1)).$$

Equation 21

In this equilibrium, at t_0 the offensive agent would find reason to attack only under extremely favorable probabilities, considering that it is not willing to let the conflict escalate. Yet the size of the strategic gain π has virtually no impact on the offensive decision making: meaning that paradoxically, for a weaker offensive agent, bold assaults could theoretically stir suspects away. Perhaps, successful attacks that were not attributed with certainty fit this scenario, yet it is impossible to make confident analyses on the subject.

Propositions for t_n:

If

$$\pi_n(\tau 1/\ \tau 1+\tau 2) - \gamma_n(2) - \varepsilon_n > - \kappa_n - \varepsilon - r(1)$$

Equation 22

and

$$0 - r_n(2) > \pi_n(\tau 2/\ \tau 1+\tau 2) - \gamma_n(1) - \phi_n,$$

Equation 23

P1 attacks for any π > 0. In repeated iterations, this equilibrium leads again to attacks for any π > 0. Yet, given the growing reputational costs for the defense, the conditions in Eq. 22 and 23 are less likely to be met as $r_n(2)$ grows with each iteration. Importantly, this difference explains one of the reasons why even powerful states may attack parsimoniously in cyberspace, as stated in *Hypothesis 2.*

If

$$\pi_n(\tau1/\ \tau1+\tau2) - \gamma_n(2) - \varepsilon_n > - \kappa_n - \varepsilon_n - r_n(1)$$

Equation 24

and

$$0 - r_n(2) < \pi_n(\tau2/\ \tau1+\tau2) - \gamma_n(1) - \phi_n$$

Equation 25

P1 attacks when

$$p_n < \pi_n/(\ -\pi_n(\tau2/\ \tau1+\tau2) + \gamma_n(2) + \varepsilon_n + \pi_n).$$

Equation 26

In repeated iterations, this equilibrium is severely impacted in two ways: first, the defensive state will act less aggressively as the condition in Eq. 25 is more likely to be met. Secondly, the condition in Eq. 26 will also be drastically affected. Reiterations lead to shifts in the probability distribution, where $p(t_{n+1}) \geq p(t_n)$ always. Thus, to satisfy the condition $p_{n+1} < \pi_{n+1}/(-\pi_{n+1}(\tau2/\ \tau1+\tau2) + \gamma_{n+1}(2) + \varepsilon_{n+1} + \pi_{n+1})$ P1 can only reduce the size of the strategic gain in following *t*s, or wait. This scenario leads to a particular form of restraint, where agents are compelled to aim for smaller gains, as an offensive strategy offers solely diminishing returns (as stated in *Hypothesis 2),* or wait for *p* to return to its original value: a course of actions that lead to restraint in terms of pace.

If

$$\pi_n(\tau1/\ \tau1+\tau2) - \gamma_n(2) - \varepsilon_n < - \kappa_n - \varepsilon_n - r_n(1)$$

Equation 27

and

$$\kappa_n - \phi_n > 0 - r_n(2)$$

Equation 28

P1 attacks when

$$p_n < \pi_n/(\pi + \kappa_n + \varepsilon_n + r_n(1)).$$

Equation 29

In this equilibrium, P1 would find a reason to attack only under favorable probabilities, which are progressively shrinking with each reiteration. While, again, the size of the strategic gain π has virtually no impact on the offensive decision making, an offensive strategy becomes essentially unviable as *p* approaches the value of 1.

3. Case studies

To corroborate this research's findings, the author has compared empirical cases taken from the short history of cyber conflict with the model's forecasts. The selected cases are among the most noteworthy cyber incidents and have been picked to present a heterogeneous selection of interactions based on the power differentials between the agents: Cyberattacks on Estonia, Operation Olympic Games, Operation Ababil, and the Sony Hack. Each interaction has been simulated with the inclusion of data from *The World Bank Group* and SIPRI for the relevant year (Table 4), and the outcome of each simulation has been compared with how the conflict played out in reality.

Table 4: Casa Studies' Data

	π, k, ε	φ	$\Delta\tau$	r	γ
Russia$_{Est}$	2	0.5388934307	0.9880272362	2.453155086	2.615519426
Estonia$_{Est}$			0.01197276381	(1.28342149)(p)	0.02349639568
USA$_{Oly}$	3	1.443459652	0.9584494652	3.161450979	4.611713798
Iran$_{Oly}$			0.04155053483	(3.459613441)(p)	0.4562432134
Iran$_{Ab}$	1.5	0.2678562346	0.05809660829	1.719436906	0.05555189754
USA$_{Ab}$			0.9419033917	(1.559605102)(p)	1.666118402
DPRK$_{So}$	1	0.1	0.01645447262	1.78980528	0.02116041354
USA$_{So}$			0.9835455274	(1.527370494)(p)	0.9788395865

The 2007 DDoS campaign against Estonia (π_{Est}=2), largely attributed to the Russian Federation, and Operation Olympic Games against Iran (π_{Oly}=3), commonly attributed to the United States, are both examples of Super Powers exploiting the unconventional nature of conflict in cyberspace against conventionally weaker actors, as described in *Hypothesis 1*. Although the two campaigns differ in terms of intensity and technical proficiency, they support the model's forecasts that, regardless of attributional difficulties, significantly stronger states can employ offensive strategies in cyberspace successfully: in both instances, by backward induction, at the last node the offensive agents (the Russian Federation and the United States) would opt to escalate the conflict rather than acquiesce, given their superior capabilities (Eq. 14). At the previous node, both defensive agents (Estonia and Iran) would rather suffer the attack's damage and reputational costs than escalating the conflict (Eq. 15). As it occurred in reality, the interaction ends at this node: the offensive agent is given the choice between preserving the status quo or acquiring π without suffering immediate consequences, thus choosing to strike.

Two cyber campaigns against the United States, Operation Ababil (π_{Ab}=1.5) and the Sony Hack (π_{So}=1) respectively attributed to Iran and North Korea, show how weaker agents fare against a conventionally dominant Power.

Operation Ababil is commonly understood as a delayed response to Operation Olympic Games, and only as such, it supports the model's forecasts. If we consider Operation Ababil a response to Operation Olympic games then it would be reasonable to assume that the escalation of conflict caused by Iran would have been significant, considering both the damage suffered from Stuxnet and the potential American response to Operation Ababil (Eq. 11). The offensive agent (Iran, in this iteration) would have opted to fight at the final node (Eq. 15, with conventional damage expressed by Eq. 11). At the previous node, instead, the defensive agent (now the United States) would suffer from the strike and its reputational costs (which are low, given the modest intensity of the strike), but also weigh the strategic gains already obtained through Operation Olympic Games, thus acquiescing, given the negligible gains obtainable through escalation. As the defensive agents would prefer to acquiesce, the interaction ends at this node: the offensive agent is given the choice between not attacking or acquiring π without suffering further consequences, thus choosing to strike.

Stuxnet had achieved, as suggested by Lindsay (2015), a very specific objective, to slow down Iran's nuclear program whilst avoiding escalation: an assertive response to Operation Ababil may have once more jeopardized the subtle equilibrium between the United States, Israel, and Iran. From this perspective it would have, therefore, been reasonable for the United States to forego any response to the 2012 attack, having already achieved a much larger success previously.

The Sony Hack is the only interaction, among those here presented, which could be considered a strategic mistake by the offense: through backward induction, even if the offensive agent (North Korea) would have preferred escalating the conflict (Eq. 16), the defensive agent (The United States) would have also opted to escalate rather than suffer reputational costs (Eq. 17). At this node, the choice for the offensive agent would have been between not acting, and preserving the status quo, or losing the conflict. The optimal strategy would have been to avoid attacking, since the condition in Eq. 18 cannot be met because of the significant power gap favoring the defense.

Instead, the erroneous decision to strike can be traced back to a variety of factors: the DPRK overestimating its own conventional weapons' deterring ability, the regime's misinterpretation of how conflicts in cyberspace play out, and perhaps excessive confidence in the agents' ability to hide their tracks. Indeed, even though the attackers prevented the cinematic release of the movie "The Interview", the reported internet outage (π_{SoR}=2 approx.) that struck the country as a response to the 2014 hack surpassed North Korea's offensive operation in terms of severity, thus supporting *Hypothesis 1*.

Interestingly, the strike against North Korea and Operation Ababil show that counterstrikes in cyberspace are perhaps the best response to cyberattacks, as they circumvent many obstacles that impede the deployment of lawful countermeasures. Another key inference in support of *Hypothesis 1* can be drawn from these two counterstrikes: while Iran had to significantly lower the response's intensity to prevent the escalation of conflict, the United States' counterstrike far surpassed the Sony and still avoided the threshold of retaliation.

4. Conclusions

This research has developed a scale for the objectives of cyberattacks, which is based on their significance to the security of a state, as posed by Lindsay (2015) and Slayton (2017). This scale was then implemented, following the technology-plus approach proposed by Tang (2010), into a game theoretical model that simulates cyberattacks.

Through this methodology, the author was able to challenge the belief that cyber weapons function best as asymmetric weaponry, and to illustrate part of the reason why states that find success in cyberspace do not resort to increasingly severe cyberattacks over time: the same features that make cyber weapons effective progressively lessen when offensive strategies are repeated.

These findings suggest that restraint, both in pace and intensity, tends to emerge as a consequence of the way states exploit attributional difficulties: although attributional difficulties seem to allow the proliferation of attacks of low intensity, the opposite occurs for severe attacks.

The author has then compared cases taken from the history of cyber conflict with the model's forecasts, proposing a few observational claims: counter strikes in cyberspace are perhaps the best response to cyberattacks, regardless of power differentials. Attacks and their responses in cyberspace don't differ excessively in terms of severity, as suggested by *The Dyadic Cyber Incident and Dispute Dataset*, although the principle of proportionality appears to be non-binding for stronger agents facing weaker ones.

Above all, this research advances that, as states' interactions in cyberspace become more common, sustained restraint interrupted by sporadic attacks of medium-to-high intensity, rather than the escalation and proliferation of attacks, is likely to remain the norm.

References

Adams, K. R. (2004) Attack and Conquer? International Anarchy and the Offense-Defense-Deterrence Balance, International Security 28, no. 3.

Attiah, A., Chatterjee, M., and Zou C.C. (2017) A Game Theoretic Approach to Model Cyber Attack and Defense Strategies, ICC18.

Bell, C. H. (2018) Cyber Warfare and International Law: The Need for Clarity, Cyber Warfare and International Law, Towson University of International Affairs, VOL. LI No.2.

Białoskórski, R., Kiczma, L., and Sułek, M (2019) The National Power Rankings of Countries 2019, Research Network Warsaw.

Biddle, S. (2001) Rebuilding the Foundation of Offense-Defense Theory, Journal of Politics, Vol. 63, No. 3.

Cyber Warfare in the 21st Century: Threats, Challenges, and Opportunities, Committee on Armed Services, United States House of Representatives, One Hundred Fifteenth Congress, First Session (2017).

Glaser, C. L. and C. Kaufmann (1998) What Is the Offense-Defense Balance and Can We Measure It?, International Security 22, no. 4.

Hathaway, O. A., Crootof, R., Levitz, P., Nix, H., Nowlan, A., Perdue, W., and Spiege, J. (2012) The Law of Cyber-attack, California Law Review.

Huntington, S. P. (1987) U.S. Defense Strategy: The Strategic Innovations of the Reagan Years, in Joseph Kruzel, ed., American Defense Annual, 1987–1988 (Lexington: Lexington Books, 1987), pp. 23–43.

Jervis, R. (1978) Cooperation Under the Security Dilemma, World Politics 30, no. 2.

Jormakka, Jorma and Molsa, J.V.E.. (2005) Modelling information warfare as a game. Journal of Information Warfare. 4. 12-25.

Krepinevich, A. (2012) Cyberwarfare: A 'Nuclear Option?', Washington, D.C.: Center for Strategic and Budgetary Assessments.

Libicki, M. (2009) Cyberdeterrence and Cyberwar, Santa Monica, CA, RAND.

Lieber, K. A. (2001) Grasping the Technological Peace: The Offense-Defense Balance and International Security, International Security, Vol. 25, No. 1.

Lieberthal, K. and Singer, P. W. (2012) Cybersecurity and U.S.-China Relations, Washington, D.C.: Brookings Institution.

Liff A. P. (2012) Cyberwar: A New 'Absolute Weapon'? The Proliferation of Cyberwarfare Capabilities and Interstate War, Journal of Strategic Studies, 35:3

Lindsay, J. R. (2015) Stuxnet and the Limits of Cyber Warfare, Security Studies.

Locatelli, A. (2013) The Offense/Defense Balance in Cyberspace, ISPI.

Locatelli, A. (2015) La reazione in legittima difesa di uno Stato a fronte di un attacco cyber, CeMISS.

Maness R. C. and Valeriano, B. (2015) The Impact of Cyber Conflict on International Interactions, Armed Forces & Society.

Maness R. C., Valeriano B., and Jensen B. (2019) The Dyadic Cyber Incident and Dispute Dataset.

Nye Jr., J. S. (2010) Cyber Power, Cambridge, Mass.: Belfer Center for Science and International Affairs, John F. Kennedy School of Government, Harvard University.

Roscini, M. (2014) Cyber Operations and the Use of Force in International law. Oxford: Oxford University Press.

Schmitt, M.N. (2011) Cyber Operations and the Jus in Bello: Key Issues, Naval War College International Law Studies.

Slayton, R. (Winter 2016/17) What Is the Cyber Offense-Defense Balance? International Security, Vol. 41, No. 3.

Sułek, M. (2013) The Power of Nations. Models and Applications, Rambler, Warsaw.

Tang, S. (2010) Offence-defence Theory: Towards a Definitive Understanding, The Chinese Journal of International Politics, Vol. 3.

van Evera, S. (1998) Offense, Defense, and the Causes of War, International Security 22, no. 4.

Strategic Cyber Threat Intelligence: Building the Situational Picture with Emerging Technologies

Janne Voutilainen and Martti Kari
University of Jyväskylä, Finland
japevout@student.jyu.fi
martti.j.kari@jyu.fi

DOI: 10.34190/EWS.20.030

Abstract: In 2019, e-criminals adopted new tactics to demand enormous ransoms from large organizations by using ransomware, a phenomenon known as "big game hunting." Big game hunting is an excellent example of a sophisticated and coordinated modern cyber-attack that has a significant impact on the target. Cyber threat intelligence (CTI) increases the possibilities to detect and prevent cyber-attacks and gives defenders more time to act. CTI is a combination of incident response and traditional intelligence. Intelligence modifies raw data into information for decision-making and action. CTI consists of strategic, operational, or tactical intelligence on cyber threats. Security event monitoring, event-based response, and anomaly and signature-based detection can create the basis of the situation in cyberspace. To achieve a uniform situational picture, long-term assessment is required. Strategic CTI informs broad or long-term issues and provides situation awareness as well as an analyzed overview of the threat landscape and early warning of cyber threats. This paper describes how the implementation of artificial intelligence (AI) and machine learning (ML) can be utilized in strategic CTI. The results were arrived at using the design science research methodology. We propose a solution that uses AI as a component of strategic CTI. Furthermore, the paper is a literature survey, integrating research literature on intelligence, cybersecurity, and AI. The paper presents the concept of CTI and its relation to the situational picture of cyberspace. It also addresses the possibilities of natural language understanding for large-scale content analysis and introduces a solution in which an existing enriched dataset provided valuable strategic-level information about an ongoing malicious cyber event. The paper is part of Ph.D. research concerning comprehensive CTI. Other articles in the dissertation discuss emerging technologies in operational and tactical CTI.

Keywords: Strategic Cyber Threat Intelligence, Machine Learning, Artificial Intelligence

1. Introduction

During the 2000s, cyberspace matured into the fifth domain of warfare, and the volume of criminal activity in cyberspace increased considerably. Cyberspace is not only a technical issue, but it also has a strategic dimension. Cyberspace has expanded warfare to a global scale, beyond the traditional use of military force. The use of force in cyberspace to cause a strategic-level malicious impact does not require state-level resources. The cyber domain differs from other domains of warfare because actors in cyber conflicts are not always military. State actors, criminals, and terrorists attack state authorities, the media, companies, universities, and critical infrastructure. State actors and private cybersecurity companies are fighting against these attackers. The detection of attacks, intelligence collection on them as well as the analysis and attribution of the attacker is often difficult or impossible to do.

A cyber-attack is defined as "an attack, via cyberspace, targeting an enterprise's use of cyberspace for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure, or destroying the integrity of the data or stealing controlled information."

Ransomware is malicious software that can be used for cyber -attack. With ransomware infection, the computer system is manipulated so that the victim can not access the data on it. In most cases, shortly after infection victim receives a blackmail message, that includes the demand to pay ransoms to regain access to the encrypted filesystem. Ransomware often disseminates via phishing emails. Also, complicated propagation mechanisms exist against high profile targets such as large organizations (ENISA 2020).

Cybersecurity is a process of protecting information by preventing, detecting, and responding to cyber-attacks. The cyber threat is an event or condition that has the potential to cause asset loss and the undesirable consequences or impact of such loss (NISTIR 7298, 2019). Cybersecurity has been based on incident response and security operations (ENISA, 2018) where the focus has been on monitoring security events and incidents and in reacting to detected incidents. Internal threat monitoring has not always provided sufficient time for the defender to be proactive.

Incident response has been the methodology of cybersecurity since the 2010s. Incident response or incident handling means the mitigation of an adverse event in a computer system or network caused by the failure of a security mechanism or an attempted or threatened breach of these mechanisms. In cyber defense based on incident response, the attacker always seems to be a few steps ahead. The prevailing methods of cybersecurity, such as security event monitoring, event-based response, and the anomaly or signature-based detection, will not improve the situation. The development and implementation of tactics, techniques, and procedures (TTP) of cyber threat intelligence, including intent-based response and detection based on an attacker's behaviour, improve the situation and give a defender the possibility to get ahead of the threat actor.

CTI consists of strategic, operational, or tactical intelligence on cyber threats. Security event monitoring, event-based response, and anomaly and signature-based detection can create the basis of the situation in cyberspace. To achieve a comprehensive situation picture, long-term assessment is required. Strategic CTI informs broad or long-term issues and provides situation awareness as well as an analyzed overview of the threat landscape and early warning of cyber threats. This paper describes how the implementation of artificial intelligence (AI) and machine learning (ML) can be utilized in strategic CTI. The results were arrived at using the design science research methodology. We propose a solution that uses AI as a component of strategic CTI. In 2019 the Finnish Parliament passed new intelligence laws, the implementation of which requires new methods to achieve the goals of national CTI. For that reason, the call for the study stems from practical needs.

2. Methodology

The methodology used in this paper is a combination of the intelligence question and the design science research process (DSRP). The study focuses on the phenomenon known as big game hunting. The intelligence question for the study was provided by the National Cyber Security Centre Finland (NCSC-FI). The purpose of the intelligence question was to initiate the intelligence process.

The DSRP is a set of analytical techniques and perspectives for information systems (IS) research. There are two activities in DSRP for improving knowledge in the IS domain: the generation of knowledge through the design of new and innovative things or processes and the analysis of things and processes via reflection and abduction. In the scope of DSRP, an example of things and processes mean algorithms, human/computer interfaces, and system design methods or languages. A common term for objects and processes in the DSRP is an artifact (Vaishnavi, Kuechler et al., 2004).

The DSRP cycle begins with awareness of the problem. Usually, the dilemma comes from industry or various areas within the information technology sector. In this study, the problem was initiated by the National Cyber Security Centre Finland (NCSC-FI): the need for a comprehensive threat analysis of the emerging cyberspace phenomenon known as big game hunting. The proposal for solving the problem was researched concerning emerging technologies in the CTI process.

The second phase of DSRP is the suggestion and tentative design. For the study, the second phase included the utilization of IBM cloud capabilities in collecting, processing, and analyzing information on strategic cyber intelligence processes.

The development of an artifact is based on various questions concerning IBM cloud capabilities. According to McDowell, the data collection for strategic intelligence should be comprehensive, from various sources, with the collected data often being qualitative (McDowell, 2009). For that reason, the selected artificial intelligence subtype was IBM Watson Natural Language Understanding (NLU). As a result of this process, an artifact is created. The artifact in this study was a combination of an IBM Watson NLU machine learning model and the strategic cyber threat intelligence process.

The artifact was evaluated with performance measures. The purpose of the study was to evaluate the suitability of IBM Watson when used in CTI. The performance measures were based on the following questions: Can the artifact find related information from the data? Can the artifact analyze and process the collected data? Can the artifact provide enough information for the intelligence direction and its sub-questions?

The conclusion from the first DSR cycle was that the ML model created according to the intelligence direction did not provide enough reasonable information for the intelligence question. As a result of insufficient data, a new DSR cycle was initiated. According to DSRP knowledge flows in figure 1 (Vaishnavi, Kuechler et al., 2004), in the development of the improved artifact the missing properties were improved by using a IBM Watson Discovery News pre-enriched dataset. The second solution fulfilled performance measures. As a result of the queries conducted from the data, new and valuable information about big game hunting was found.

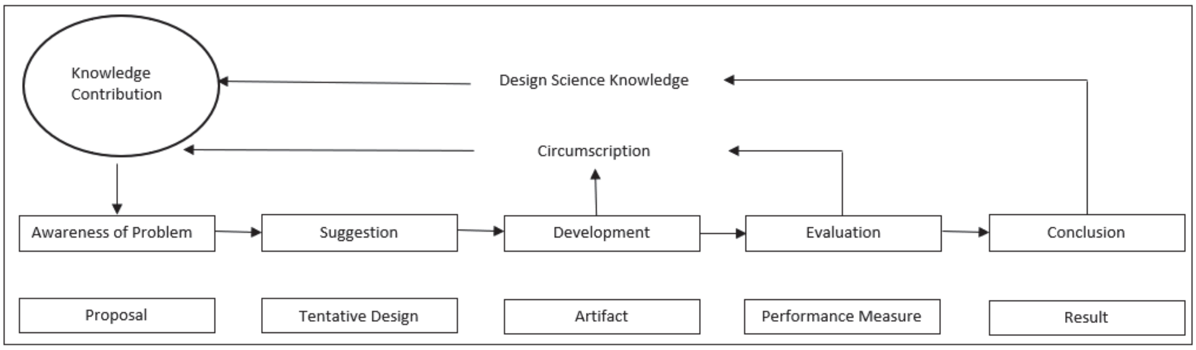


Figure 1: Design science research process (Vaishnavi, Kuechler et al., 2004)

3. Cyber Threat Intelligence

According to European Union Agency for Cybersecurity (ENISA, 2018), cyber threat intelligence, even though it is still evolving and in the early adoption phase, is becoming a critical part of cybersecurity capabilities and activities. There is no widely accepted definition of cyber threat intelligence (CTI). ENISA's definition of CTI is as follows: "Cyber threat intelligence is the process and product resulting from the interpretation of raw data into information that meets a requirement as it relates to the adversaries that have the intent, opportunity, and capability to harm" (ENISA, 2018).

CTI includes information on a threat actor's TTPs, threat indicators on impending or ongoing cyber-attacks or on successful compromise, and security alerts and threat reports (Jasper, 2017). TTPs describe the tactical and technical activities of an actor, including their tendency to use a specific malware variant, order of operations, attack tool, delivery mechanism (e.g., phishing or watering hole attack), or exploit. Tactics are the description of the way an adversary carries out his attack from the beginning, i.e., from reconnaissance to the end, command, control, and action on objectives inside the compromised information system. Techniques give a more detailed description of behaviour in the context of a tactic, and procedures an even lower-level, more highly detailed description in the context of a technique.

The challenge for CTI is the enormous volume, velocity, and diversity of different data sources. An organization's network can generate petabytes of data per second and CTI has to be capable of detecting a threat from among all of this traffic. The data are collected from many sources, and the format of the collected data may vary as well. This requires a process for collecting and analyzing data in a timely fashion and creating CTI products that adequately support decision-making. The sources of CTI are typically open data sources, commercial feeds, shared intelligence, and asset information (Doerr, 2018).

CTI can be considered a combination of incident response and traditional intelligence. The purpose of CTI-based defense is, in the best case, to get ahead of the malefactor or to at least be ready if a cyber-attack occurs. Alan Breakspear, the President of the Canadian Association for Security and Intelligence Studies (CASIS), has stated that intelligence is misunderstood and too often focused only on threats, missing opportunities for advantage. His proposal for a definition of intelligence is the following: "Intelligence is a corporate capability to forecast change in time to do something about it. The capability involves foresight and insight, and is intended to identify impending change which may be positive, representing opportunity, or negative, representing a threat" (Breakspear, 2012).

Intelligence is a process, described by the intelligence cycle. The intelligence cycle is the process of developing raw data into information and delivering this information (i.e., intelligence) to policymakers to use in decision-making and action. The aim is to increase the knowledge and, in the best case, the wisdom of decision-makers by providing them with objective intelligence at the right time. The intelligence cycle consists of several steps, with Mark Lowenthal presenting a cycle that consists of seven phases. These phases are identifying and setting requirements, collection, processing and exploitation, analysis and production, dissemination, consumption, and feedback. In this paper, the information collection, procession, exploitation, and analysis are included in the evaluation of the AI capabilities.

Identifying and setting requirements means defining and prioritizing requests for information by the clients, that is, the questions to which intelligence is expected to answer. Collection means information or data gathering from diverse sources, including open-source data. Processing and exploitation mean the process of converting the collected data for analysis. Conversion can include translations, decryption, interpretation, and evaluation of data. Evaluation means assessment of the type, quality, and format of data (information) in a way that promotes reliability and prevents supposition. In the analysis and production process the collected data is processed by appropriate analysis techniques and tools to inform and compiled to the intelligence product. In dissemination, the intelligence product, responding to the client's request for intelligence, is distributed to clients. A dialogue between clients and intelligence producers should continue during the whole process. Feedback from the client indicates how well their intelligence requirements are being met.

CTI is the process of collecting, analyzing, and disseminating intelligence on cyber threats to protect information and information systems by preventing and detecting cyber-attacks. The purpose of CTI is to understand the attacker and attacks, help anticipate future actions and plan a response and cyber defense. CTI can be illustrated and explained in the seven-phase process described by Lowenthal. Like all intelligence, CTI produces accurate, timely, and relevant intelligence, improves cyber threat information and supports the client in identifying threats and opportunities. CTI can be strategic, operational, or tactical.

Strategic cyber threat intelligence is typically non-technical, risk-based intelligence on broad or long-term issues and provides an overview of the threat landscape. It might also provide early warning of threats. Strategic cyber threat intelligence constructs a situation picture of the intent and capability of cyber threats. A strategic cyber threat picture can include information on geopolitical events and trends, combining cyber-attacks on geopolitical conflicts and events, information on malicious actors, threat actor tactics and targets, tools, and Tactics, Techniques and Procedures (TTP)s and their changes over time, and trends and patterns of emerging threats and risks (CIS, 2019). Strategic CTI analyses trends, emerging risks while creating and updating a strategic-level situation picture of the possible broad impacts and consequences of cyber-attacks. Strategic CTI differs from other CTI categories because it is mainly non-technical and produced for senior leadership instead of technical personnel. Typical strategic CTI products are policy papers, white papers, assessments, and threat reports which present a strategic cyber threat picture.

Strategic CTI requires a human element because it is time-consuming to evaluate and test new adversary TTPs against existing security controls. Parts of the process of strategic threat intelligence can be automated, but an intelligence analyst is needed for the production. Strategic CTI focuses on assessing and mitigating current and future risks to operations and businesses by answering the questions of who is causing the cyber threat and why.

The products of strategic CTI include assumptions about the nature and role of the cyber-attacks and the attacker, and about the threat posed by the attacker. The aim is to support decision-making about which defensive measures will be the most effective. Strategic CTI also supports security policy planning and implementation as well as resource allocation.

Operational CTI assesses specific, potential incidents related to events, investigations, or activities. Operational CTI relates to specific attacks or campaigns and answers the questions of how the threat is created and where it is projected. It helps defenders understand the nature, intent, and timing of a specific attack, and also provides insight into the nature and sophistication of the groups) responsible. In many cases, however, only partial context can be obtained (Recorded Future, 2018). Operational intelligence is often related to campaigns, malware, and tools, and may come in the form of forensic reports. Operational intelligence is also referred to as technical threat intelligence because it can include technical information about cyber attacks. Technical information can include attack vectors, exploited vulnerabilities, and command and control domains used by the attacker.

Tactical CTI is intelligence on real-time events, investigations, or activities. It answers the question of why and provides support for day-to-day operations and events, such as the development of signatures and indicators of compromise. Typical indicators of compromise are the presence of malware, signatures, exploits, vulnerabilities, and IP addresses. Indicators of an attack, such as code execution, persistence, stealth, command control, and lateral movement within a network, are early warning signs that an attack may be underway or has already occurred.

The cyber kill chain model, published by Lockheed Martin in 2011, has been the most used framework to explain the phases and process of a cyber-attack. In the kill chain model, the cyber-attack consists of seven phases. In reconnaissance, the attacker selects and researches a target and identifies vulnerabilities of the target network. In weaponization, the attacker creates remote access malware weapons tailored for the identified vulnerability. The attacker delivers the malware to the target system where the malware exploits the identified vulnerability and installs access points to the target system. The attacker then uses this access point to gain command and control of the target system and starts actions to achieve their goals as data exfiltration, data destruction, or encryption for ransom (Hutchins, Cloppert & Amin, 2011). The ultimate goal of the CTI process is to produce a situational picture of cyberspace. Notably, it is the changes in events that belong to the scope of the process.

Since the beginning of the decade, defining a cyberspace situational picture has been one of the most popular topics in research on cyberspace. According to Bartch et al (2018)**,** with a decent situational picture, it is possible to obtain situational awareness that contains the latest information about risk landscape, vulnerabilities, and cyber defense capability readiness (Bartch et a,l 2018).

According to JP 3-12, a detailed, accurate and comprehensive situational picture of cyberspace operations is the key element for correct decision-making in a dynamic and continuously developing environment. The observation of adversarial behavior is based on signals intelligence (SIGINT) and analysis of exploitations (Joint Chiefs of Staff, 2018)

As in Figure 2, the JP 3-12 definition of the observation of adversarial behavior positions itself as tactical CTI, analyzing the dataflow. Tactical CTI is the most critical way to make observations from cyberspace, but operational and strategic CTI should not be underrated.

4. AI-based solutions for strategic CTI

According to Brown and Lee, the area of automation is a growing area of interest in CTI. Still, nowadays, the organizations use legacy management tools, such as emails and spreadsheets for managing the strategic CTI. The overall capability of CTI can easily be increased by researching the new use cases and the possibilities of modern technologies. (Brown, Lee 2019)

When the phases of the strategic cyber intelligence process are observed from the perspective of AI-based technologies, the first phase, intelligence direction, requires human capabilities to initiate the intelligence process. Despite the continually growing speed of the technology, initiation requires a human. Intelligence direction is usually a task that is shaped as an intelligence question. In the 2019 cyber threat landscape, the malicious phenomenon known as big game hunting appeared. It refers to the launching of ransomware attacks against large organizations to obtain a significant amount of money in bitcoins (Feeley, Hartley et al., 2019). The intelligence direction from decision-makers in large organizations might be, for example, "Provide information about the risk of big game hunting against my organization." The idea of using AI in strategic intelligence is presented in *Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management* (Liebowitz, 2006) It states that AI can improve knowledge management for making strategic decisions in various organizations.The first point in the strategic intelligence cycle where AI-based technology can be used is information collection. According to McDowell, the volume of collected data for strategic intelligence might be significant and the data requirements are complicated. Furthermore, the collection should happen from all possible sources and the collected data might be qualitative, anecdotal, or even impressionistic (McDowell, 2009).

In this study, two alternatives for utilizing AI for strategic cyber intelligence are presented. Both solutions were arrived at using the DSRP. The first solution did not pass the performance measures. For that reason, the DSR cycle was initiated again. In the second solution, the approach was dissimilar. Instead of using limited source data, a dataset called IBM Watson Discovery News, which updates continuously, was used. As a result, the second solution proved to be suitable for strategic CTI.

4.1 The first solution

In the first solution, the source data were provided by NCSC-FI. The data included 1,300 cybersecurity documents from various internet sources. For analyzing the data and documents, an ML-type system was created in the IBM cloud. The type system consists of entities and their relations. The selection of entities and relations was based on an estimate of how the type could support the intelligence question in the best way. The type system used is presented in Figure2.

The type system was annotated with reliable ransomware-related documents published by the European Union Agency for Cybersecurity (ENISA). Annotation is a task where the training data words and relations are linked to the correct entity and relation as in figure 3. Finally, the model was trained to find entities and relations from new documents. Once adequate accuracy was achieved, the model was deployed to Watson Discovery, where the queries were conducted from the 1,300 documents that NSCS-FI provided.

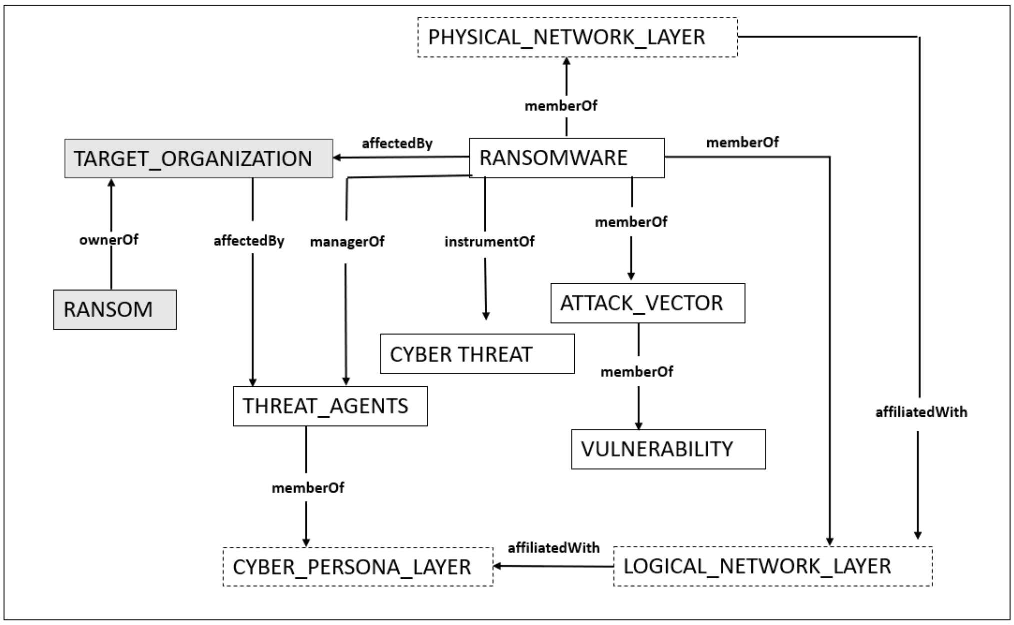


Figure 2: Type system (Voutilainen 2019)

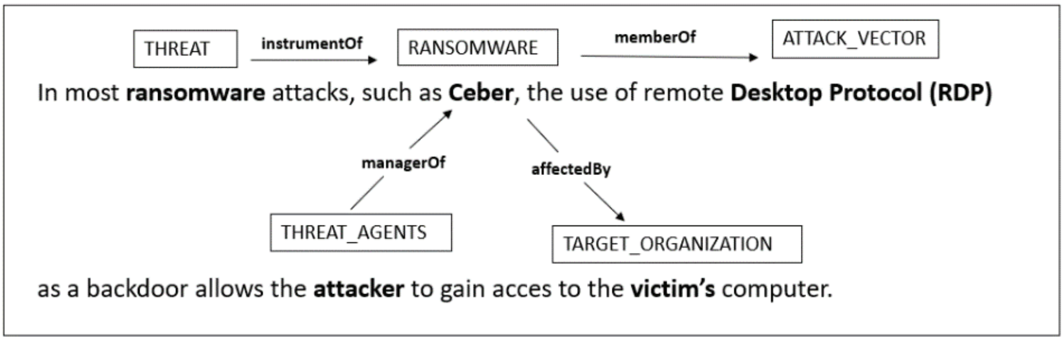

THREAT
instrumentOf
RANSOMWARE
memberOf
ATTACK_VECTOR
In most **ransomware** attacks, such as **Ceber**, the use of remote **Desktop Protocol (RDP)**
managerOf
affectedBy
THREAT_AGENTS
TARGET_ORGANIZATION
as a backdoor allows the **attacker** to gain acces to the **victim's** computer.

Figure 3: Annotation (Voutilainen 2019)

As a result, the model was able to find new information from the documents. An interesting observation was the passages. Passages are short sentences that match the best way to the query. For example, when the data was queried with the words *ransomware* and *big game hunting*, Watson returned passage: *"Both INDRIK SPIDER (with BitPaymer ransomware) and GRIM SPIDER (with Ryuk ransomware) have made headlines with their high-profile victims and ransom profits, that big game hunting is a lucrative enterprise."* The disadvantage in the first solution was the amount of data. The 1,300 documents used were insufficient for the required in-depth analysis of strategic cyber intelligence. Instead, the corresponding approach would be suitable for operational cyber threat intelligence, where the source data might be, for example, forensic reports. The organizations that require high security for data might use similar solutions for tracking phenomena similar to big game hunting by using their data. The conclusion from the first solution was that IBM Watson can be used for information processing and analyzing phases of the strategic intelligence process. In the solution, the collection phase of the strategic intelligence process was not tested since NCSC-FI initially vetted the dataset.

4.2 The second solution

The approach to the second solution was different and it proved to be more efficient. The IBM Watson Discovery News dataset was used as source data for the second solution. The data are public, including over 14,000,000 documents, and the sources for the dataset are in five languages. The majority of the documents are from English news sites that are updated continuously. The oldest information in the Watson Discovery data is 60 days old. The data changes every day, with IBM adding approximately 300,000 new articles from news and blogs daily (Scherping, 2017).

The main difference compared to the data that were used in the first solution is that Watson Discovery News does not support corresponding custom models that were used in the first solution. Further, Discovery News cannot be trained, so there is no possibility to add documents and the dataset cannot be configured for specific use cases. Still, according to IBM documentation, the proposed use cases for Discovery News are news alerts, event detection and obtaining data from trending topics (IBM, 2019). These features are a good fit for strategic cyber threat intelligence. The data in Watson Discovery News include multiple issues from various areas, in addition to cybersecurity-related documents.

Various queries were tested in the second solution. When the correct queries were found, Watson Discovery News quickly provided the required information. In addition, the queries can be made with a graphical interface. The following example query was used to obtain information about the intensity, geographical location, and time perspective of the phenomenon: "Find the number of documents that include the words *big game hunting* and *ransomware*. Aggregate the results by the publishing date and the country where the document has been published and show the top five results." The returned information is presented in figure 4. The result gives a rough guideline of how big game hunting has developed during the first half of 2019. The obtained trend proved to be correct even when the time and geolocation were compared to the number of documents.

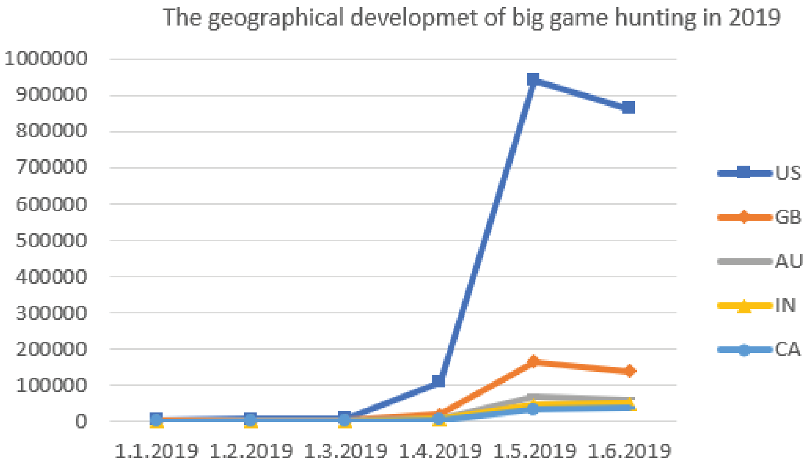


Figure 4: The geographical development of big game hunting in 2019 (Voutilainen 2019)

5. Conclusions

To achieve a uniform situational picture from cyberspace events, the need for emerging technologies is evident. All three components – strategic, operational and tactical – are equally important. Strategic CTI requires information from all available sources. The amount of data is enormous and, with human capabilities, it might be extremely difficult to find the required intelligence information from the source data.

In the study, IBM cloud was used as a platform to collect, process and analyze the data for intelligence direction. Both presented solutions based on machine learning. In the first solution, *An ML model that is planned according to intelligence direction to provide information about big game hunting*, the restricted dataset was queried with an NLU-based model from initially vetted data. The model functioned, but it did not provide enough information that could be used for the intelligence question. It was based on supervised learning, where the data labels are created for the training data. When new data was ingested in Watson Discovery News, the service classified the documents according to the labels. Moreover, during the ingestion, the service created labels using unsupervised learning. The queries that were conducted with unsupervised learning based on labels produced during ingestion provided almost identical answers to the queries. From that perspective, unsupervised learning is a better alternative, because creating a supervised learning-based model is laborious. The model in the first solution might be used in organizations that use closed environments and require high data security.

The second solution, *Targeted queries from Watson Discovery News to answer an intelligence question concerning big game hunting*, proved to be useful for strategic CTI. First, the amount of data met the requirement for source data. Second, the data are updated continuously, which means the ability to obtain an almost real-time situational picture of the investigated phenomenon. Finally, the pre-enriched dataset does not require a separate ML model. In the second solution, the ML capabilities of Watson Discovery News happens when data are crawled from the available sources. The documentation of the IBM CLOUD does not provide details on how the source documents are labeled. Yet, depending on the multiple possibilities of the queries, valuable information within the scope of the intelligence question can be obtained.

The main finding of this paper is the suitability of NLU and text analytics for strategic CTI. There exist multiple similar cloud services where corresponding queries can be made with various documents. In the future, the possibilities to combine refined intelligence information from strategic-, operational-, and tactical level should be researched. Also, the precisely tailored solution from the neural network level to NLU for strategic CTI is a potential research topic. In future research, data security needs to be considered in each solution.

References

Bartsch, M., Frey, S, ed, (2018). Cybersecurity Best Practices. 1 edn. Bern, Swizerland: Springer Vieweg.

Breakspear, A. (2012) A New Definition of Intelligence. *Intelligence & National Security*, Volume 28, 2013 - Issue 5, p678-693. [Online]. Available at: https://doi.org/10.1080/02684527.2012.699285 (Accessed: 05 November 2019).

Bown, R., Lee, R.M.,(2019). The Evolution of Cyber Threat Intelligence (CTI): 2019 SANS CTI Survey. SANS. Available at :https://wow.intsights.com/rs/071ZWD900/images/The%20Evolution%20of%20Cyber%20Threat%20Intelligence%20%28CTI%29%202019%20SANS%20CTI%20Survey.pdf (Accessed: 17 February 2020)

CIS. (2019) Center for Internet Security, Inc. What is Cyber Threat Intelligence? [Online]. Available at: https://www.cisecurity.org/blog/what-is-cyber-threat-intelligence/ (Accessed: 05 November 2019)

Doerr, C. (2018) Cyber Threat Intelligence Standards. - A high-level overview. TU Delft. [Online]. Available at: https://www.enisa.europa.eu › cti-eu-2018-presentations (Accessed: 21 December 2019).

ENISA. (2018) Exploring the opportunities and limitations of current Threat Intelligence Platforms. [Online]. Available at: https://www.enisa.europa.eu/publications/exploring-the-opportunities-and-limitations-of-current-threat-intelligence-platforms (Accessed: 05 November 2019).

ENISA, 2020-last update, Ransomware. Available at: https://www.enisa.europa.eu/topics/csirts-in-europe/glossary/ransomware (Accessed: 17 February 2020).

Feeley, B., Hartley, B. and Frankof, S., -03-06T00:04:48+00:00, 2019-last update, PINCHY SPIDER Adopts “Big Game Hunting” to Distribute GandCrab. Available at: https://www.crowdstrike.com/blog/pinchy-spider-adopts-big-game-hunting/ (Accessed: 05 July 2019).

Hutchins, E, Cloppert M and Amin, R 2011. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. *Lockheed Martin Corporation*. [Online]. Available at: https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf (Accessed: 05 November 2019).

IBM. (2019) IBM Cloud Docs. Available at: https://cloud.ibm.com/docs/services/discovery?topic=discovery-about (Accessed: 02 July 2019).

Jasper, S. (2017) *Strategic Cyber Deterrence: The Active Cyber Defense Option*. London: Rowman & Littlefield Publishers.

JOINT CHIEF OF STAFF, 2018-last update, JP 3-12, Cyberspace Operations, 8 June 2018 - jp3_12.pdf. Available: https://fas.org/irp/doddir/dod/jp3_12.pdf. (Accessed: 10 November 2019)

Liebowiz, J., (2006) *Strategic Intelligence : Business Intelligence, Competitive Intelligence, and Knowledge Management.* London: Auerbach Publications.

Lowenthal, M. (2006) Intelligence Cycle and Process. [Online]. Available at: https://www.e-education.psu.edu/sgam/node/15 (Accessed: 07 November 2019).

McDowell, D., (2009) *Strategic intelligence : a handbook for practitioners, managers, and users.* Rev. edn. Lanham (Md.): Scarecrow Press.

NIST. (2018) SP 800-150 GUIDE TO CYBER THREAT INFORMATION SHARING. [Online]. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-150.pdf (Accessed: 05 November 2019)

NISTIR. (2019) NISTIR 7298. National Institute of Standards and Technology. [Online]. Available at: https://csrc.nist.gov/glossary/term/Cyber-Attack (Accessed: 05 November 2019).

Recorded Future. (2018) How Operational Threat Intelligence Blocks Attacks Before They Happen. THE RECORDED FUTURE TEAM. [Online]. Available at: https://www.recordedfuture.com/operational-threat-intelligence/ (Accessed: 05 November 2019).

Scherping, J., -04-13, 2017-last update, Out of the box, Discovery provides a pre-enriched collection of 2 months of internet news content. Available: https://www.ibm.com/blogs/watson/2017/04/box-discovery-provides-pre-enriched-collection-2-months-internet-news-content/ (Accessed: 05 May 2019).

Vaishnavi, V., Kuechler, B. and Petter, S., 2004. Design Science Research in Information Systems. (2004/17), pp. 1-62.

Voutilainen, J., 2019. Machine learning and intelligence cycle: enhancing the cyber intelligence process, University of Jyväskylä. Available at: https://jyx.jyu.fi/bitstream/handle/123456789/65229/URN%3ANBN%3Afi%3Ajyu-201908143828.pdf?sequence=1&isAllowed=y

Event Tracing for Application Security Monitoring in Linux

Jack Whitter-Jones, Mark Griffiths and Ross Davies
University of South Wales, Pontypridd, UK
jack.whitter-jones@southwales.ac.uk
mark.griffiths@southwales.ac.uk
ross.davies@southwales.ac.uk

DOI: 10.34190/EWS.20.072

Abstract: The modern corporate landscape is one that encourages the deployment of intrusion detection and incident response as a core part of an organisation's security strategy. With the growing sophistication of cybercrime, organisations are looking at more ways to provide early indicators and contextual understanding to their security analysts to aid in mitigating security breaches. This paper presents the use of event tracing for intrusion detection within the Linux operating system via the Extended Berkley Packet Filter and gives an evaluation of its effectiveness. Within the research, one primary aim and one secondary aim is established. The primary aim of the research is to evaluate the use of event tracing to provide intrusion detection capabilities within the Linux operating system, with a further evaluation on how event tracing could compliment current monitoring techniques, e.g., Log collection. The secondary aim of this research is to discuss the possibilities of using event tracing within the emerging and embedded technologies market. To provide answers to these aims the research involves an experimental method to evaluate how event tracing can be used to detect two web-based attacks: SQL injection and path traversal. The results of the experiment present three key findings. Firstly, the use of event tracing within the Linux operating system can provide clear indicators of SQL Injection and web traversal attacks. Secondly, event tracing can introduce another way of interacting with running applications to provide forensic data, while maintaining confidentiality and integrity. Finally, event tracing can complement traditional logging techniques to supply further context and indicators relating to web-based attacks. Future work will target intrusive behaviour through the adoption of event tracing to supply indicators of attack and compromise within the emerging and embedded technologies market. Such work within the field could help advance the security community in monitoring very ad-hoc security practices.

Keywords: Event Tracing, Application Security, Continuous Monitoring

1. Introduction

The modern corporate landscape is a growing industry that encourages intrusion detection and incident response as a core part of an organisation's security strategy. With the growing sophistication of cybercrime, organisations are looking at more ways to provide early indicators and contextual understanding to their security analysts to aid in mitigating cyber-attacks (Janos and Dai, 2018). As data is integral to forensic investigation (Onwubiko, 2018), offensive methodologies aim to disrupt the integrity and availability of artefacts that may be recorded during a cyber-attack (Hutchins, Cloppert and Amin, 2011), for example, deleting log files to hide the installation of Malware. Event Tracing for Windows (ETW) a traditional debugging application within the Windows operating system (Microsoft, 2019) has been augmented to aid intrusion detection capabilities (Microsoft, 2017) without generating a log file which could have its integrity and availability compromised. The primary aim of the research is to evaluate the use of event tracing to provide a proof of concept regarding intrusion detection and potential contextual understanding that could aid traditional log collection techniques within the Linux operating system. The secondary aim is to discuss the possibilities of using event tracing within the emerging and embedded technologies market. With these aims in mind, the following sections include: the methodology of experimental implementation; results and analysis; discussion around key findings; and concluding remarks and next steps.

2. Method

To best represent a Linux environment in production, a model of a Linux web server will be used as the experiment base, as illustrated in Figure 1. Of the web servers worldwide, 71% run Linux (W3Techs, 2020a), 79% use PHP (W3Techs, 2020c), 42% use Apache2 (W3Techs, 2020b), and 39% use MySQL (Scalegrid, 2019). Using these popular technologies, a virtual machine was constructed through VMWare Workstation 15.0.3 with the following applications installed using default settings:

- **Linux: -** x86_64 18.04.3 LTS, Kernel Version: 4.15.0-65-Generic,
- **Apache2: -** 2.4.29 Ubuntu,
- **MySQL: -** 14.14 Distribution 5.7.28 for Linux,
- **PHP 7.2: -** Zend Engine 3.2.0.

To carry out web-based attacks a **Kali 2019.4** virtual machine created in **Virtual Box 6.1**. To simulate the path traversal attack, **Dirbuster** was used, with the following settings changed from a standard launch:

- **Wordlist**: - directory-list-2.3.small.txt
- **File Extension**: - php, html, txt

To simulate an SQL injection, connection to the MySQL server was established to generate SQL queries. Log files that are generated will be compared to event trace data that are captured by a **Python-3.7 Middleware** and sent to **Splunk,** a popular Security Incident and Event Management system. Finally, to watch the effect of the middleware, a separate **Python-3.7 Systems Monitor** script was used to collect CPU and Real Memory over a set time limit. The primary aim of the experiment setup is to generate synthetic Apache2 errors and MySQL log data and capture event traces targeted at Apache2 errors and SQL queries to be captured and analysed for intrusive behaviour.

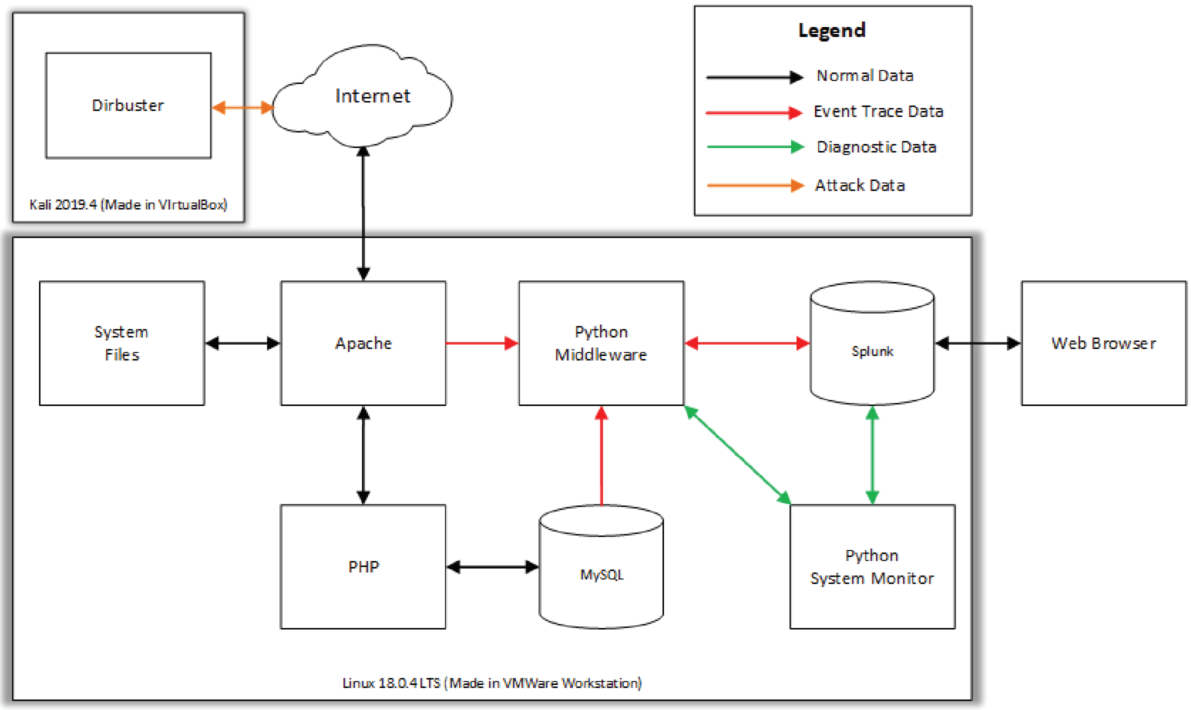


Figure 1: Experiment setup for capturing attack data through event tracing

3. Results

The following section presents two areas of information. Firstly, an evaluation of event tracing alongside log files will be conducted to analyse potential indicators of threat regarding SQL injection and path traversal. Secondly, an analysis of the event tracing process before (Figure 2) and after (Figure 3) an attack will be conducted to show a potential indicator of intrusive behaviour.

3.1 Evaluation of Event Tracing

Table **1** presents a comparison of an event trace of Apache2 errors and SQL queries alongside log file entry for the applications Apache2 and MySQL. The evidence found that event tracing of Apache2 errors and SQL queries supply understanding of what content is being accessed without showing any post data. This approach can help analysts refine their search queries and provide clarity in their results. In comparison to traditional log collection, the standard Apache2 log does provide more information of web data such as user agent and response code, however, this use case could be improved through further development of the event tracing script used. Reviewing the standard MySQL log file, presents limitations in general contextual understanding of the query being made.

Table 1: Comparative view of event traces and log files from Apache2 and MySQL

	Apache2	**MySQL**

Event Trace	{
	"MACAddress": ["00:50:56:9c:52:12"],
	"pid": 18076,
	"hostname": "jacksphd",
	"request": "GET /robots.txt HTTP/1.1",
	"IPAddress": ["192.168.171.124"]
	}
		{
		"MACAddress": ["00:50:56:9c:52:12"],
		"query": "DELETE FROM mysql.db WHERE Db='test' OR Db='test\\_%'",
		"hostname": "jacksphd",
		"pid": 30478,
		"IPAddress": ["192.168.171.124"]
		}

Log File	192.168.171.124 - - [20/Mar/2019:11:02:34 +0000] "GET /robots.txt HTTP/1.1" 404 486 "-" "Mozilla/5.0 (compatible; Nmap Scripting Engine; https://nmap.org/book/nse.html)"	2019-03-08T15:05:26.263542Z 6 [Note] Access denied for user 'root'@'localhost' (using password: NO)

Figure 2 presents the CPU and Real Memory usage of the event tracing process during its idle time. As it can be seen the event trace does not require any intermediate polling during downtime of an event. This can be effective for resource constrained devices such as emerging and embedded technologies.

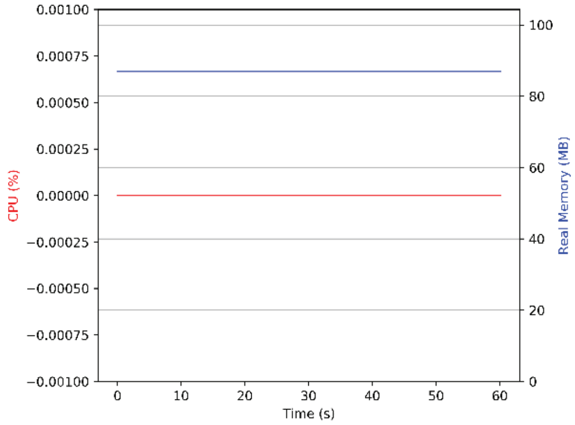


Figure 2: Event Tracing CPU and Real Memory usage over a 60 second period during idle time

Figure 3 presents the CPU and Real Memory usage of the event tracing process during a path traversal attack carried out on the Kali machine. The event tracing application provides real time polling of all traces of Apache2 errors that have been captured. A comparison of Figure 2 and Figure 3 can be used as an indicator of intrusive behaviour that can be linked to path traversal due to the increase in Apache2 requests to pages that relate to 404 errors. A spike occurred which could be due to an increase in requests being made. It is hypothesised that this application could be improved further with software engineering optimisation. The limitations of this finding are that there is no realistic traffic being induced into the web server.

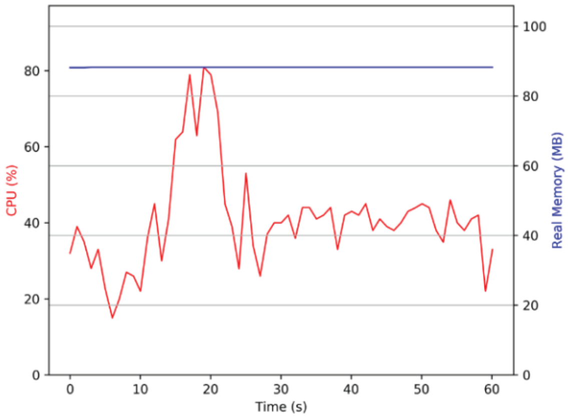


Figure 3: Event Tracing CPU and Real Memory usage over a 60 second period during a web directory traversal attack

4. Discussion

The results demonstrate three findings. Firstly, event tracing shows capability within the Linux domain to provide detectors of potential cyber-attacks. Secondly, event tracing can attach to applications without causing further development or downtime. Finally, employing an event trace in addition to log files could help analysts with a greater range of understanding of potential cyber-attacks, for example, MySQL queries, however, this must be caveated with the storage location being classified appropriately due to potential sensitive queries. Additionally, event tracing demonstrates an improvement on the confidentiality, and integrity of forensic data. Confidentiality is provided to an event trace through the delivery of the result directly to the SIEM, removing the need for a file to be created. The integrity of an event trace once generated cannot be manipulated as no file artefact is generated, this is analogues of a network packet being sent. With the accessibility of event tracing within the Linux domain, and the growing research into the uses within the field of Intrusion Detection. This could help security analysts monitor bespoke applications that provide no forensic data within the emerging and embedded technologies.

5. Conclusion

In conclusion, two aims have been achieved from this study. Firstly, a review of event tracing via eBPF within the Linux domain has provided key insights, namely, the ability to provide further context to cyber-attacks while improving on the confidentiality and integrity of forensic data. Secondly, the potential implementation of the technique within the context of security monitoring could be used within the field of emerging and embedded technologies. However, further experimentation with event tracing implemented on emerging and embedded technologies would target intrusive behaviour that is common within these technologies, e.g., remote login (Antonakakis *et al.*, 2017). Providing indicators of attack and compromise within this field could help advance the security community in monitoring very ad-hoc security practices that are common within such devices (Barcena, Wueest and Metz, 2015).

References

Antonakakis, M. *et al.* (2017) 'Understanding the Mirai Botnet', in *Proceedings of the 26th USENIX Security Symposium*, pp. 1093–1110. Available at: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis.

Barcena, M. B., Wueest, C. and Metz, R. (2015) *Insecurity in the Internet of Things*, *MIT Technology Review*. 1. Available at: https://www.symantec.com/content/dam/symantec/docs/white-papers/insecurity-in-the-internet-of-things-en.pdf.

Hutchins, E. M., Cloppert, M. J. and Amin, R. M. (2011) 'Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains', *6th Annual International Conference on Information Warfare and Security*, (July 2005), pp. 1–14. Available at: http://papers.rohanamin.com/wp-content/uploads/papers.rohanamin.com/2011/08/iciw2011.pdf%5Cnhttp://www.lockheedmartin.com/content/dam/lockheed/data/corporate/documents/LM-White-Paper-Intel-Driven-Defense.pdf.

Janos, F. D. and Dai, N. H. P. (2018) 'Security concerns towards security operations centers', in *SACI 2018 - IEEE 12th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*. IEEE, pp. 273–278. doi: 10.1109/SACI.2018.8440963.

Microsoft (2017) *Hidden Treasure: Intrusion Detection with ETW*. Available at: https://docs.microsoft.com/en-us/archive/blogs/office365security/hidden-treasure-intrusion-detection-with-etw-part-1 (Accessed: 9 January 2020).

Microsoft (2019) *Event Tracing for Windows*. Available at: https://docs.microsoft.com/en-us/windows/desktop/ETW/about-event-tracing (Accessed: 10 May 2019).

Onwubiko, C. (2018) 'CoCoa: An ontology for cybersecurity operations centre analysis process', in *2018 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, CyberSA 2018*. IEEE, pp. 1–8. doi: 10.1109/CyberSA.2018.8551486.

Scalegrid (2019) *2019 Database Trends – SQL vs. NoSQL, Top Databases, Single vs. Multiple Database Use*. Available at: https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/ (Accessed: 9 January 2020).

W3Techs (2020a) *Usage of Operating Systems for Websites*. Available at: https://w3techs.com/technologies/overview/operating_system (Accessed: 9 January 2020).

W3Techs (2020b) *Usage of Web Servers*. Available at: https://w3techs.com/technologies/overview/web_server (Accessed: 9 January 2020).

W3Techs (2020c) *Usage statistics of PHP for websites*. Available at: https://w3techs.com/technologies/details/pl-php (Accessed: 9 January 2020).

Masters Research Papers

Enabling Rapid Response and Service Restoration with Machine Learning Techniques

Andre Lopes and Andrew Hutchison
University of Cape Town, Rondebosch, South Africa
LPSAND004@myuct.ac.za
Andrew.Hutchison@t-systems.com

DOI: 10.34190/EWS.20.108

Abstract: The speed with which incidents can be detected has a major impact on the extent of associated damage. Similarly, the time within which incidents can be resolved, and services restored to active state, can minimize the impact on business or transactional activities facilitated by the service under attack. The focus of this work is to see to what extent machine learning can improve rapid response and service restoration in attack situations. The hypothesis is that machine assisted attack response and resolution can minimise non-availability. By applying different scenarios and machine learning approaches, preliminary results have been achieved which show the effectiveness of machine learning within an attack minimisation strategy. The paper describes the scenarios considered and provides results of different machine learning approaches.

Keywords: Machine learning, Automated intrusion response

1. Introduction

The complexity of Intrusion Response Systems (IRSs) has increased to compete with the growing sophistication of intrusions. This has resulted in sophisticated systems which combine varying components to create more effective and robust responses to intrusions. This sophistication can however introduce new unforeseen problems if not managed appropriately.

A particular example of these unforeseen problems is the *deactivation issue*, whereby response systems automatically attempt to deactivate a service under attack so that the problem can be resolved, and the system then returned to its normal operating behaviour. IRSs that incorporate such behaviour could become susceptible to manipulation from attackers, as this component could be used to continuously activate and deactivate a response, causing a degradation in service performance by means of the IRS itself if it is not accurate in detection / response conditions and circumstances.

This work serves to investigate an IRS system where the response behaviour is determined by the severity of the attack. In such a system, sophisticated responses are applied as necessary -- reducing the risk associated with unseen vulnerabilities. As such, an IRS is created with multiple responses, each with a unique purpose. This work will assess the applicability of responses which range from only *notifying* network administrators about an attack, to responses which are *fully automated*, and will activate and deactivate services where necessary.

The research questions of this work are as follows:

- Can an Intrusion Response System (IRS) be created such that it will respond appropriately according to attack type?
- Can an Intrusion Response System (IRS) be created such that it utilises machine learning to determine when to activate and deactivate network responses?

2. Background

2.1 Intrusion response

IRSs have become increasingly complex with various components fulfilling different objectives to provide for more security. The complexity of IRSs can be categorised according to their level of automation, with *notification systems* only providing alerts about an intrusion, to *manual response systems* providing potential responses, to *automated response systems* which act independently of network administrators. This section, with the aid of Figure 1, will explain the further distinctions of automated responses systems and will illustrate the complexity of IRSs.

Components of automated intrusion response systems

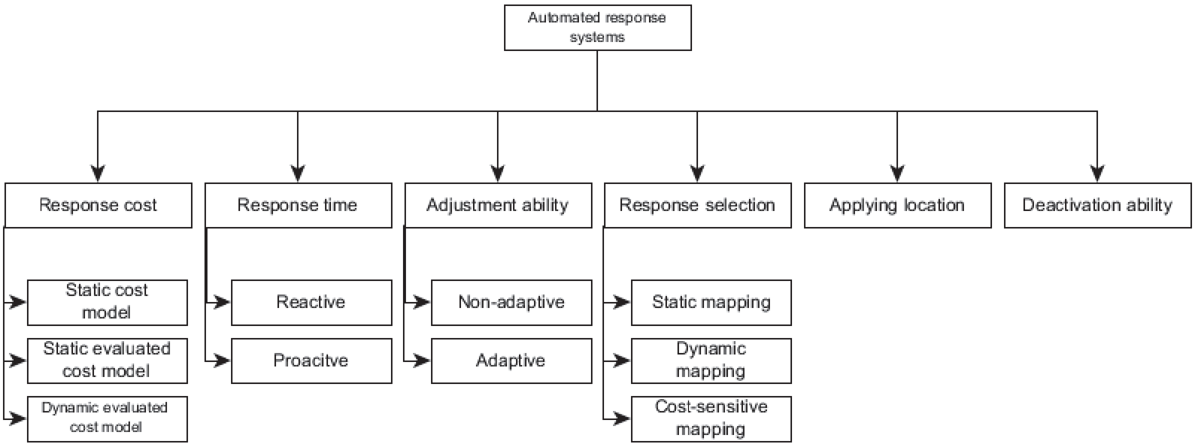


Figure 1: Our assessment of an IRS taxonomy

2.1.1 Adjustment ability

The ability to learn from previous responses, and how effective they were, determines whether a response system is able to adjust its behaviour or not. A response system can either be non-adaptive or adaptive (Inayat, et al., 2016) (Shameli-Sendi, et al., 2012) (Stakhanova, et al., 2007) (Shameli-Sendi, et al., 2014). *Non-adaptive* systems will always respond in the same way for the remainder of their execution. This does not change unless an administrator manually intervenes. *Adaptive* systems will respond differently throughout the duration of the system's execution.

2.1.2 Response selection

An automated IRS needs a way of responding to an attack. It can do this either statically, dynamically or through a cost-sensitive measure (Inayat, et al., 2016) (Shameli-Sendi, et al., 2012) (Stakhanova, et al., 2007) (Shameli-Sendi, et al., 2014). *Static mapping* determines a response based on the type of attack. *Dynamic mapping* responds to an attack by taking into account the overall state of the network. *Cost-sensitive mapping* is a unique approach, as it tries to balance the damage caused by attackers, with the damage caused by the response system, since both can lead to a disrupted network.

2.1.3 Response time

A response to an attack can occur either before or after an attack has happened. Intrusion response systems can be classified into reactive or proactive, depending on their approach (Inayat, et al., 2016) (Shameli-Sendi, et al., 2012) (Shameli-Sendi, et al., 2014). *Reactive* systems respond to an attack after it has occurred while *proactive* systems make predictions about attacks.

2.1.4 Response cost

All responses to attacks have the ability to disrupt a system. It is possible to associate a response cost to these responses, which can then be used by other systems (such as in section 2.1.2). Response cost can be calculated using a static cost model, static evaluated cost model or a dynamic evaluated cost model (Shameli-Sendi, et al., 2014). *Static cost model* systems use expert knowledge to assign costs to each response. *Static evaluated cost model* systems use an evaluation mechanism to assign cost to responses. *Dynamic evaluated cost model* systems evaluate cost based on the state of the network.

2.1.5 Applying location

Intruders in a network will need to have an attack path associated with their intrusion. The attack path determines how the intruder managed to reach the target machine on the network. Attack paths consist of four points: the start point (or intruder's machine), the firewall point, the midpoint (associated with intermediary machines that are exploited), and the end point (the target) (Shameli-Sendi, et al., 2014).

2.1.6 Deactivation ability

It is often the case that a response will only need to be applied for a temporary period of time, after which, the response should be deactivated, and the system restored to its normal operational order. Deactivation ability is not a common feature included in most IRSs, since it is a complex problem to decide when and how to deactivate a response (Shameli-Sendi, et al., 2014) (Kanoun, et al., 2010).

2.2 Reinforcement learning algorithms

It is not always possible to create an algorithm that performs well in any given problem (Wolpert, 1995), which is why many reinforcement learning algorithms adapt to the problem by learning a policy. Examples of such algorithms include: Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO), which have both proven to perform well on a variety of problems, irrespective of the problem's complexity. Unfortunately, not all reinforcement learning algorithms have managed to obtain the same adaptability that is achieved with TRPO and PPO, and will instead perform, for example, very well on complex problems, but poorly on simple problems, or, they will perform well on problems with discrete action spaces, but not on problems with continuous action spaces (where actions chosen are not limited) (Van Hasselt, 2016), (Schulman, et al., 2015), (Schulman, et al., 2017).

TRPO is an algorithm designed to be scalable, allowing it to perform well with a large amount of parameters, and reliable, allowing it to be useful on a number of problems without hyper-parameter optimization (Schulman, et al., 2015). *PPO* expands the capabilities of TRPO by retaining its scalability and reliability, but also improves by reducing algorithm complexity and increasing compatibility with other reinforcement learning architectures (Schulman, et al., 2017).

3. Design

In this section the design of an intrusion response system in the context of network protection is considered. The structure of the IRS (section 3.2), along with surrounding components, such as the IDS (section 3.1) and the test network (section 3.3), are introduced in this section. The explanation will follow in order of the events that are expected to occur, with an IDS alert triggering first; the IRS responding; and the network reacting dynamically to the responses controlled by the IRS. This procedure is further demonstrated in
Figure 2.

3.1 Intrusion detection system

The Intrusion Detection System (IDS) used is a simulated component which for test purposes triggers an alert either upon manual instruction, or in intervals according to a specific distribution. As a consequence of this, no network attacks are performed, nor is live network traffic analysed for attacks. The function of this component is to allow for controlled testing of the IRS, and it makes use of three alert types to do so:

- The DoS alert type, when executed, will continuously alert the IRS about dubious network traffic. It is assumed that during a DoS attack, multiple alerts will be sent to the IRS. It is also further assumed that multiple attacks will occur, all of varying intensity. The intensity of each attack will follow a distribution, which the IRS can learn in order to mitigate future attacks. The intention of this alert type is to test fully automated responses.
- The firmware altered alert is triggered upon manual intervention and provides an alert to the IRS about the firmware of a device on the network being unexpectedly changed. The intention of this alert is to show the applicability of no automation, since it is assumed that if a periodic scan discovers altered firmware on a device, it will require further investigation. For scenarios requiring digital forensics, it is arguably preferable that no automation takes place, so as to preserve the circumstances and configuration of the attack.
- The virus alert type also triggers upon manual intervention and provides a single alert to the IRS about the potential of a virus being transferred to a computer on the network. The intention of this alert type is to test limited automation, since it is assumed that while a quick response to a virus will result in less damage to network integrity and confidentiality, a fully automated response with learning, such as which program to run during a virus attack, is not preferable, as these responses could allow the network or computer to undergo irreversible damage if, for instance, the program selected is incorrect.

Test-bed structure

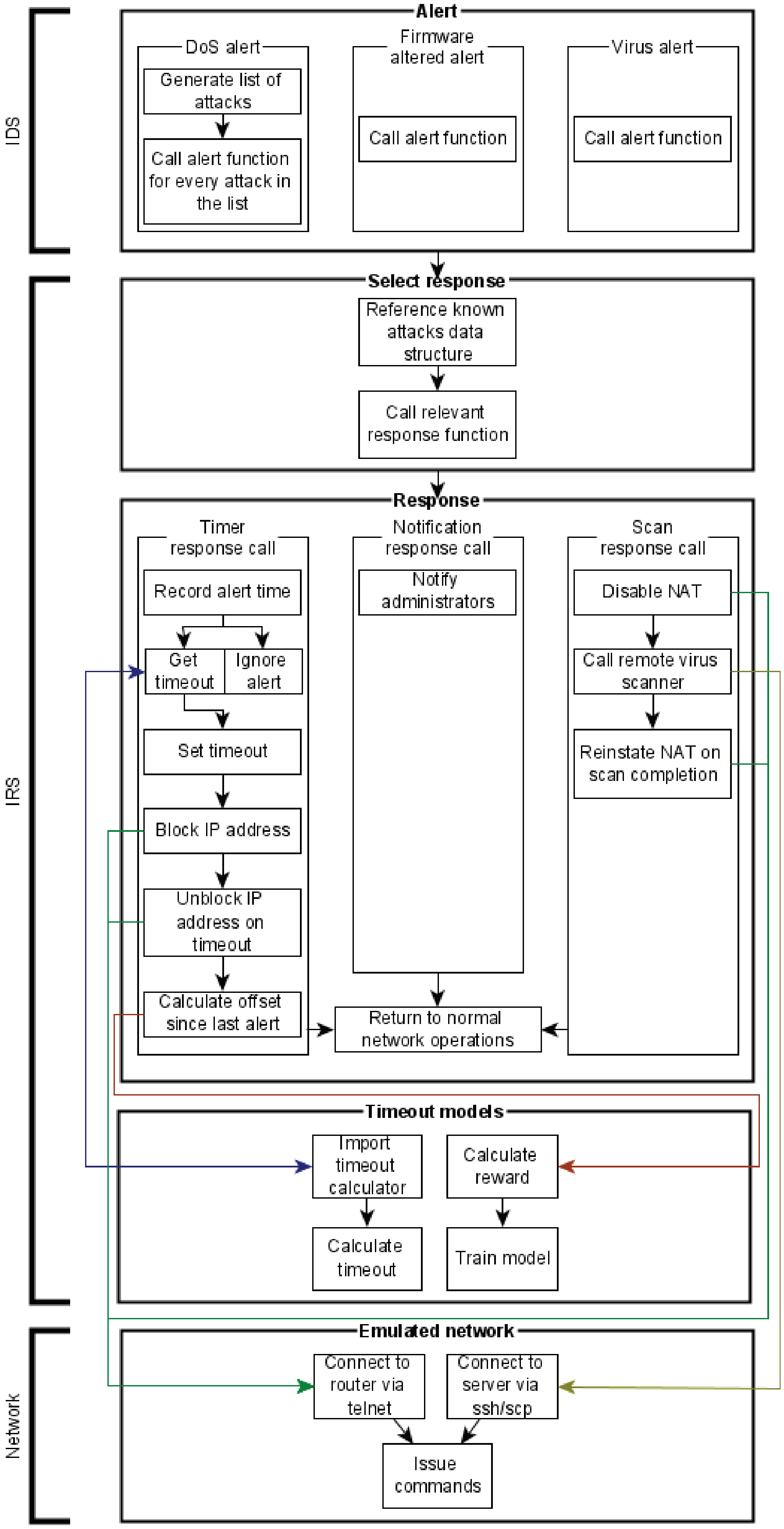

IDS
IRS
Network
Alert
DoS alert
Generate list of attacks
Call alert function for every attack in the list
Firmware altered alert
Call alert function
Virus alert
Call alert function
Select response
Reference known attacks data structure
Call relevant response function
Response
Timer response call
Record alert time
Get timeout
Ignore alert
Set timeout
Block IP address
Unblock IP address on timeout
Calculate offset since last alert
Notification response call
Notify administrators
Return to normal network operations
Scan response call
Disable NAT
Call remote virus scanner
Reinstate NAT on scan completion
Timeout models
Import timeout calculator
Calculate timeout
Calculate reward
Train model
Emulated network
Connect to router via telnet
Connect to server via ssh/scp
Issue commands

Figure 2: Components and procedure of the test-bed

3.2 Intrusion response system

The Intrusion Response System (IRS) created for this work was designed to respond appropriately according to the severity of the attack. As such, this section will present varying responses that aim to counteract the example attacks given in the IDS section (section 3.1). The IRS will do this by selecting an appropriate response, enacting the response, and using external components, such as the timeout models, for furthering the response's effectiveness if necessary.

3.2.1 Select response

The select response component serves as the first component of the IRS and is responsible for receiving alerts and choosing the appropriate responses. Each alert received contains a parameter indicating the type of attack that was identified. Response selection is performed statically (see static mapping in section 2.1.2), according to the parameter in the alert. The DoS, firmware altered and virus alerts (see section 3.1) map to the timer, notification and scan responses (see section 3.2.2), respectively. The following data structure represents this information:

```
knownAttacks={
  `DoS' : `timer_response'
  `firmware altered' : `notification_response'
  `virus' : `scan_response'
}
```

It is possible to map a response to multiple attack types, as shown below, however, for the purposes of this experiment, each response will remain mapped to one attack type, as shown above.

```
knownAttacks={
  `virus' : `scan_response'
  `worm' : `scan_response'
}
```

Once a response is selected, the appropriate response function is called.

3.2.2 Response

The response component serves as the second and main component of the IRS, and is responsible for enacting the response chosen in the select response component (see section 3.2.1). Responses can be enacted in one of three approaches: timer, notification or scan, each of which can be seen in Figure **2**. It should be noted that some of these approaches interact with the emulated network that contains routers, a client and a server (see section 3.3):

- The *timer* response, when enacted, will prevent communication between the server in the network and a user outside the network by blocking the user's IP address. This response attempts to minimize server downtime by making sure that the server is online when no attacks are present, while still keeping it offline when attacks are present. The intention of this response is to test fully automated responses, which is why it is enacted when a DoS alert is detected, as it is assumed that damage from such an attack won't result in irreversible damage to the network. The response operates by recording the times of each alert received, determining if a response is already active, and if not, will make use of a reinforcement learning algorithm, that has been trained on past attacks, to estimate the downtime required by the server. The server will then be isolated for that time and when brought back online, will be compared to recent attacks for training purposes. If the estimated downtime was too short and enabled attacks to reach the network, the reinforcement learning algorithm will be trained to increase its future estimates, and likewise will decrease future estimates if the downtime was too long.
- The *notification* response simply outputs a message to the console, or network administrator, about an attack. The intention of this response is to test no automation, and is used when a firmware altered alert is received. Though it is possible to automate some aspects of this response, such as permanently blocking communications on the affected device until a network administrator can unblock it, it is assumed that such automation could be harmful when digital forensics is required. This response serves

to complete the IRS by providing it with the option to not interact with the network, which can be beneficial in preventing undue harm to the network through its own responses.

- The scan response automatically isolates the server in order to run a remote anti-virus tool. This response tests limited automation, where a specific program is automatically executed, and is used when a virus alert is received. This response operates by disabling the Network Address Translation (NAT) settings on the gateway router (router R1) that allows the server to communicate with external IP addresses. The connection between the IRS and the server, however, remains unchanged, thus allowing the IRS to communicate with the server. Once the server is isolated from external communication, the IRS copies an anti-virus tool to the server, and remotely commands it to perform a scan. For the purposes of this experiment, the anti-virus tool used is lynis. The IRS is able to wait until the tool is finished, after which, it will reinstate the NAT settings on the router.

3.2.3 Timeout models

The timeout models component is an addition to the IRS for responses that require further complexity than interacting with the network. With the given example, this component provides the IRS with the reinforcement learning capabilities required by the timer response (see section 3.2.2), and determines the length of time that communications are disabled for. Through the use of three algorithms, this component is able to learn from past attacks. The three algorithms implemented are: Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO) (see section 2.2) and a random algorithm for comparison purposes.

Default parameters for both of the algorithms were used in all cases, with exception of the learning rate for TRPO and the actions exploration parameter for both algorithms. The library used to implement each of the algorithms was tensorforce (Kuhnle, et al., 2017), and an example of the parameters for the algorithms can be seen below:

```
“Type”: “ppo_agent”
“learning_rate”: 1e-3
“discount”: 0.99
“actions_exploration”: {
        “initial_epsilon”: 0.5
        “timesteps”: 100
}
```

3.3 Network

The IRS makes use of an emulated network to show its functionality. The emulated network contains 5 nodes, two of which are emulated Cisco c3745 layer 3 switches, and two of which are debian docker images that represent the client and server. The client in the network is used to ensure that responses are active by issuing ping requests, while the server exists to respond to them. Should responses be enacted, the ping requests will either be filtered or delayed / dropped, depending on the response used. The last node, called the IDS/IRS, is the host itself, and allows communication between the emulated network and the physical host. It is through this node that response commands (see section 3.2.2) are sent to, and executed on, the emulated network devices. The network within GNS3, a network modelling program, can be seen in Figure 3.

Test-bed network

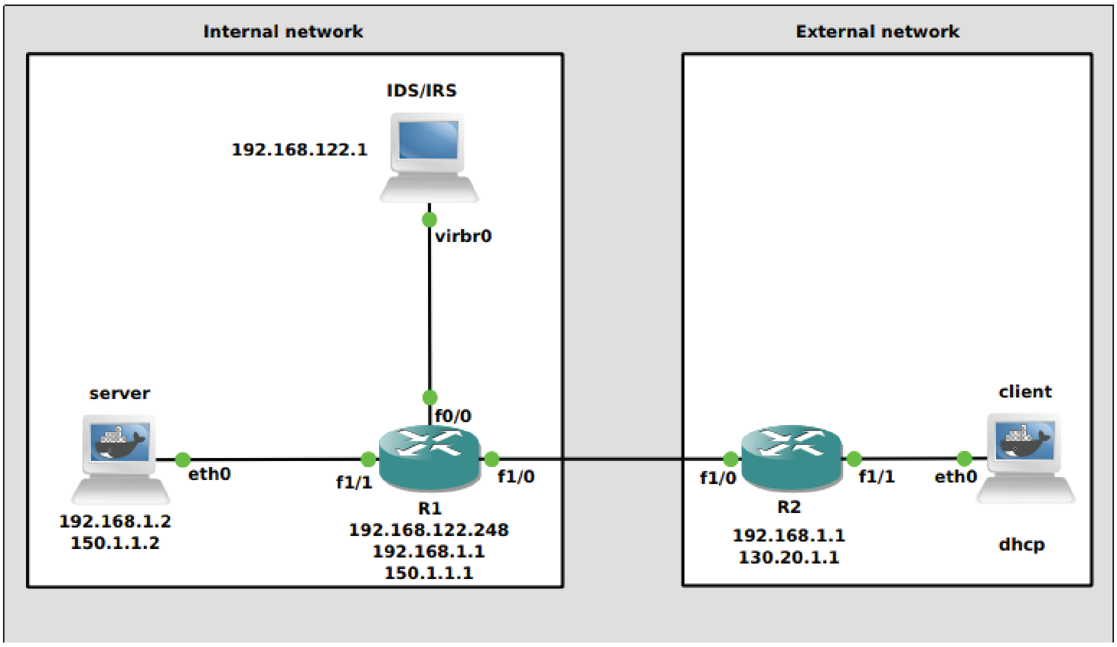


Figure 3: Diagram of network where testing was performed on

4. Testing

Testing was performed for each type of response identified in the previous section. Each response was designed to counter a specific attack. This section will comment on the effectiveness of each response while providing resulting log / performance information. Table 1 has been added to summarise the various responses according to the taxonomy seen in section 2.1.

Response summary

Table 1: Responses as applied to our taxonomy

Category/Feature	Level of automation	Response cost	Response time	Adjustment ability	Response selection	Applying location	Deactivation ability
Timer	Automated	None	Proactive	Adaptive	Static	Firewall	Supported
Scan	Automated	None	Reactive	Non-adaptive	Static	Firewall, Target	Supported
Notification	Manual	N/A	N/A	N/A	N/A	N/A	N/A

The first response considered is that of the notification response and was designed to respond against a firmware altered alert. It was intended to provide for scenarios where no automation is used, so that investigations could be performed. The following log entry provides the message outputted when the firmware altered alert is received. No other actions are performed.

```
Info: A hash mismatch has been detected on router R1
```

The next response considered is the scan response, which was designed to respond against virus alerts. It is intended to provide for scenarios of limited automation and will automatically isolate the affected device (the server in our test-bed) and will run anti-virus software (lynis). The following logs provide an account of the downtime experienced when this response is active. It can be seen that while the client attempts to ping the server, an interruption occurs, which increases the response time of the ping requests. The server the scan response is executed on is a minimal system, and as such, scans are performed quickly, enabling the packets to remain delayed at 568 and 630ms, as opposed to being dropped. Logs of the actual scan have not been provided, as no viruses were used since the IDS is simulated.

```
PING 150.1.1.2 (150.1.1.2) 56(84) bytes of data.
```

```
64 bytes from 150.1.1.2: icmp_seq=1 ttl=62 time=30.2 ms
64 bytes from 150.1.1.2: icmp_seq=2 ttl=62 time=29.1 ms
64 bytes from 150.1.1.2: icmp_seq=3 ttl=62 time=23.6 ms
64 bytes from 150.1.1.2: icmp_seq=4 ttl=62 time=27.4 ms
64 bytes from 150.1.1.2: icmp_seq=5 ttl=62 time=21.4 ms
64 bytes from 150.1.1.2: icmp_seq=6 ttl=62 time=25.8 ms
64 bytes from 150.1.1.2: icmp_seq=7 ttl=62 time=29.6 ms
64 bytes from 150.1.1.2: icmp_seq=8 ttl=62 time=23.9 ms
64 bytes from 150.1.1.2: icmp_seq=9 ttl=62 time=568 ms
64 bytes from 150.1.1.2: icmp_seq=10 ttl=62 time=630 ms
64 bytes from 150.1.1.2: icmp_seq=11 ttl=62 time=45.4 ms
```

The last response considered is the timer response which was designed to respond against DoS attacks. This response type was intended to provide for scenarios of full automation. As a result, this response will automatically interrupt communications between the server and the client, determine how long the interruption should last in seconds through the use of machine learning algorithms, and restore communication afterwards. Table 2 provides the results for each of the algorithms used while they learned at different numbers of attack instances (100, 500, 1000, 2000). Each intersection of algorithm and instance reflects the average inaccuracy of the guesses made by the algorithm as to when to restore communications. As indicated by the table, PPO obtains the lowest inaccuracy of 9 seconds, after only learning from 1000 attack instances, suggesting that it is the most effective algorithm to use for trying to reduce server downtime while mitigating exposure to attacks. Further details of the results obtained from this response can be seen in our previous works (Lopes & Hutchison, 2019).

Timer response results

Table 2: Learning rates at various attack instances

Algorithm/Instances	100	500	1000	2000
PPO	30	24	9	9
TRPO	29	28	27	21
Random	30	30	29	31

5. Conclusion

In this work, we have created multiple approaches to dealing with incidents on a network. These approaches are applied in one of the following ways: no automation (notification response), limited automation (scan response) and full automation (timer response). Table 1 summarizes these responses according to the IRS taxonomy (Figure 1) seen in the intrusion response section (2.1). Two of the methods (the scan and timer response) automatically deactivate themselves after repairing the network, thus optimizing service availability. In fact, on minimal systems like the one tested, these responses don't disrupt network traffic, and merely delay the packets sent, as seen with the scan response in the Testing section (5). However, for more complicated responses that require adjustments to their parameters to optimize service availability, machine learning is used. This can be seen with the timer response in the Testing section (5). Two algorithms were tested against a random algorithm, and it can be seen that PPO improves from a 30 second guessing inaccuracy to a 9 second inaccuracy, proving that the algorithm not only learnt, but improved service availability by 21 seconds per attack on average. Though the performance is promising, a 9 second accuracy is the best this algorithm achieves as it does not improve after learning from another 1000 attack instances. Therefore, while the response works, improvements could still be made to increase its performance.

References

Cohen, M. I., 2008. PyFlag--An advanced network forensic framework. *Digital investigation,* Volume 5, pp. S112--S120.

Garfinkel, S. & Cox, D., 2008. Finding and archiving the internet footprint. *The British Library.*

Garfinkel, S. L., 2010. Digital forensics research: The next 10 years. *digital investigation,* Volume 7, pp. S64--S73.

Huber, M. et al., 2011. *Social snapshots: Digital forensics for online social networks.*, ACM, pp. 113-122.

Inayat, Z. et al., 2016. Intrusion response systems: Foundations, design, and challenges. *Journal of Network and Computer Applications,* Volume 62, pp. 53-74.

Kanoun, W., Cuppens-Boulahia, N., Cuppens, F. & Dubus, S., 2010. *Risk-aware framework for activating and deactivating policy-based response.*, IEEE, pp. 207-215.

Kuhnle, A., Schaarschmidt, M. & Fricke, K., 2017. *Tensorforce: a TensorFlow library for applied reinforcement learning.* [Online]
Available at: https://github.com/tensorforce/tensorforce

Lopes, A. & Hutchison, A., 2019. *Experimenting with Machine Learning in Automated Intrusion Response.* Saint Petersburg, Intelligent Distributed Computing XIII, pp. 505-514.

Nance, K., Hay, B. & Bishop, M., 2009. *Digital forensics: defining a research agenda.* Hawaii, IEEE, pp. 1-6.

Schulman, J. et al., 2015. *Trust Region Policy Optimization*., s.n., pp. 1889-1897.

Schulman, J. et al., 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347.*

Shameli-Sendi, A., Cheriet, M. & Hamou-Lhadj, A., 2014. Taxonomy of intrusion risk assessment and response system. *Computers & Security,* Volume 45, pp. 1-16.

Shameli-Sendi, A., Ezzati-Jivan, N., Jabbarifar, M. & Dagenais, M., 2012. Intrusion response systems: survey and taxonomy. *Int. J. Comput. Sci. Netw. Secur,* Volume 12, pp. 1-14.

Stakhanova, N., Basu, S. & Wong, J., 2007. A taxonomy of intrusion response systems. *International Journal of Information and Computer Security,* Volume 1, pp. 169-184.

Van Hasselt, H., Guez, A. & Silver, D., 2016. Deep reinforcement learning with double q-learning. *Thirtieth AAAI Conference on Artificial Intelligence.*

Wolpert, D. H., Macready, W. G. & others, 1995. *No free lunch theorems for search,* Santa Fe: Technical Report SFI-TR-95-02-010.

Using Ethereum Smart Contracts for Botnet Command and Control

Alexandre Oliveira, Vinícius Gonçalves and Geraldo Rocha Filho
University of Brasília, Brasília, Brazil
alexandrevvo@gmail.com
vpgvinicius@unb.br
geraldof@unb.br

DOI: 10.34190/EWS.20.065

Abstract: Since the Bitcoin emergence in 2008, the blockchain capabilities have been drawing attention in several areas, including, most recently, cybersecurity applications. At the same time, botnets are since the 1990s, a concerning source of cyber incidents, ranging from financial losses to national security threats. Regarding botnets, the blockchain technology has been studied with both defensive and offensive approaches. On the offensive strategies, the use of blockchain as command and control (C&C or C2) for botnets is a promising field. In this way, while most proposals use the Bitcoin implementation, this paper demonstrates the use of Ethereum's smart contract feature to act as C&C. The smart contract gives the flexibility of its programming language, leveraging the blockchain features to implement a decentralized botnet communication infrastructure. Therefore, while the smart contract is distributed in all blockchain nodes, and provides restrictions to its interactions, it enables a resilient and reliable C&C for the botnet. As the C&C is the most critical node of a botnet, keeping it protected is a major concern of botmasters, and its takedown is the main objective for the defensive actors. Hence, the proof of concept held in this paper contributes to the cybersecurity community, helping to understand and anticipate cybercriminals' tactics in a way to improve the defense capabilities of security agents.

Keywords: blockchain, botnet, ethereum, smart contract, command and control

1. Introduction

Even though the blockchain technology is not exactly new (Haber and Stornetta, 1991), it was with the emergence of Bitcoin in 2008 (Nakamoto, 2008) that it became widely known. Since then, the peculiarities of this technology have been attracting the attention of many business and research areas, such as banking (Finck, 2018), logistics (Tijan *et al.*, 2019), and healthcare (Zhang *et al.*, 2018). Regarding the cybersecurity field, the studies focusing on the blockchain are even more recent, with papers published only around 2015 (Taylor *et al.*, 2019). A more specific area in cybersecurity, where researchers have already identified the promising use of blockchain, is the botnet universe.

The blockchain and botnets interaction has been studied in a variety of scenarios, mainly over the Bitcoin implementation. Many studies aim at using the blockchain capabilities for the prevention and defense against the botnets (Sagirlar, Carminati and Ferrari, 2018; Falco *et al.*, 2019; Shafi and Basit, 2019; Zarpelão, Miani and Rajarajan, 2019). Another field of study is the offensive use of blockchain, focusing on how the technology properties can be leveraged to empower the botnets (Roffel and Garrett, 2014; Ali *et al.*, 2015, 2018; Frkat, Annessi and Zseby, 2018). This paper describes an offensive strategy, proposing the use of Ethereum's smart contract feature as a command and control (C&C or C2) to deliver high resilience to the botnet.

1.1 Blockchain and Smart Contracts

A blockchain consists of a distributed ledger, composed of a chain of blocks where transactions are registered. Each block is generated based on the previous one in the chain. This structure provides the immutability feature to the technology, along with the mechanism of validating data to be inserted in the blocks. Those main characteristics of being distributed, immutable, and other cryptographic features, which gives it the authenticity trait, make the blockchain very attractive to a variety of applications (Nofer *et al.*, 2017).

The blockchain can be used both in public or private modes. The public method is the known Bitcoin and Ethereum's chains in which people and companies exchange cryptocurrency around the world. On the other hand, the private implementation consists of using the blockchain platform, restricting access to specific nodes. It can be done in a local area network or through the internet.

Smart contracts were first theorized by Nick Szabo (Szabo, 1997) in the late 1990s. The Bitcoin protocol implemented it in a limited way but was in the Ethereum platform that the concept was actively developed (Buterin, 2014). The smart contract is a trustless and automated agreement which have predefined behaviours given specific conditions (Bashir, 2017; Harris, 2019). In the Ethereum environment, it is a piece of code, usually written in a programming language called solidity, which runs over the Ethereum Virtual Machine (EVM). As running over a blockchain structure, it brings with it all the decentralized, secure, and immutable characteristics.

1.2 Botnet

Botnets are networks of infected devices, commonly called bots or zombies, which are remotely controlled by a botmaster. The central point of communication, from where the botmaster sends instructions to the bots, is the command and control. Therefore, the botmaster uses the C2 to send messages to the bots to execute orchestrated actions on a large scale. Since the C&C is the central communication node with all bots, the maintenance of this point is vital to the existence of the botnet.

Initially used for vandalism and demonstration of hacking skills, this kind of network became an unsettling tool used by the cybercriminals. Distributed Denial of Service, extortion, phishing campaigns, and most recently, cryptocurrency mining, are currently the primary use of the botnets (Barford and Yegneswaran, 2007; Ali *et al.*, 2015). Recent records indicate that 500 million computers are infected annually, causing global losses of approximately $110 billion (Ali *et al.*, 2018).

Initially, since the 1990 decade, the primary mechanism used for botnet communication was the Internet Relay Chat (IRC) protocol. The IRC is a client/server protocol used for exchange text messages in an environment where this could be done in chat channels, allowing the sent of messages in one-to-one or one-to-many standard. However, the extensive use of this communication method forced the security professionals to develop filters and other mechanisms to block the botnet controlling through IRC (Karim *et al.*, 2014). The *PrettyPark*, emerged in 1999, was one of the first botnets using this method of communication (Canavan, 2005).

Therefore, the attackers promptly developed alternatives to exchange instructions between bots and the botmaster, without being detected. The strategies evolved in several different ways. From the protocols HTTP (Nazario, 2007) and SMB (Symantec Security Response, 2015), to the use of social networks (Nagaraja *et al.*, 2011; Compagno *et al.*, 2015), the attackers keep enhancing their bots communication.

In addition to the variety of protocols used, mechanisms to avoid the botnet takedown were developed or leveraged, such as fast-flux, with the assignment of multiple IP address to a domain, or the Domain Generation Algorithm (DGA) (Vormayr, Zseby and Fabini, 2017).

2. Related work

The most practical approaches to the utilization of blockchain as a command and control for botnets were developed in (Ali *et al.*, 2015, 2018; Frkat, Annessi and Zseby, 2018), all those focusing on the Bitcoin platform. The Zombiecoin project (Ali *et al.*, 2015, 2018) discusses some possible strategies to send instructions through Bitcoin transactions and develop its proof of concept using the OP_RETURN field. The OP_RETURN is an output script function, intended for metadata, implemented in the transaction data message, and with the capacity of 80 bytes of data. In this work, the authors discuss the use of ECDSA key leakage, and subliminal channels to send instructions inside bitcoin transactions. In (Frkat, Annessi and Zseby, 2018), the authors also based the communication on a key leakage method and the creation of a subliminal channel to insert C&C messages in bitcoin transactions. One advantage of the method is that it does not require any specific blockchain field and relies on the digital signature scheme, which is common to all blockchains.

In both papers cited above, the communication mechanisms depend on bots continuously monitoring all blockchain transactions for botmaster activity. After they identify a transaction from the known botmaster address, the bot will analyze it and take the programed actions. This strategy can represent a substantial consumption of computing power and network activity, which can more easily expose the bot. In our proposal, there is no need for this kind of monitoring. The bots can be designed to interact with the contract in a specified time frame without risking to lose data.

The only paper where the Ethereum is studied as a botnet C&C tool is in (Baden *et al.*, 2019). Here, a theoretical approach is used to suggest the utilization of the Whisper communication protocol, part of the Ethereum platform, in order to transmit messages to the blockchain nodes, leveraging the anonymity capabilities of the protocol. This strategy has good potential. Nevertheless, the paper stays on the conceptual area and does not implement a proof of concept.

In this paper, we propose and demonstrate the use of Ethereum's smart contracts acting as C&C of a botnet. Assuming that all the bots are blockchain members, all of them can interact with the contract, which will be registered in the chain. Thanks to the Ethereum signature mechanisms, it is possible to limit these interactions in a way that only the botmaster will be able to set some variable contents in the contract. In the same way, the bots, as members of the blockchain, will be able to request data from the contract. Using this data to store instructions for the malware installed in a bot can be an effective way to command and control a botnet. Since the contract is deployed in the blockchain as a transaction, all nodes have a copy of it. Therefore, the smart contract, as the C&C, is distributed in all blockchain nodes, becoming virtually impossible to deactivate it, thus providing high resilience to the botnet.

3. Proposal

Our work demonstrates a novel strategy in the utilization of blockchain technology as C&C for botnets. Unlike the other strategies, this one was carried out with the Ethereum platform, leveraging the smart contract features. As will be demonstrated in this paper, this proposal gives high resilience and flexibility, offering significant advantages when compared to other C&C communication methods. Here follows a brief description of the method:

- We assume that the devices were already infected by a malware, which would enable an instance of the Ethereum blockchain.
- The botmaster, as a node of the blockchain, deploys the smart contract.
- After the inclusion of the smart contract in the blockchain, all nodes will have a copy of it.
- The bots, also as nodes of the blockchain, would be able to interact with the contract, calling functions to retrieve instructions that will be executed in the device.
- Only the contract owner, the botmaster, can interact with specific functions in the contract to set variables with new instructions to the bots.

As discussed previously, the blockchain can be implemented in both public or private mode. The public chain would interact with the regular Ethereum network around the world, spending "real" ether. Therefore, the private implementation can be ran composed only by our controlled nodes, with all mining parameters previously defined by us, and without worrying with the expense of cryptocurrency. For this work, we used the private model.

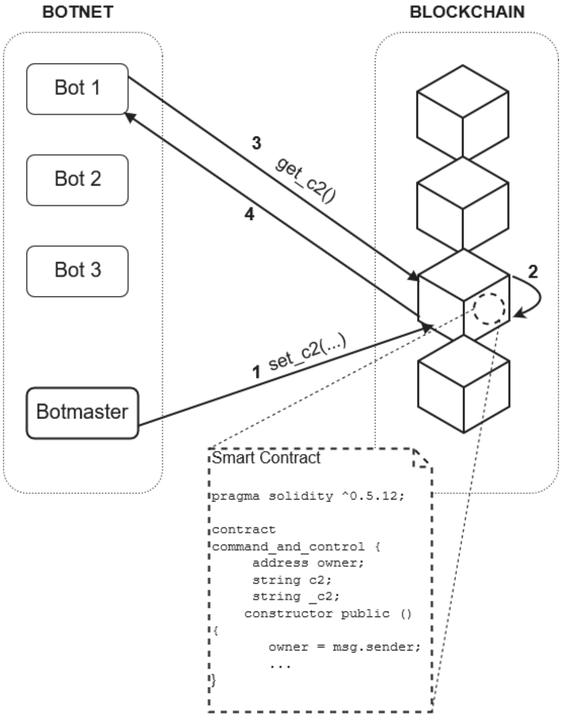
To deploy and interact with the blockchain, we used the truffle console and ganache, both part of the Truffle suite (Truffle, 2019). This tool enables a development environment, testing framework, and asset pipeline using the Ethereum virtual machine. In the first device, we deployed an instance of the blockchain using ganache and called a truffle console to interact with it. In another device, we used the truffle console to connect with the provider enabled by ganache. This way, we had two nodes of the same blockchain, in different devices. This scenario emulates a real situation where the nodes can be in any part of the world, communicating in a peer-to-peer standard. With this scenario set up, we can deploy the contract in one node and call its functions from the other one.

The contract was implemented in solidity with simple functions, aiming to proof the feasibility of the communication through it. The main functions are the *set_c2* and *get_c2*.

```
function set_c2 (string memory _c2) public {
   if (msg.sender != owner) return;
   c2 = _c2;
}
function get_c2() public view returns (string memory) {
   return c2;
}
```

In this example, the *c2* variable can store an internet domain where the bot should access in order to retrieve more data. On the other hand, the smart contract, due to its flexibility, could be built in a much more sophisticated way, sending instructions already designed to interact with the malware, which is infecting the bots. In any of those scenarios, the resilience to the C&C in guaranteed. Even storing an internet domain in the *c2* variable, in the case of a takedown of this domain, the bots would keep recurring to the contract looking for an update.

Figure 1 shows the sequence of basic communication. With the contract deployed by the botmaster, he can call the *set* function to define a new instruction to the bots. After this transaction has been validated, all nodes will have a copy of the new state of the contract. In this way, the bots can interact with the contract calling the *get* functions to retrieve information from the contract.



```
pragma solidity ^0.5.12;

contract
command_and_control {
     address owner;
     string c2;
     string _c2;
    constructor public ()
{
        owner = msg.sender;
        ...
}
```

Figure 1: Command and control communication

Another critical piece of the code is the definition of the contract owner. We defined the variable "owner" with the address of the message sender in the constructor of the contract. Assuming that the botmaster is deploying the contract, the "owner" variable will be set with his address. Therefore, this information can be used in the other functions, like the *set_c2*, to verify who is calling the function and allow some operations only to the botmaster.

```
constructor public () {
   address owner;
   ...
   owner = msg.sender;
}
```

This method is responsible for the authenticity in the interactions with the contract. Since Ethereum uses the ECDSA (Elliptic Curve Digital Signature Algorithm) to sign its transactions, all the interactions with the smart contract rely on the nodes' public and private keys. The node address, in turn, is created based on its public key (Bashir, 2017). This encryption system is essential for the resilience we are looking for in the control of the botnet through the smart contracts.

Hence, with this authentication method, even if a defender identifies all the contract functions and understand how the bots are interacting with it, he would not be able to change the *c2* variable through the *set_c2* function. This takeover would only be possible in the case of obtaining the botmaster's private key and deploying a node with the original botmaster's address.

3.1 Costs

Regarding the expenditure of cryptocurrency, it is essential to understand the differences between the public and private implementations. In the case of using a private blockchain, the attacker would not have the inconvenience of spending ether. On the other hand, the public method, despite the ether expense, brings the advantage of being part of a vast network where malicious activities would be better disguised.

In the case of a public blockchain, which would be the most resilient strategy, all transactions would involve Ether expense. Table 1 demonstrates how much Ether would be spent to store 1 kB of data in a variable of the contract, for example. The values were deducted from the Ethereum yellow paper (Wood, 2014) where the author specifies that 20,000 gas would be the fee for a SSTORE operation to save a 256 bit word. The Gas price is variable and depends on the network mining activity. At last, the USD value was based on the rate Ether/USD at the time of writing this paper.

Table 1: Costs

Transaction	Gas cost	Gas price	Ether expense	USD
Store 1kB	640k	3.1	0.001984	0.29

This way, 1 kB would cost 640,000 gas. During the development of this work, the gas price was around 3.1 Gwei. Since one Gwei corresponds to 1 x 10^(-9) ETH, 1kB would cost 0.001984 ETH, or $0.29 according to the considered Ethereum/USD rate.

Those values are illustrative, and a real operation would involve other costs and the variation of the data exchanged. In any case, the proposal of using the contract to store a domain, or domain list in text format, would use a small amount of data, probably less than 1 kB.

Furthermore, most of the communication will rely on the request information functions. Since those functions are "public view" kind, those calls do not change the blockchain nor the contract state; hence, it will not cost any ether.

3.2 Resilience

Given the blockchain distributed computing characteristics, the creation of a botnet where each infected machine is a blockchain node ensures high resilience to the entire structure. This happens because, as the smart contract acts as the C&C and is hosted in the blockchain, the usual strategy of analysts seeking to takedown the botnet by discovering the C2 address and deactivating it or confiscating web domains, becomes ineffective.

In order to stop the C&C, all the blockchain nodes should be deactivated. It is essential to understand, as shown in Figure 1, that the smart contract, once deployed, is stored in all blockchain nodes, enabling data exchange between bots and C2. In this way, the botmaster can update the contract variables, setting new instructions to the bots with the confidence that all nodes will have access to the new information.

Furthermore, the deactivation of one or more nodes does not imply in the botnet takedown. Even with the bots been compromised, the instructions sent to the smart contract remain available to all other nodes and cannot be changed unless by the botmaster.

Even not demonstrated here, other actions can be developed to enhance the obfuscation of the data exchanged through the contract. One alternative could be restricting the access even to the *get* functions and controlling the bots activation by the botmaster, in order to authorize in the contract their interaction with it.

3.3 Weakness and Countermeasures

Even providing a novel proposal to ensure high resilience to the botnet, this strategy has its weaknesses and points of attention, which both the offensive and defensive actors should be aware of.

3.3.1 51% attack

In the case of using a private blockchain for the C&C, a critical aspect of this implementation would be the reduced number of nodes. If an adversary succeeds in reversing the malware which deployed the blockchain in the victim's device, he could obtain the genesis block data. With this information, it would be possible to instantiate multiple nodes of the blockchain. In case the opponent has enough nodes to achieve 51% of the blockchain computing power, this would allow the "51% attack" (Lin and Liao, 2017). In this attack, the adversary has control of the blockchain and can modify transaction data or prevent the validation of new blocks.

The best option to mitigate this vulnerability would be using the public Ethereum blockchain. Even with the ether expenditure, this implementation would ensure the most resilient infrastructure for the botnet.

3.3.2 Anonymity

As the transactions in the blockchain remain clear to all nodes, the botmaster C&C transactions will expose its IP address. Hence, an intruder node could trace the botmaster activity.

Nevertheless, the use of conventional anonymization techniques such as proxies and VPNs should be enough to handle this point.

3.3.3 Botnet Intruder

If an adversary takes control of a bot, he would have access to the malware code and would be able to reverse it in order to decode the contract instructions. Even in the case of a private blockchain implementation, after taking control of the bot, the adversary could have access to the genesis file, network ID, and all necessary information to join the blockchain and interact with other nodes.

This vulnerability can represent an information leak to the adversary, but it would hardly cause the botnet takedown. Regardless, this can be an opportunity for defensive actors to search for other peers interacting with the contract, identify the botnet size, and act on the malicious instructions as soon as they are issued.

4. Conclusion

In this paper, we proposed a novel strategy for the command and control of botnets, leveraging the Ethereum smart contracts characteristics. The proof of concept held in this work was carried out by implementing a private Ethereum blockchain with only two nodes. One of the nodes served as the botmaster, deploying the contract, and the other node, through the blockchain, interacted with the contract retrieving instructions. The implementation has proved itself feasible and demonstrated the application of all the security, flexibility, decentralization, and resilience concepts we have assumed initially.

Thanks to the distributed characteristic of blockchain, our proposal becomes a promising infrastructure for hosting the most critical node of the botnet. While most papers that study this topic focuses on the Bitcoin implementation, we demonstrated the feasibility and advantages of using the Ethereum platform. The Ethereum's smart contract feature proved to be a powerful tool in providing the needs of a botnet C&C. Through the programming language flexibility and cryptographic features, it was possible to demonstrate how the smart contract can act as the command and control being distributed in all blockchain nodes.

The proposal can be implemented in both private or public blockchain model. Both scenarios have their benefits and drawbacks, with the private model being the most convenient one, and the public model being the most resilient strategy to the proposed objective.

Here we demonstrated a simple contract as a proof of concept of the proposal. This model can be escalated, and other measurements can be done over its performance. Furthermore, we discussed some weaknesses of the implementation, and further works can enhance the proposal to achieve a more robust infrastructure.

In any case, the proof of concept held in this paper contributes to the cybersecurity community attempting to foresee the possible strategies the attackers might use in the botnet controlling. Understanding and anticipating the cybercriminals' tactics are a way to improve the defense capabilities against the aforementioned severe impacts of botnets.

References

Ali, S. T. *et al.* (2015) *ZombieCoin: powering next-generation botnets with bitcoin*. UK. doi: 10.1007/978-3-662-48051-9_3.

Ali, S. T. *et al.* (2018) 'ZombieCoin 2.0: managing next-generation botnets using Bitcoin', *International Journal of Information Security*. Springer Berlin Heidelberg, 17(4), pp. 411–422. doi: 10.1007/s10207-017-0379-8.

Baden, M. *et al.* (2019) 'Whispering Botnet Command and Control Instructions', in *2019 Crypto Valley Conference on Blockchain Technology (CVCBT)*. IEEE, pp. 77–81. doi: 10.1109/CVCBT.2019.00014.

Barford, P. and Yegneswaran, V. (2007) 'An inside look at Botnets', *Malware Detection*, 27, pp. 171–191. doi: 10.1007/978-0-387-44599-1.

Bashir, I. (2017) *Mastering Blockchain*. Birmingham: Packt Publishing.

Buterin, V. (2014) 'Ethereum White Paper', *Etherum*, (January), pp. 1–36. Available at: http://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf.

Canavan, J. (2005) 'WHITE PAPER : SYMANTEC SECURITY RESPONSE The Evolution of Malicious IRC Bots'. Available at: https://www.symantec.com/avcenter/reference/the.evolution.of.malicious.irc.bots.pdf.

Compagno, A. *et al.* (2015) 'Boten ELISA: A novel approach for botnet C&C in Online Social Networks', in *2015 IEEE Conference on Communications and Network Security (CNS)*. IEEE, pp. 74–82. doi: 10.1109/CNS.2015.7346813.

Falco, G. *et al.* (2019) 'NeuroMesh: IoT Security Enabled by a Blockchain Powered Botnet Vaccine', in *Proceedings of the International Conference on Omni-Layer Intelligent Systems - COINS '19*. New York, New York, USA: ACM Press, pp. 1–6. doi: 10.1145/3312614.3312615.

Finck, M. (2018) 'Blockchain Technology - Application in Indian Banking Sector', *Blockchain Regulation and Governance in Europe*, 19(2), pp. 1–33. doi: 10.1017/9781108609708.001.

Frkat, D., Annessi, R. and Zseby, T. (2018) 'ChainChannels: Private Botnet Communication over Public Blockchains', *Proceedings - IEEE 2018 International Congress on Cybermatics: 2018 IEEE Conferences on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology, iThings/Gree*, pp. 1244–1252. doi: 10.1109/Cybermatics_2018.2018.00219.

Haber, S. and Stornetta, W. S. (1991) 'How to Time-Stamp a Digital Document', *Journal of Cryptology*, 3(2), pp. 99–111.

Harris, C. G. (2019) 'The Risks and Challenges of Implementing Ethereum Smart Contracts', in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, pp. 104–107. doi: 10.1109/BLOC.2019.8751493.

Karim, A. *et al.* (2014) 'Botnet detection techniques: review, future trends, and issues', *Journal of Zhejiang University SCIENCE C*, 15(11), pp. 943–983. doi: 10.1631/jzus.C1300242.

Lin, I. C. and Liao, T. C. (2017) 'A survey of blockchain security issues and challenges', *International Journal of Network Security*, 19(5), pp. 653–659. doi: 10.6633/IJNS.201709.19(5).01.

Nagaraja, S. *et al.* (2011) 'Stegobot: A Covert Social Network Botnet', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 299–313. doi: 10.1007/978-3-642-24178-9_21.

Nakamoto, S. (2008) *Bitcoin: A Peer-to-Peer Electronic Cash System*. doi: 10.1007/s10838-008-9062-0.

Nazario, J. (2007) 'BlackEnergy DDoS Bot Analysis', p. 11. Available at: http://atlas-public.ec2.arbor.net/docs/BlackEnergy+DDoS+Bot+Analysis.pdf.

Nofer, M. *et al.* (2017) 'Blockchain', *Business & Information Systems Engineering*, 59(3), pp. 183–187. doi: 10.1007/s12599-017-0467-3.

Roffel, D. and Garrett, C. (2014) *A Novel Approach For Computer Worm Control Using Decentralized Data Structures What is the Blockchain ? Technical Implementation of the Blockchain*. Santa Cruz. Available at: https://pdf.yt/d/E2ZwuLAVfC44kEQk.

Sagirlar, G., Carminati, B. and Ferrari, E. (2018) 'AutoBotCatcher: Blockchain-Based P2P Botnet Detection for the Internet of Things', in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, pp. 1–8. doi: 10.1109/CIC.2018.00-46.

Shafi, Q. and Basit, A. (2019) 'DDoS Botnet Prevention using Blockchain in Software Defined Internet of Things', in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, pp. 624–628. doi: 10.1109/IBCAST.2019.8667147.

Symantec Security Response (2015) 'Regin: Top-tier espionage tool enables stealthy surveillance Symantec Security Response'. Available at: https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/regin-top-tier-espionage-tool-15-en.pdf.

Szabo, N. (1997) 'Formalizing and Securing Relationships on Public Networks', *First monday*, 2(9), pp. 1–21. doi: 10.5210/fm.v2i9.548.

Taylor, P. J. *et al.* (2019) 'A systematic literature review of blockchain cyber security', *Digital Communications and Networks*. Elsevier Ltd, (June 2018). doi: 10.1016/j.dcan.2019.01.005.

Tijan, E. *et al.* (2019) 'Blockchain Technology Implementation in Logistics', *Sustainability*, 11(4), p. 1185. doi: 10.3390/su11041185.

Truffle (2019) *Truffle Suite*. Available at: https://www.trufflesuite.com/docs.

Vormayr, G., Zseby, T. and Fabini, J. (2017) 'Botnet Communication Patterns', 19(4), pp. 2768–2796. doi: https://doi.org/10.1109/COMST.2017.2749442.

Wood, G. (2014) *Ethereum: a secure decentralised generalised transaction ledger*, *Ethereum Project Yellow Paper*.

Zarpelão, B. B., Miani, R. S. and Rajarajan, M. (2019) 'Detection of Bitcoin-Based Botnets Using a One-Class Classifier', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Nature Switzerland AG 2019, pp. 174–189. doi: 10.1007/978-3-030-20074-9_13.

Zhang, P. *et al.* (2018) 'Blockchain Technology Use Cases in Healthcare', in *Advances in Computers*. 1st edn. Elsevier Inc., pp. 1–41. doi: 10.1016/bs.adcom.2018.03.006.

Towards Cyber-User Awareness: Design and Evaluation

Toyosi Oyinloye, Thaddeus Eze and Lee Speakman
University of Chester, UK
t.oyinloye@chester.ac.uk
t.eze@chester.ac.uk
l.speakman@chester.ac.uk

DOI: 10.34190/EWS.20.055

Abstract: Human reliance on interconnected devices has given rise to a massive increase in cyber activities. There are about 17 billion interconnected devices in our World of about 8 billion people. Like the physical world, the cyber world is not void of entities whose activities, malicious or not, could be detrimental to other users who remain vulnerable as a result of their existence within cyberspace. Developments such as the introduction of 5G networks which advances communication speed among interconnected devices, undoubtedly proffer solutions for human living as well as adversely impacting systems. Vulnerabilities in applications embedded in devices, hardware deficiencies, user errors, are some of the loopholes that are exploited. Studies have revealed humans as weakest links in the cyber-chain, submitting that consistent implementation of cyber awareness programs would largely impact cybersecurity. Cyber-active systems have goals that compete with the implementation of cyber awareness programs, within limited resources. It is desirable to have cyber awareness systems that can be tailored around specific needs and considerations for important factors. This paper presents a system that aims to promote user awareness through a flexible, accessible, and cost-effective design. The system implements steps in a user awareness cycle, that considers human-factor (HF) and HF related root causes of cyber-attacks. We introduce a new user testing tool, adaptable for administering cybersecurity test questions for varying levels and categories of users. The tool was implemented experimentally by engaging cyber users within UK. Schemes and online documentations by UK Cybersecurity organisations were harnessed for assessing and providing relevant recommendations to participants. Results provided us with values representing each participants' notional level of awareness which were subjected to a paired-T test for comparison with values derived in an automated assessment. This pilot study provides valuable details for projecting the efficacy of the system towards improving human influence in cybersecurity.

Keywords: Human-Factor, Awareness cycle, Interconnected devices, Cyber-attacks, Cyber Awareness.

1. Introduction

Interconnected devices have ubiquitous presence having become integral parts of our everyday lives, serving various essential and convenient purposes. Internet-enabled gadgets offer users possibility of capturing, processing, storing, and rapidly transmitting extremely high volumes of data. Information and communication technologies (ICT) are highly used as they offer enormous convenience (Öğütçü *et al.,* 2016). They are essential repositories for valuable data, making them attractive to cybercriminals, whose activities take huge tolls on the financial and reputational strength of victims. Stakeholders within the cyber world desire technological habitats with sustained data confidentiality, Integrity and availability. This can be achieved by proactively protecting cyber-active systems from detrimental activities but according to Joinson *et al.* (2010), efforts towards cybersecurity is not as progressive as deterrence resulting from successful cyber-attacks. Hackers exploit technical vulnerabilities to attack systems, but human errors have been identified as greater risks to cybersecurity. Users within the cyber-chain are either originators or recipients of data. Humans are prone to errors, often being secondary vectors that foster further exploits (Kelly, 2017). It is generally accepted that inculcating cybersecurity culture in cyber-users would significantly improve cybersecurity and inevitably incur additional costs on systems management. Deliberate steps towards cost-effective, easily accessible and consistent cyber awareness programs are needful. According to Chang and Bresciani (2019) awareness efforts are impaired when stakeholders consider them as one-off processes with a one-size-fits-all mentality.

This study was inspired by the desire to contribute to on-going cyber-awareness campaigns by using an accessible, acceptable and cost-effective process within a user awareness cycle. Factors that could raise concerns for system administrators were considered in building and implementing a new user testing application within this cycle by capitalising on HF and HF related root causes of cyber-attacks. The application was designed to administer cybersecurity test questions with adaptability to various systems and categories of users. Our experimental use of the tool sought to determine the level of cyber literacy of users and their reactions to threat circumstances. Elements within the tool provided specific feedbacks to participants. Steps in the awareness cycle are intended for use in cyber-active systems as pathways towards effective awareness pattern. For case study, 66 cyber-users within the UK participated in the experiment via online surveys providing us with values representing each participants' assessed level of cyber awareness and their notional level of awareness. A paired-T test was introduced to make a comparison between these 2 assessments to evaluate the efficacy of the system. Guidance provided by UK governmental and Cybersecurity organisations was harnessed in establishing realistic model for the experiment, setting baselines for assessing participants and making recommendations. Projections are made towards future works and improvements to the project.

2. Background

Popularity of interconnectivity, cyber-attack track records and adversities due to upcoming technologies signal urgent needs to enhance cybersecurity up to the barest lines of defence. In the past 2 decades, users have been exposed to cyber-threats and continue to seek ways of mitigating risks and attacks. 2017 witnessed historic attack incidents in form of Distributed Denial of Service (DDoS) attacks on Netflix, Twitter, Amazon and The New York Times; A data breach on Equifax; and WannaCry Ransomware. Back then, a quarter of the world's population had interconnected devices (World Bank, 2017) but by 2019, the population of interconnected devices was 17 billion (Lueth, 2018) far outweighing the human population of 8 billion. Gartner (2019) predicted 25 billion distinct interconnected gadgets by 2020 and Newman (2019) further predicts 64 billion of these devices by 2025. These predictions might be realistic with improvements to connectivity. For example, 5G networks could foster malicious exploits that could be initiated and concluded almost without trace, due to high communication speed. According to Mathews (2019), 5G would accelerate the capacity of botnets, which are commonly used by attackers to induce DDoS attacks and steal data. Human intervention will truncate lingering effects of attacks as prompt detection of intrusions would prevent further intrusions and foster victims' recovery. Gartner (2019) declared it nearly impossible to curb targeted attacks unless they are promptly detected. Users' Influence can impact prevention/detection of cyber-attacks but Kritzinger and von Solms (2010) argued that promptness in prevention/detection also depends on technical measures, policies implemented, and domain that systems reside.

The security strategy of systems remains only as strong as their weakest link (IT Governance, 2019). PWC (2018) observed that human errors largely impact cybersecurity and suggested regular user awareness to mitigate risks. They attributed HF vulnerabilities to unpreparedness of cyber-users for cyber-attacks. Unlike their victims, cyber-attackers are not unprepared. Many cybercriminals consider it lucrative and are usually motivated by thoughts of rewards in form of money, political, ethical, intellectual or social incentives. The most sophisticated security facilities are incapacitated by unrelenting attackers especially in the absence of reliable human interventions (Abawajy, 2014).

Losses incurred by cyber-active systems in events of successful cyber-attacks far outweigh the cost of implementing user awareness. According to Albrechtsen and Hovden (2010), user awareness is the most cost-effective step towards cybersecurity. This was affirmed in an experiment by Dodge Jr. *et al.* (2007) on improvements in user awareness after training users against phishing attacks. Fraudwatch international (2018) also reported that expensive advanced security techniques would not suffice without cultured users. Users can be cultivated into security sensors for detecting and proactively reporting suspicious activities (Rogers, 2014). According to Bardia and Kumar (2017), humans naturally tend to proactively protect themselves and their belongings to the best of their knowledge and security awareness. This implies that the human sense of responsibility and tendency to be protective can be subtly stimulated against cyber-attacks. Educational institutions are good avenues for inculcating cybersecurity culture in societies but according to Grachis (2014), the concept of human involvement in cybersecurity is yet extensively integrated into educational curricula.

3. Scope and structure of the research

Firstly, in setting benchmarks and metrics for evaluating our awareness model, we categorise users according to the nature of their regular engagement which largely exposes them to cyber-threats. The categories include:

- Management- Individuals in administrative roles who make use of technological devices for official/personal purposes.
- Technical users- Individuals with expertise in technological devices who make use of them for personal/official purposes.
- End-users- Individuals who are not necessarily tech-savvy but make use of technological devices for personal/official purposes.

We introduced a cyber awareness cycle with test questions built around HF-related factors in cyber-attacks, considering the most effective methods of interacting with users during awareness processes. Processes in this research are structured for evaluating and stimulating the awareness of users to cybersecurity, thereby enhancing their capacity in early detection/aversion of cyber-attacks. Test questions describing scenarios that expose users to security issues they might encounter in the cyberspace, without subjecting them to dangers of actual incidents, were delivered to participants. Outcomes derived from implementing the tool were used for statistical analysis in proving efficacy of the testing tool. In the end, it is presented that user orientation through effective periodic awareness cycle can improve cybersecurity. It is noteworthy to mention that the scope of this study does not eliminate the use of automated techniques in ensuring cybersecurity but promotes active inclusion of users in the big picture.

4. Methodology

An in-depth literature survey on related works afforded us insight on methods for conducting further research on the topic. With an objective to synthesise existing knowledge in an action research exercise, we adopt quantitative and qualitative methods for obtaining valuable details and performing investigations. Tools for investigations include new user testing application and surveys, techniques leveraged the human sense of responsibility and ability to make decisions and choices. This was achieved by using relevant questions to obtain responses that indicate participant's opinion and choices of action in given scenario.

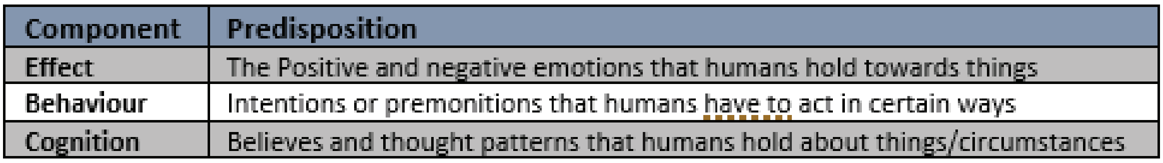
4.1 Approach for the research

Study on best practices for implementing awareness highlighted factors relevant for consideration. Rantos *et al.* (2015) identifies that consistency and continuity are vital towards Information Systems Awareness (ISA) while aiming to reach targeted audiences at the lowest possible cost. Chang and Bresciani (2019) also emphasized the need for continuity presenting that systems require security awareness that are adaptable to them. We have designed our awareness cycle in this respect, as an affordable and relatable model. To cut costs, we designed and implemented the new testing tool adaptable to evolving systems and related factors.

Qualitative approach to our study employed the law of Social psychology which harnesses human attributes exhibited through behaviour, effects of experiences and cognition (Table1) in addressing HF in cyber awareness. Every user is responsible for their point of contact within the cyber-chain. According to Widdowson and Goodliff (2015), a scam on Barclays and Santander banks in 2013 was successfully due to human error within the authentication chain. This attitude was described by Rosenbaum and Blake (1955) as the tendency for humans to slip into *bystander effect* or diffusion of responsibility. Salient factors, including users' attitudes to password security, social engineering, browsing, updates, and malware/viruses, involve human influence. Eagly and Chaiken (1984) attributed some of these risks to the human tendency to trust people they prefer. PWC, (2015) reports that this tendency is easily leveraged for social engineering attacks, tricking people into taking detrimental actions or revealing personal data.

In an investigative research, Thompson and von Solms (1998) defined the components of human attributes in another dimension (Table 2), proposing that people can be persuaded to behave in certain ways to effect desired changes, regardless of their attitude to subjects being addressed. Zimbardo and Leippe (1991) agreed that commending people for taking appropriate actions and otherwise reprimanding them, enforces desired changes. Figure 1 shows samples from 30 multiple-choice questions used to interact with participants to obtain their likely actions/response to situations. Schlienger and Teufel (2003) employed social psychology via cognition component of human psychological attribute, suggesting that cultivated predispositions of humans can affect behaviours. This comprises the individual's attitude to security within their environment, their perception of systems' values and security policies, and what they consider as best solutions to problems. This trichotomy was adopted in composing our test questions.

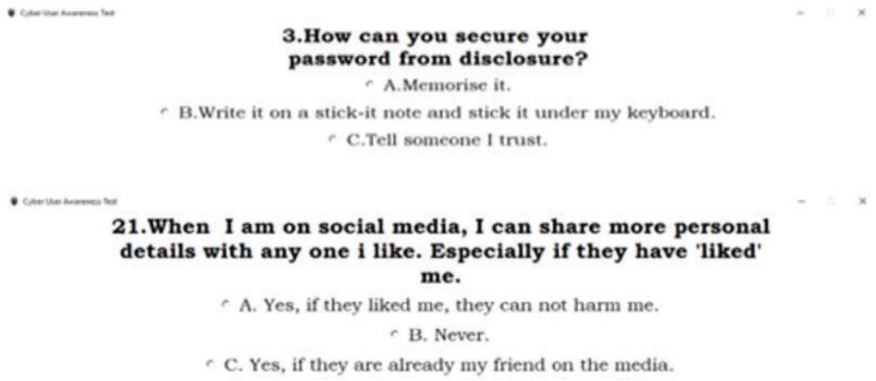
Table 1: Components that influence human dispositions to things/situations



Component	Predisposition
Effect	The Positive and negative emotions that humans hold towards things
Behaviour	Intentions or premonitions that humans have to act in certain ways
Cognition	Believes and thought patterns that humans hold about things/circumstances

Table 2: Another perspective to components that influence human dispositions to things/situations

Component	Description
Knowledge	What does an individual know about a thing/situation
Attitude	What does an individual feel about a subject/event
Behaviour	What does the individual do about a subject/situation



3.How can you secure your password from disclosure?

○ A.Memorise it.

○ B.Write it on a stick-it note and stick it under my keyboard.

○ C.Tell someone I trust.

21.When I am on social media, I can share more personal details with any one i like. Especially if they have 'liked' me.

○ A. Yes, if they liked me, they can not harm me.

○ B. Never.

○ C. Yes, if they are already my friend on the media.

Figure 1: Sample questions from testing tool

Outcomes of investigations were assigned scores to provide numerical values for assessing participants and efficacy of our model. Veseli's (2011) idea of conducting surveys to investigate cyber-users for changes resulting from ISA implementation was imbibed in this study in obtaining users' self-assessments of themselves as values comparable to our automated assessment. This was applicable for projecting efficacy of our system. Users' self-assessment technique was also geared towards evaluating participant's knowledgeability as well as motivate their participation in the investigation. According to Hughes *et al.* (1985) self-assessment improves learning and behaviours.

4.2 Setting Baselines

In the UK, the National Cyber Security Centre provides free sourced facilities to promote cyber-awareness, respond to incidences and reduce risks being faced in various sectors. Schemes and documentations provided by NCSC (2018) were considered in this research for setting awareness baselines and describing risk scenarios. Links to informative webpages were selected and displayed for recommendations to participants.

4.3 Delivery of awareness program

Methods of delivering awareness affects its outcome. According to Abawajy (2014), success of an awareness scheme is highly dependent on how it was delivered to targeted audience. This was hinged on the importance of engaging participants long enough to impact them. Veseli (2011) implemented a module around classroom, discussion-group, and web-based survey to obtain responses from participants. The user testing tool does not require any complex delivery to enhance interactivity to get users engaged for long periods of time because each session lasts approximately 15 minutes. Kruger and Kearney (2006) administered test sessions in their study using prints but here, we present a computer-based method which is cost-effective, flexible and accessible, offering ease of retrieving participants' responses.

Similar investigations by Alotaibi et al (2016) performed web-based surveys for cost-effective and immediate retrieval of responses. In this study, the use of desktop-based model of delivery was preferred to establish its efficacy as well as make distinctions to participants between assesement and being questioned for their points of view. Subsequent versions can be built into web platforms. Online surveys are void of complexities and enable easy retrieval of participants' responses prior to, and after the test.

Idea of gamification for user awareness was applied by Ruboczki (2015) leveraging users' addiction to gaming stating that gamification is useful in sustaining audiences' attention long enough to impact desired knowledge, but admitted that it is effective mostly for low sample population. Harnessing addictions and passions of users to gamification has pros and cons in real life application. Considering the high cost of production and time required to impact targeted audiences, gamification for impacting knowledge was not found valuable for this study.

Alotaibi *et al.* (2014) conducted investigations placing importance on how to present questions to participants with clarity and precision. They adopted a 2-pilot process that engaged 2 different sets of academicians who examined the test questions to confirm their suitability for use in assessing participants with diverse backgrounds. Feedbacks from 2-pilot process guided them to interpret questions appropriately for their Saudi-Arabian recruits. The process for building the testing software relates with this idea by taking agile steps of production in an iterative cost/time effective process.

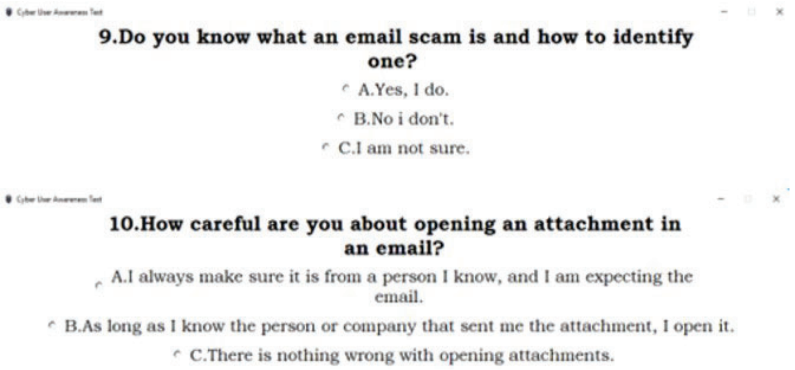
Reaching participants to obtain relevant sample size is part of the delivery process. Posters and survey links were sent to remote participants by using snowball technique inspired by Alotaibi *et al.* (2014). This involves prompting volunteers to forward invitation emails to their friends/families.

4.4 Data retention

Obtaining and retaining as much data samples as possible is valuable for making realistic analysis and deductions. Ridzuan *et al.* (2012) built their questions with options that encourage participants to answer as much questions as possible with intentions to aid data retention. This study established data retention by ensuring that all questions were answered. The testing application is designed as graphical user interface with built-in interactive buttons controlling navigation. Participants could only proceed whenever they had selected an option in response to current questions.

4.5 Data processing and analysis

Surveys were processed by representing speculations numerically with values assigned using Likert scales while the testing tool automatically compared responses with inbuilt correct answers, allocating 10 marks for each correct answer. Kruger and Kearney's (2006) analysed changes in risk profiles by continuously measuring participants' improvement and then using a widely used additive model in a scorecard approach. They suggested the use of a likert scale as an enhanced means for evaluating participants' responses. Abawajy's (2014) study employed a reward/penalty granting scheme matching real-world consequences of falling prey to phishing. Peker *et al.* (2018) applied a reverse scoring strategy to regularise participants' performance prior to, and after tests. They designed an e-learning module characterised by penalties for careless cyber-user practices. In this study, a combination of these schemes was adopted for avoiding false positives/negatives when scoring participants. 10 questions were selected and tagged in scoring, not for any priority other than being related questions. Example is shown in Figure 2 where question 9 tests user's knowledge of e-mail scams as a terminology and question 10 describes a related possible threat scenario.



9.Do you know what an email scam is and how to identify one?

○ A.Yes, I do.

○ B.No i don't.

○ C.I am not sure.

10.How careful are you about opening an attachment in an email?

○ A.I always make sure it is from a person I know, and I am expecting the email.

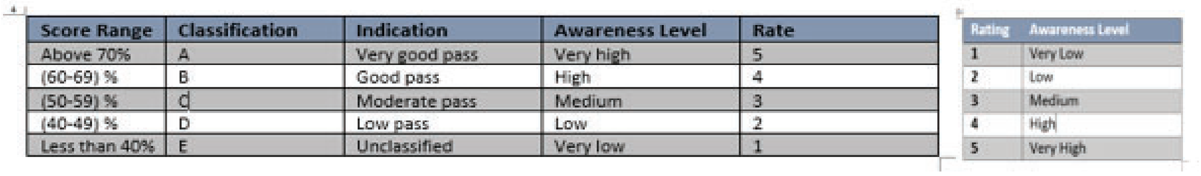
○ B.As long as I know the person or company that sent me the attachment, I open it.

○ C.There is nothing wrong with opening attachments.

Figure 2: Example of tagged questions

Participants' response to question 10 would confirm or dismiss their initial claim. Reversals were issued where succeeding responses contradicted initial ones while scores were assigned for preceding responses that although contradictory, ultimately indicated knowledgeability. At the end of the test, total marks scored were graded in percentage (Total-score/300*100) and classified on a scale of 1 to 5 (Table 3)

Table 3: Score classification with scale



Score Range	Classification	Indication	Awareness Level	Rate
Above 70%	A	Very good pass	Very high	5
(60-69) %	B	Good pass	High	4
(50-59) %	C	Moderate pass	Medium	3
(40-49) %	D	Low pass	Low	2
Less than 40%	E	Unclassified	Very low	1

Rating	Awareness Level
1	Very Low
2	Low
3	Medium
4	High
5	Very High

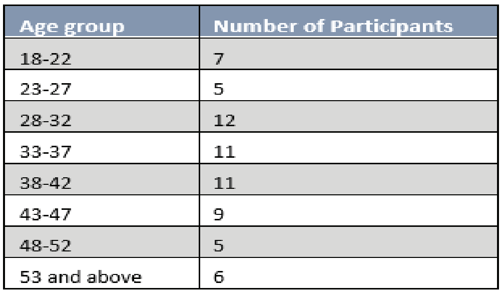
4.6 Evaluating the system

One of the aims of this study is to project effectiveness of the model. In a similar study, Atolaibi *et al.* (2016) measured effectiveness of their model by applying a Chi-Square test to data retrieved from respondents and obtained results indicating that Internet skills influence cybersecurity practices from the users' end. Data retrieved from our investigations were suitable inputs for conducting statistical analysis using Paired-T test which establishes a null hypothesis indicating the efficacy of the new testing tool. Null hypothesis in this case represents the absence of significant difference when automated and self-assessment outcome of the experiment are compared. This difference is denoted by the probability (p-value) of having the same outcome on both assessments and is compared to conventional significance level (.05). P-value lower/greater than .05 leads us to accept/reject our hypothesis.

4.6.1 Pre-assessment survey

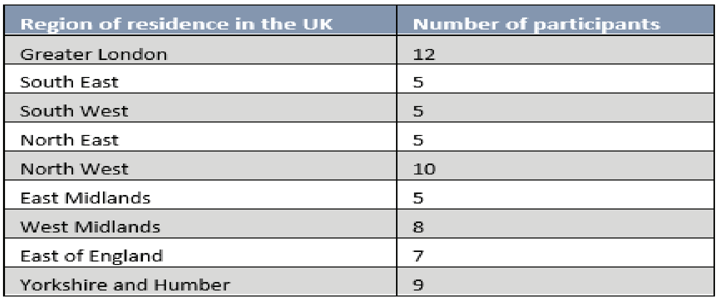
Details retrieved from 66 participants prior to taking tests include demographics (Table 4), participants' region of residence (Table 5), user category, and participants' notional level of cyber awareness (Table 6), to characterise the sample in use and give investigation a structure. Participants were assigned identification numbers after completing the pre-assessment survey.

Table 4: Number of participants within age group



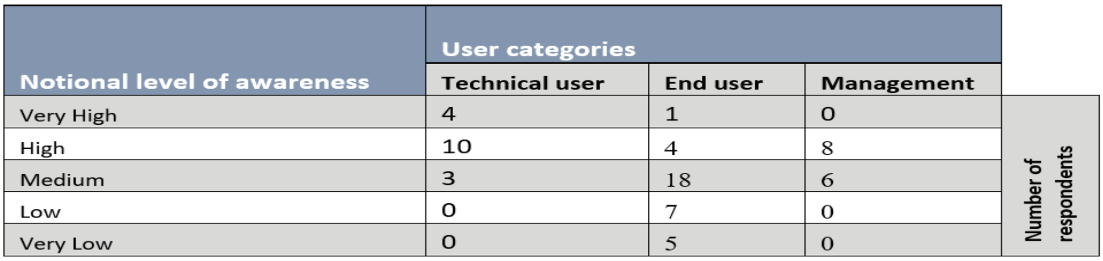
Age group	Number of Participants
18-22	7
23-27	5
28-32	12
33-37	11
38-42	11
43-47	9
48-52	5
53 and above	6

Table 5: Region of residence



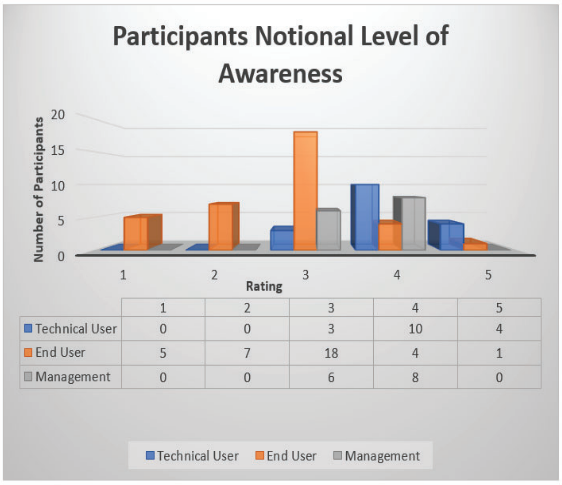
Region of residence in the UK	Number of participants
Greater London	12
South East	5
South West	5
North East	5
North West	10
East Midlands	5
West Midlands	8
East of England	7
Yorkshire and Humber	9

Table 6: Participants' notional level of awareness within user categories

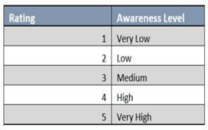


Notional level of awareness	User categories			
	Technical user	End user	Management	
Very High	4	1	0	Number of respondents
High	10	4	8	
Medium	3	18	6	
Low	0	7	0	
Very Low	0	5	0	

Participants' notional level of awareness (Figure 3) varied with each individual. Most technical users indicated medium to very high levels of awareness. Management users indicated medium to high levels of awareness while most end-users claimed a medium level of cyber-awareness. No participants in management and technical jobs indicated low awareness and no management worker claimed very high level of awareness.



	1	2	3	4	5
🟦 Technical User	0	0	3	10	4
🟧 End User	5	7	18	4	1
⬜ Management	0	0	6	8	0



Rating	Awareness Level
1	Very Low
2	Low
3	Medium
4	High
5	Very High

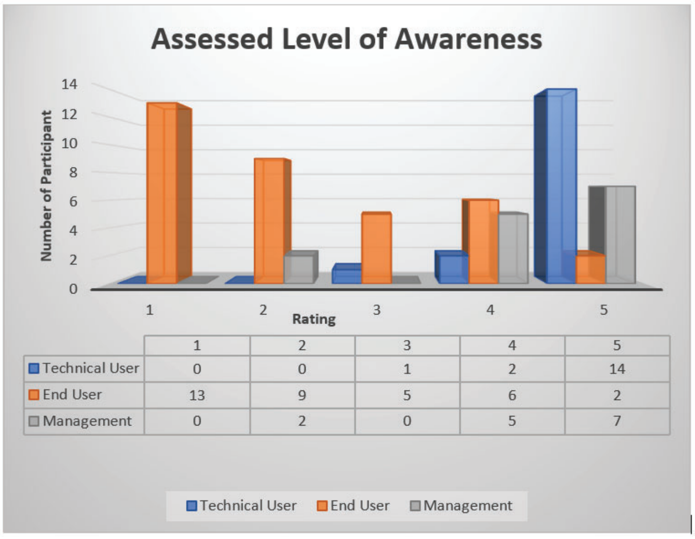
Figure 3: Representation of participants' notional level of awareness

4.6.2 Implementation of the model

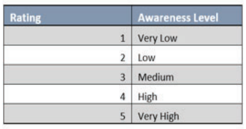
The testing application was implemented by engaging participants in the test. Performance of participants varied (Figure 4). A log of outcomes of each participants' interaction with the model reveal notes of interest. Highest score was 93% and lowest 17%, while 50% of total participants scored above average. An interesting finding was the outstanding performance of management users. This is desirable as workers in management have been identified to be mostly at risk of being targeted for cyber-attacks.

Veseli (2011) emphasized needs for management support and involvement in ISA within organisations considering that ISA successes are highly dependent on their input. Most Technical users performed highly. This is expected based on their relatively higher level of ICT education compared to other users. Awareness of most end-users appear to be very low, which raises concerns. Low performance indicates low level of awareness.

Inappropriate responses to threats equate falling prey to attacks. Specific recommendations shown in Figure 5, were displayed for each participant at the end of their test sessions. Other findings revealed responses like "*I do not know*" and "*I am not sure*". For example, participant *004* who scored 20% made several picks of "*I am not sure*". Such responses according to Ridzuan *et al.* (2012) indicates that they are not knowledgeable in question areas. Individuals with low performances, who besides private use, regularly interact with computerised devices at work were identified. Examples are *5843*, *5866* and *004* performing below average showing that regular exposure to devices does not necessarily equate knowledgeability. Responses to questions based on salient factors also revealed areas of weakness/strength. For example, participant *004* selected "p1ssword" as their strong password, which though is easily memorised, is a very weak choice. *009* selected password formed out of a combination of their first and last names which can eaily be traced to them by familiar attackers. *009*'s choice to open an email bearing familiar sender's name might not be safe because phishers are known to camouflage phishing mails using familiar names. *009* and *005* consider updates unnecessary, whereas updates are released by software vendors to enhance previous versions or as mitigations for known vulnerabilities. Participants *001* and *003* would shut down their PC if they got virus alerts rather than resolving it. Participant *004*'s response to Social engineering tactics would permit strangers into systems without adequate checks while *016* considers social media acquaintances trustworthy especially if they seemed to share common interests.

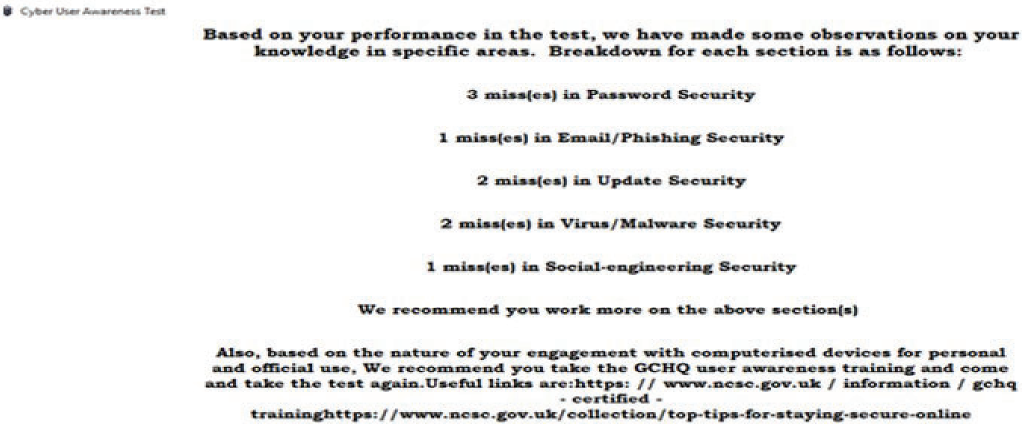


	1	2	3	4	5
Technical User	0	0	1	2	14
End User	13	9	5	6	2
Management	0	2	0	5	7



Rating	Awareness Level
1	Very Low
2	Low
3	Medium
4	High
5	Very High

Figure 4: Performance of participants by user-categories



Based on your performance in the test, we have made some observations on your knowledge in specific areas. Breakdown for each section is as follows:

3 miss(es) in Password Security

1 miss(es) in Email/Phishing Security

2 miss(es) in Update Security

2 miss(es) in Virus/Malware Security

1 miss(es) in Social-engineering Security

We recommend you work more on the above section(s)

Also, based on the nature of your engagement with computerised devices for personal and official use, We recommend you take the GCHQ user awareness training and come and take the test again.Useful links are:https: // www.ncsc.gov.uk / information / gchq - certified - traininghttps://www.ncsc.gov.uk/collection/top-tips-for-staying-secure-online

Figure 5: Sample recommendation with direct link to informative site

4.6.3 Handling false positives and false negatives

Knowledgeability of users in cybersecurity terminologies could foster adequate information rendering for raising alerts or reporting incidents. Performances of participants in tagged questions showed that most participants were conversant with cyber terminologies but not necessarily with actual circumstances. There are factors that could contribute to this, for example, the tendency for participants to dash through questions without proper consideration of the given scenario. According to Luce and Khan (1999), this could result in false positive/negative outcomes and require the test to be repeated. Our scoring scheme double-checks responses to rectify false positives/negatives. For example, *003* was penalised for claiming to know what email scams are and then not recognising an email scam scenario.

4.6.4 Participants' notional level of awareness vs assessed level of awareness

Investigations have afforded us two different perspectives of assessing each cyber-user. Participants' notional level of awareness in this context represents individuals' self-assessment while assessed level of awareness represents automated ratings assigned to each participant based on their performance in the test. With these values, we can test, and accept/reject hypothesis for evaluating the new testing tool. Figure 6 depicts comparison between these outcomes. Differences appear to be sparse, making them statistically relevant. Mean difference of these sets can be applied in calculating p-value which is the level of marginal significance within statistical hypothesis.

Hypothesis: *There is significant difference between assessed and notional assessment outcome of each participant.*

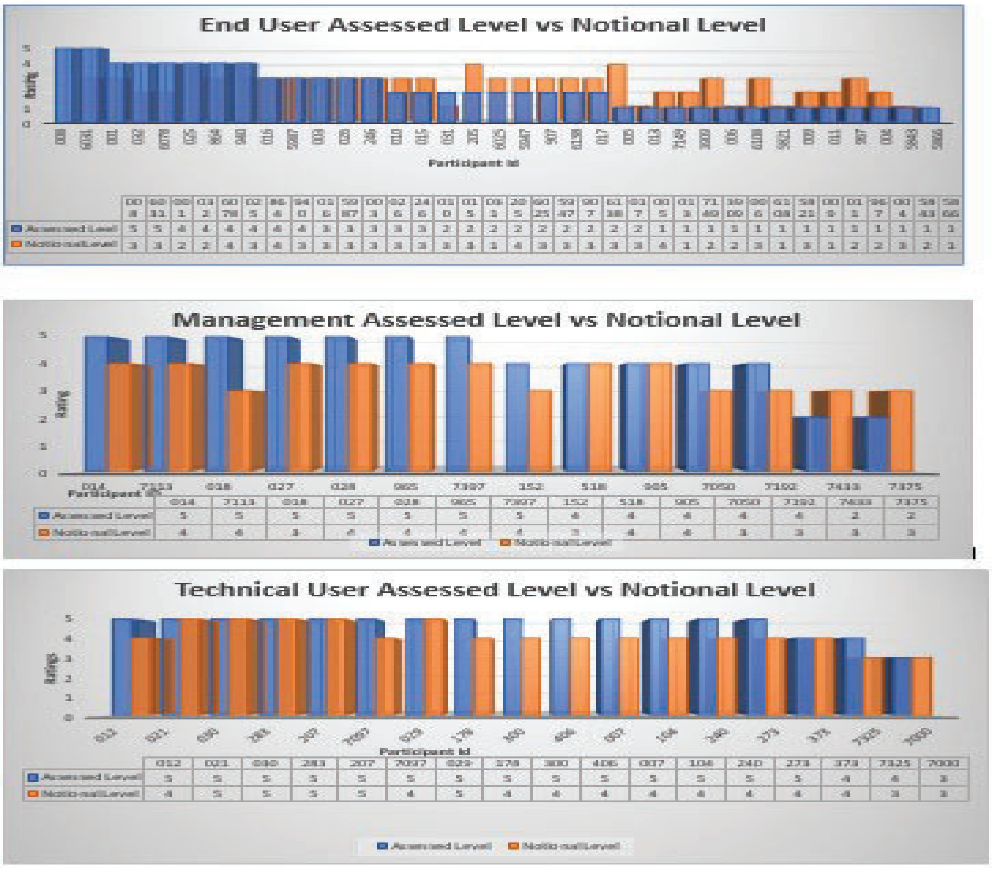


Figure 6: Comparison of assessments in each user-category

Paired-T test:

Calculate difference d_i from each pair y_i and x_i.

$$d_i = y_i - x_i$$

Where $y_i = Assessed\ level\ of\ awareness$ and $x_i = Notional\ level\ of\ awareness.$

Mean difference $\bar{d} = 0.11$ standard deviation $(s_d) = 1.11$

Calculate standard error of mean difference: $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{1.11}{\sqrt{66}} = 0.137$

Calculate t-statistic $T = \frac{\bar{d}}{SE(\bar{d})}$ which follows t-distribution with $n - 1$ degrees of freedom under null hypothesis.

Obtain p-value $T = \frac{0.11}{0.137} = 0.81$, p >.05 (reject hypothesis)

Probability of getting difference between y and x is highly insignificant indicating strong confidence of having same values from both feedbacks, indicating efficacy of the testing tool.

4.6.5 Post assessment survey

Participants' evaluation informs us of the impression made on them and their readiness to make better choices. An ideal awareness program must influence users' behaviours to produce measurable benefits (Abawajy, 2014). Figure 7 shows majority of participants accepting effectiveness of the testing system.

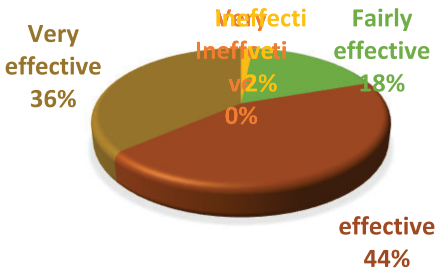


Figure 7: Participants' evaluation of testing application

Participants indicated their agreement/disagreement with the automated assessment. Most participants (Figure 8) agree that scores given them truly represents their awareness. An indication of their acceptability of the tool.

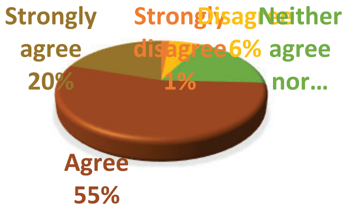


Figure 8: Participants' acceptability of testing application

5. Discussion

While considering the relevance of human influence in cybersecurity, we can equate each miss in our experiment as inroad for exploits. Responses to tagged questions indicate that users are often conversant with cyber terminologies, which is relevant for proper reporting of incidents or threats but does not amount to awareness. Ridtzuan *et al.*'s (2012) made similar observations where participants claimed knowledgeability in online spam, and were indeed conversant with ICT related terminologies, but further investigations revealed that they misconstrued what online spam entailed.

This study has shown that user testing is impactful towards improving user awareness. Data obtained in the process of this research were found useful to conduct our investigations and relevant for future studies. For example, regions of residence can be used in analysing this topic from a regional perspective. In addition to this, focus on demographics highlights statistical distributions for further investigations.

6. Conclusion

The essence of our research is to contribute to the big picture, towards promoting cybersecurity. Phases in the lifecycle of this project have presented insight on the status quo and possible solution for cyber-active systems. Investigations have demonstrated circumstances surrounding HF-related factors of cyber-attacks and how users' choices of actions impact cybersecurity. The paired-T test derived a high confidence in the efficacy and reliability of the user testing tool. Free sources of guidance on safety of cyber-users were highlighted for improvements in noted areas of concern. Means of sensitisation like e-mail snowball technique and use of posters were identified as effective for informing participants. Positive feedbacks show that users can relate to continual need for awareness and appreciate recommendations given them. This one-off process has impactful potential that can be repeated. A wholesome awareness program considers all relevant factors and identifies objectives that increase/reduce values affecting its effectiveness. Figure 9 depicts the cycle created in this research.

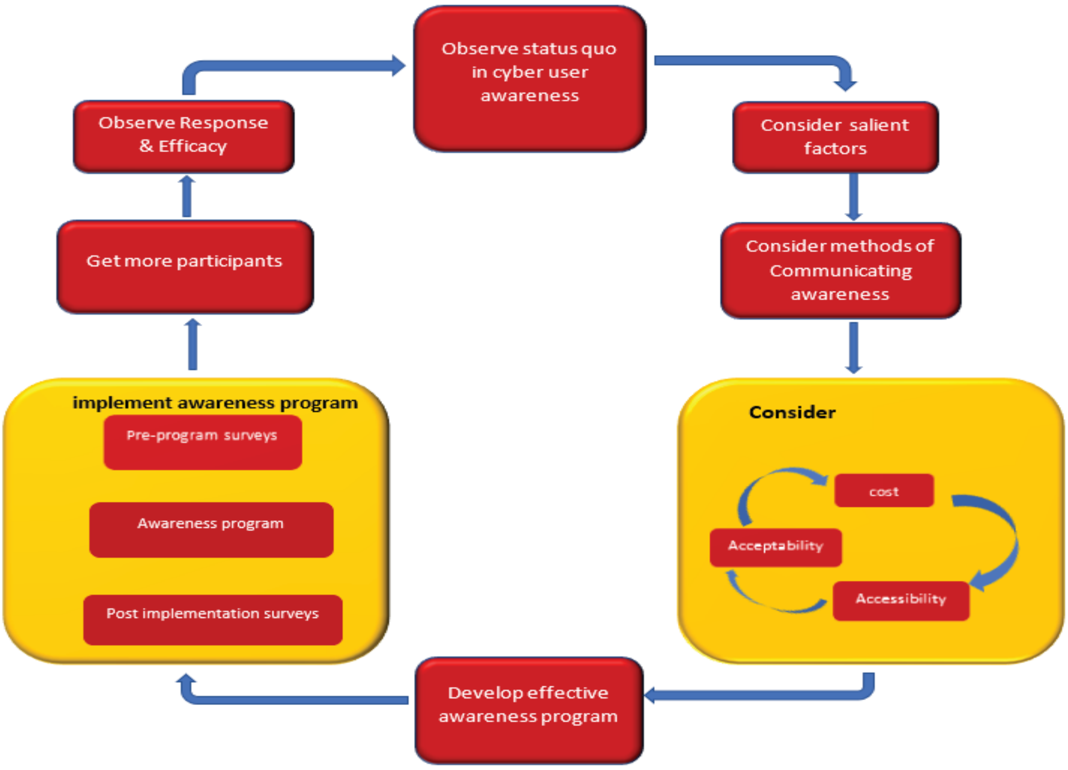

Observe status quo in cyber user awareness
Consider salient factors
Consider methods of Communicating awareness
Consider
cost
Acceptability
Accessibility
Develop effective awareness program
implement awareness program
Pre-program surveys
Awareness program
Post implementation surveys
Get more participants
Observe Response & Efficacy

Figure 9: Cycle of an effective user awareness program

User awareness process channelled strategically could avert imminent dangers. The human instinct for consistently sustained sense of wholesomeness can be constantly refreshed to empower users in their readiness to handle situations from a safety conscious perspective, continually cultivating them into 'cyber-detectives'.

References

Abawajy, J., 2014. User preference of cyber security awareness delivery methods. *Behaviour and Information Technology,* 33(3), pp. 237-248 .

Albrechtsen, E., and Hovden, J., 2010. Improving information security awareness and behaviour through dialogue, participation and collective reflection. An intervention study. *Computers and Security,* June, 29(4), pp. 432-445.

Alotaibi, F., Funnel, S.,Stengel, I., and Papadaki, M., 2016. *A Survey of Cyber-Security Awareness in Saudi Arabia.* Barcelonia, Spain, IEEE, pp. 154-158.

Bardia, V., and Kumar, C., 2017. *End Users Can Mitigate Zero Day Attacks Faster.* Hyderabad, India, IEEE, pp. 935-938.

Chang, C. and Bresciani, M., 2019. *Does Your Security Awareness Program Put People First?.* [Online] Available at: https://securityintelligence.com/posts/does-your-security-awareness-program-put-people-first/ [Accessed 31 December 2019].

Dodge Jr., R.C., Carver, C., and Ferguson, A. J., 2007. Phishing for user security awareness. *Computers and Security,* February, 26(1), pp. 73-80.

Eagly, A. H., and Chaiken, S., 1984. Cognitive Theories of Persuasion. *Advances in Experimental Social Psychology,* 1 December, 17(C), pp. 267-359.

Fraudwatch international, 2018. *Blog:Fraudwatch international.* [Online]Available at: https://fraudwatchinternational.com/security-awareness/what-is-cyber-security-awareness-training/ [Accessed 07 August 2019].

Gartner, 2019. *Newsroom: Gartner Inc..* [Online] Available at: https://www.gartner.com/en/newsroom/press-releases/2019-08-29-gartner-says-5-8-billion-enterprise-and-automotive-io [Accessed 21 January 2020].

Grachis, G., 2014. [Online] Available at: https://www.csoonline.com/article/2134449/strategic-planning-erm/the-new-security-perimeter-human-sensors.html [Accessed 30 December 2018].

Hughes, B., Sullivan, H. J., and Mosley, M. L., 1985. External evaluation, task difficulty, and continuing motivation.. *The Journal of Educational Research,* 78(4), pp. 210-215.

ITGovernance, 2019. [Online] Available at: https://www.itgovernance.co.uk/what-is-cybersecurity [Accessed 19 August 2019].

Joinson, A. N., Reips, U-D., Buchanan, T., and Schofield, C. B. P., 2010. *Human-Computer Interaction,* 29 March, 25(1), pp. 1-24.

Kelly, 2017. [Online] Available at: https://chiefexecutive.net/almost-90-cyber-attacks-caused-human-error-behavior/ [Accessed 10 August 2019].

Kritzinger, E., and von Solms, S.H., 2010. Cyber security for home users: A new way of protection through awareness enforcement. *Computers and Security,* November, 29(8), pp. 840-847.

Kruger, H. A., Kearney, W.D., 2006. A prototype for assessing information security. *Computers and Security,* June, 25(4), pp. 289-296.

Luce, M. F., and Khan, B.E., 1999. Avoidance or Vigilance? The Psychology of False-Positive Test Results. *Journal of Consumer Research,* 01 December, 26(3), p. 242–259.

Lueth, K. L., 2018. *IOT Analytics.* [Online] Available at: https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/ [Accessed 05 April 2019].

Mak, P., 2017. [Online] Available at: https://www.vircom.com/blog/human-factors-in-cyber-security-preventing-errors/ [Accessed 10 August 2019].

Mathews, K., 2019. *Malwarebytes Labs.* [Online] Available at: https://blog.malwarebytes.com/101/2019/05/how-5g-could-impact-cybersecurity-strategy/ [Accessed 07 August 2019].

McDonald, N., 2013. Prevention Is Futile in 2020: Protect Information Via Pervasive Monitoring and Collective Intelligence. *Gartner*, 30 May, p. Online.

NCSC, 2018. *Guidance.* [Online] Available at: https://www.ncsc.gov.uk/collection/top-tips-for-staying-secure-online [Accessed 14 August 2019].

Newman, L.H., 2019. [Online] Available at: https://www.wired.com/story/whatsapp-hack-phone-call-voip-buffer-overflow/ [Accessed 21 October 2019].

Öğütçü, G., Testik, Ö. M., and Chouseinoglou, O., 2016. Analysis of personal information security behavior and awareness. *Computers and Security,* February, Volume 56, pp. 83-93.

Peker, Y., K., Ray, L., and Da Silva, S., 2018. *Online Cybersecurity Awareness Modules for College and High School Students.* Huntsville, AL, USA, IEEE, pp. 24-33.

PWC , 2015. *2015 Information security breaches survey,* London: Crown.

PWC, 2018. *PWCUk:Issues:Cyber Security: Insights:Global state of information security.* [Online] Available at: https://www.pwc.co.uk/issues/cyber-security-data-privacy/insights/global-state-of-information-security-survey.html [Accessed 23 July 2019].

Rantos, K., Fysarakis, K., and Manifavas, H., 2015. How Effective Is Your Security Awareness Program? An Evaluation Methodology. *Information Security Journal A Global Perspective,* January, 21(6), pp. 328-345.

Ridzuan, F., Potdar, Z., & Hui, W., 2012. *A survey of awareness, knowledge and perception of online spam.* Seoul, IEEE, pp. 1106-1110.

Rogers, J., 2014. *Activating Your Human DefenceNetwork(Change your employees security behaviour, or watch out!).,* Online: s.n.

Rosenbaum, M., and Blake, R. R., 1955. Volunteering as a function of field structure. The Journal of Abnormal and Social Psychology. *The Journal of Abnormal and Social Psychology,* 50(2), pp. 193-196.

Ruboczki, E. S., 2015. How to Develop Cloud Security Awareness. *10th Jubilee IEEE International Symposium on Applied Computational Intelligence and Informatics,* 21-23 May, Volume 10, pp. 323-326.

Schlienger, T., and Teufel, S, 2003. Information security culture – from analysis to chnage. *South African Computer Journal,* Volume 31, pp. 46-52.

Thompson, M. and von Solms, R., 1998. Information security awareness: educating your users effectively. *Information Management and Computer Security,* 01 October, 6(4), pp. 167-173.

Veseli, I., 2011. Measuring the Effectiveness of Information Security program. *Master's Thesis, Master of Science in Information Security, Department of Computer Science and Media Technology.*

Widdowson, A. J., and Goodliff, P. B., 2015. *CHEAT, an approach to incorporating human factors in cyber.* Bristol, UK, IET (IEEE), pp. 1-5.

World Bank, 2017. *World Development Indicators: The information society,* Online: The World Bank Group.

Zimbardo, P.G., and Leippe, M. R., 1991. *The psychology of attitude change and social influence.* New York: McGraw-Hill, New York.

OS Kernel Malware Detection through Data-Characterization of Memory Analysis

Harmi Armira Mohamad Har, Farihan Ghazali and Maslina Daud
CyberSecurity Malaysia, Cyberjaya, Selangor Malaysia
harmi.armira@cybersecurity.my
farihan@cybersecurity.my
maslina@cybersecurity.my
DOI: 10.34190/EWS.20.025

Abstract: In this modern technology, malware is rapidly evolving through various stealth techniques to avoid detection. The evolvement and advanced trick of modern malware caused the code-centric approach to become less-effective and inflexible as they (modern malware) are good at hiding themselves and cover up their tracks. Furthermore, Operating System (OS) rootkits can be varying the pattern code execution that capable of confusing behavior-based malware detector. It shows that approaches such as code-centric and malware-based behavior tend to become unreliable with the evolvement of modern malware, especially rootkits. Thus, this study proposed a quite brand-new approach, which is a data-centric approach by characterizing the OS kernel malware through memory analysis. This approach tries to detect OS rootkits based on trace patterns found in the content of the memory dump. This approach combines digital forensics disciplines to gather the required dataset for analysis. The framework proposed consists of two main stages. The first stage in this framework is a Dataset of Rootkits Characterization that will create a dataset by identifying unique characteristics in memory dump content (sample) that indicates the trace of rootkits. The second stage, which to determine the Rootkits Presence that can detect rootkits based on signature created on stage one. Through this proposed approach, a set of Data Behavior Element (DBE) is generated. There are eight (8) of the Data Behavior Element (DBE) identified throughout the data extracted from the collected samples. This DBE can be used as a signature to detect the OS rootkits. Eight (8) samples are collected that consist of clean (benign) and malicious (infected), including some unknown samples that will be used for False Positive test. As for False Positive test, the result successfully indicated that the signature was capable of showing the presence of OS rootkits. This proposed approach is the proof of concept that data-centric are capable of detecting the presence of malware, especially OS rootkits, as the existing approach mostly implements a code-centric approach.

Keywords: OS Kernel Malware, Malware Analysis, Data-Centric Approach, Memory Analysis

1. Introduction

Malicious software, also known as malware, is a type of malicious code in software that can be used to compromise computer operations, gather sensitive information, gain access to private computer resources, and do any illegitimate action on data, host, or networks. Malware can infect and exploit the resource from various system platforms. There are various classes of malware, such as viruses, worms, Trojans, bots, back doors, rootkits, etc. Malware is considered as dangerous as it can attack main security goals, which is confidentiality, integrity, and availability. Nowadays, malware creator also works on anti-antivirus techniques to give a complex challenge for researchers to detect malware. This is because most of the anti-antivirus is aims to bypass the existing antivirus system.

A rootkit is one kind of malware that essentially related to the OS. With a term that made up of two words, which ‘root’ is commonly giving the user full administrative over the system, while ‘kit’ is set of applications that carry out tasks or administrative processes (Hili et al, 2013). There is various type of rootkit available, and type rootkits suit is depending on the specific system of the OS. Rootkits are considered as one of stealth malware as it is very difficult to be detected. Moreover, it can modify and manipulate OS modules once infected and is also able to cover back their tracks. Some rootkits can infiltrate the kernel level, which gives them complete control over the OS.

There are common approaches used by the attacker to distribute rootkits. One of them is by infecting the existing and legitimate web servers that cause users unaware of the risk they confront. Another approach is where the attacker creates a common application or common files which appear as legitimate yet contain harmful script once executed.

The reasons why rootkits are difficult and complex to be detected in the attack techniques evolve along the time. Rootkits can alter the output of common systems administrative commands such as a list of a running process and all the opened port (Prakash et al, 2013). They will give an inaccurate report that indicates that everything is clean. Even though several methods help in detecting rootkits, but it still may lead to false-positive results. The main thing about rootkits is they able to modify software applications that may capable to reporting a fake results. Also, rootkits are designed to be remaining undetected, so they are experts on covering their track. Among all the common methods of detecting rootkits, behavioral analysis is more reliable. The framework proposed is conceptually based on behavior analysis of rootkit.

Rootkits can be divided into two types, generally, which are application rootkit and kernel rootkit. Application rootkits are a rootkit that establishes at the application layer, while kernel rootkits are a rootkit that can infect deep into the kernel layer. This paper focuses on kernel rootkits, which are trickier as they are difficult to be detected. Generally, kernel rootkits can hide a process and files, hiding network connection, able to redirect file execution, etc. However, the activity or the behavior of rootkit leaves a footprint that later can be used as a pattern to detect their presence.

Furthermore, in this paper, memory analysis based on volatility information in RAM is used. Memory analysis is using memory image (memory dump) to determine the running program, operating system information, and overall state of the computer. A raw memory dump is a complete snapshot of memory that records the content of system memory, data from processes that were running when the memory dump was collected.

Normally memory dump is used as evidence in court as it consists of volatile information (Amari, 2009). As mention before, memory image can be used to determine information about running programs, operating system states, and also to locate deleted or temporary information as long as the machine is on. This because the memory dump consists of volatile data that be recorded when the machine is on and will be lost when the machine is off. Several forensics tools are able to acquire memory dump, such as FTK Imager. In this paper, the FTK imager is used to acquire a memory dump from Windows 7. Moreover, several software and tools that can be used to analyze the sample of the memory dump. One of them is a tool in Kali Linux (Volatility Framework).

As mention previously, a memory dump is normally used as evidence as it may consist of valuable information that may help in the investigation. However, there are plugins in the Volatility Framework that can be used to trace the footprint of rootkits (Halleigh et al, 2014). There are supported plugins for MS Windows, Linux, and Mac OS memory dump. There are many available plugins such as process memory, kernel memory, and objects, networking, registry, malware, and file system. Thus, we need to identify which plugin is suitable to be used and can show the significant result to trace rootkits' activities and behavior.

The remaining of this paper covers as follows: Section 2 provides a problem statement, Section 3 explaining the proposed solution, Section 4 briefly describes the result, Section 5, which is concluding this paper, and lastly, Section 6 explains future direction for this work.

2. Problem Statement

Modern malware tends to become tricky and confusing against malware scanner as they can combine several characteristics of the undesirable programs from different classes (Hili et al, 2013). Besides, they are also evolved rapidly in every aspect, especially on advancing their attack strategies (Rhee et al, 2014). To avoid detection, some malware uses obfuscating techniques such as reuse a legal code, capable of modifying their structure (polymorphism), and also able to replace some routine of targeted resource (stealth virus) (Ford et al, 2007). Those evolvements and advanced tricks cause a code-centric approach to become ineffective. Especially on the OS environment, it is very difficult to detect malware such as rootkit as they are able to modify and replaced the OS modules to cover up their track (Guri et al, 2015).

Code-centric becomes more unreliable to deploy on the detection of OS kernel malware as they good in hiding themselves (Musavi et al, 2014). Moreover, not only able to trick the code-centric approach but OS kernel malware also able to circumvent detection by varying the pattern code execution that will confuse behavior-based malware detectors (Sharif et al, 2008). This shows that several approaches like code-centric and malware-based behavior, tend to become unreliable with the evolvement of malware.

The main problems that need to be explored are rootkits expert in hiding their presence and able to subvert normal operating system behavior. There are several techniques used by rootkits to subvert OS behavior such as hooking OS APIs and system call table (SDT), hiding in unused space on machine’s hard disk, and infecting the master boot record (MBR). SDT contains a data structure that is able to process system calls. By manipulating this table, rootkits can insert malicious instructions. Kernel rootkits are not only able to add new code but also have the capability to delete and replace the operating system code. This capability makes kernel rootkits becomes inevitable once infect the operating system. Therefore, by only depends on the code-centric approach, it could be impossible to detect kernel rootkits once infected. Besides, the code-centric approach normally used specific characteristics for each rootkit. This makes it less flexible when it comes to unknown or new rootkits attack.

3. Proposed Solution

In this study, Data-Centric OS Kernel Malware Characterization Framework is being proposed to detect operating system rootkits based on trace patterns found in the memory dump. As mention previously in the problem statement section, there is much previous work that proposes on OS rootkit signature based on either code-centric or data-centric approach. However, most of the researchers proposed a signature that mainly for specific rootkits or even for a specific behavior (Lin et al, 2011)(Davide et al, 2010)(Shahzad et al, 2013)(Case et al, 2015)(Case, 2016). These specific rootkits' signature is inflexible against evolved or advanced rootkits attack as they used a fixed signature. With the evolution of rootkits structure, attack strategy, and behavior, those specific signatures become less reliable to detect rootkit's presence.

Windows 7 Operating System with Service Pack 0 (SP0) is being choosen as an OS sample for memory dump. There are two types of memory dump that will be collected from Windows 7 (SP0) which is benign (clean) and malicious (infected) sample. The samples of memory dump will be collected by using FTK Imager (forensic tool). The samples later will be analysed using built-in analysis tools in Kali Linux which is Volatility Framework version 2.6.

Besides, this study proposed a standard and general signature of rootkits that aim to detect rootkits presence without depends on specific rootkits structure, attack strategy, and behavior. The general rootkits signature is proposed based on the characteristics and patterns found in the memory dump content. The characteristics and patterns are collected from various types of rootkits. This is to standardize the signature and applicable to various rootkits.This framework generally is using the pattern of suspicious information in memory dump content to create a signature. Later, the signature created will be used to detect kernel rootkits. This framework consists of two main stages. Figure 1 below shows the framework and its stage.

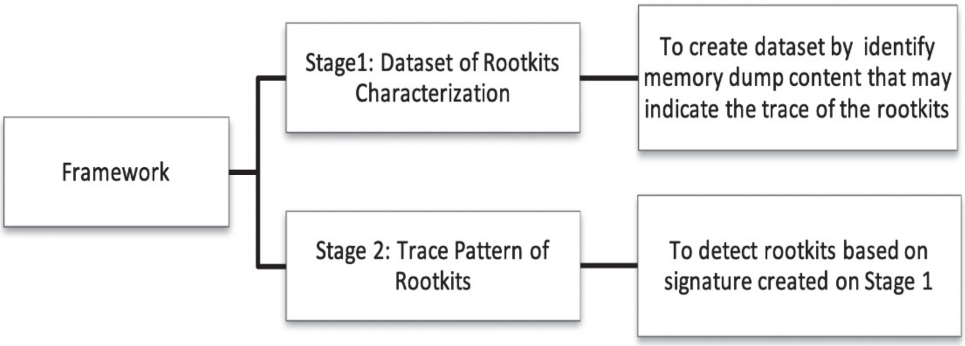

Framework
Stage1: Dataset of Rootkits Characterization
To create dataset by identify memory dump content that may indicate the trace of the rootkits
Stage 2: Trace Pattern of Rootkits
To detect rootkits based on signature created on Stage 1

Figure 1: Data-Centric OS Kernel Malware Characterization Framework

3.1 Stage 1: Dataset of Rootkits Characterization

3.1.1 Dataset of Data Behaviour Profile (DBP)

Data Behaviour Profile (DBP) is based on the information of patterns found in the memory dump content. To create a DBP, a benign sample (clean) and a malicious sample (infected) are collected and analyzed. DBP consists of Data Behaviour Profile (DBE) that is created based on the information of the pattern found in memory dump content. In other words, DBP is a representation of a set of DBE.

Static analysis is being done toward samples collected by using the existing text editor tool (DiffMerge) to identify DBE. DiffMerge can make a comparison of a few text files. Therefore, based on the differences between the clean and infected sample, the one that has a consistent pattern will be highlighted. The highlighted difference will be analyzed and will convert into DBE. Below is the list of rootkits malware used in this study:

List of rootkits used:

1. Malicious sample 1: Win32.Stuxnet
2. Malicious sample 2: Rustocks (Variants 1)
3. Malicious sample 3: Purple Haze
4. Malicious sample 4: Alureon
5. Unknown Sample 1: Win32.Sekoia Rootkits
6. Malicious sample 2: Cutwail

3.1.2 List of Plugin used

This study makes use of memory analysis as a technique to collect and retrieve valuable information in the memory dump. The information collected will be used to do analyzing and find the pattern that shows the abnormalities. Manual static analysis is being done to identify and highlight the differences between benign and malicious samples. Those differences should be consistent and valid to be used as Data Behaviour Element (DBE).

The list of the plugin used in this study as follows:

1. malfind: A plugin that designed to find and trace remote code injections
2. svcscan: A plugin that able to scan all Windows Services
3. ldrmodule: A plugin that able to detect unlinked dlls
4. devicetree: To show device tree
5. timers: Print kernel timers and associated modules DPCs
6. callbacks: A plugin that able to print system-wide notification routines

3.1.3 Generate Signature

As mention previously, the consistent differences between clean and infected samples are turned into DBE. Next, the DBE is used to create the signature for each sample. To create the signature, two (2) clean samples and four (4) infected samples are being used.

3.2 Stage 2: Trace Pattern of Rootkits

3.2.1 Tool for Data Aggregation

This Data Aggregation tool mainly to calculate the probability or the chances of rootkits presence based on sample (DBP) provided. Besides, this tool also able to store the DBP of clean and malicious samples that serve as reference samples for the next analysis.

3.2.2 Static Analysis (Generate Signature)

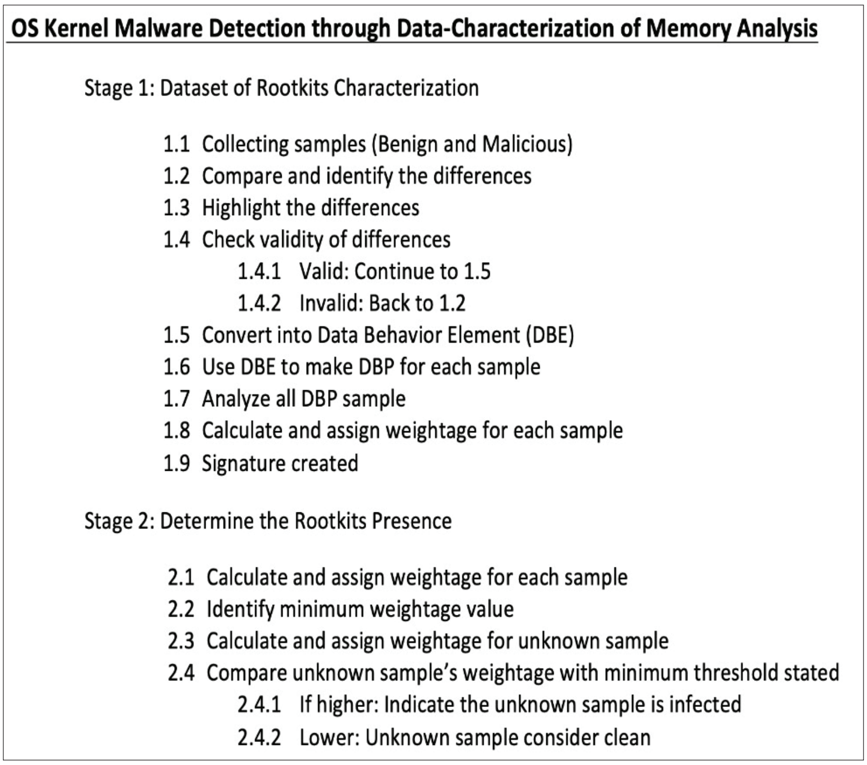
Next, based on DBE that is being converted from highlighted differences between benign and malicious samples, a signature is created. This signature is based on the DBE list. For each sample, an analysis is done to check whether the sample fulfills the DBE conditions. If the samples meet the DBE conditions, value one (1) will be assigned, while value 0 is for unfulfilled conditions. There will be a weightage value that is assigned for each DBE with a minimum of 1 and a maximum of 4. This weightage value is assigned based on the frequency of four (4) malicious samples that being chosen fulfill for that DBE.

3.2.3 Design of the Second Stage

The second stage of this framework is to determine the presence of the rootkit. This stage aims to identify and calculate the percentage of rootkits presence in the sample. To achieve the goal, simple Data Aggregator tools are created by using NetBeans IDE 8.0.2 and Java language. This tool consists of two main functions which are the first function is able to store the signature of clean samples that are label as DBP Setting. Next is Data Analyse where the unknown sample signature will be entered and is being analysed to calculate the probability of rootkits presence.

3.3 Overall Method

Overall, this framework consists of two main stages as mention previously. Thus, to get a clearer view, an algorithm for this framework as Figure 2 below. This provided algorithm states each activity/process in the framework.



OS Kernel Malware Detection through Data-Characterization of Memory Analysis

```
Stage 1: Dataset of Rootkits Characterization

    1.1  Collecting samples (Benign and Malicious)
    1.2  Compare and identify the differences
    1.3  Highlight the differences
    1.4  Check validity of differences
        1.4.1   Valid: Continue to 1.5
        1.4.2   Invalid: Back to 1.2
    1.5  Convert into Data Behavior Element (DBE)
    1.6  Use DBE to make DBP for each sample
    1.7  Analyze all DBP sample
    1.8  Calculate and assign weightage for each sample
    1.9  Signature created

Stage 2: Determine the Rootkits Presence

    2.1  Calculate and assign weightage for each sample
    2.2  Identify minimum weightage value
    2.3  Calculate and assign weightage for unknown sample
    2.4  Compare unknown sample's weightage with minimum threshold stated
        2.4.1   If higher: Indicate the unknown sample is infected
        2.4.2   Lower: Unknown sample consider clean
```

Figure 2: Data-Centric OS Kernel Malware Characterization Framework Algorithm

4. Result

The result of the first phase in the first stage is a dataset of DBP that consists of 8 Data Behaviour Element (DBE). As mention before, each attribute in DBE is being chosen based on valid differences of a benign and malicious sample. The list below is the findings of this study. The output retrieved from memory extraction is being analyzed based on the eight DBE.

List of Data Behaviour Element (DBE):

1. Suspicious/hidden injected code
2. Additional/Hidden service
3. Service state changing
4. The binary path is hidden
5. Suspicious process
6. Unknown driver attach to a device
7. Unknown module set a timer
8. Callbacks by unknown module

This study is being established in a controlled environment where a few variables are set as constant for each trial. In each trial, the same setting and configuration are used to ensure the retrieval information of memory extraction should produce similar output. However, during the analysis done against the information retrieved, there are differences captured that lead to a pattern. This pattern indicates the abnormalities activities caused by rootkits executed and is considered as a signature. All samples are being analyzed based on eight DBE and signature for each sample, as shown in Table 1 below.

As shown in Table 1, 6 samples are collected (2 clean, 4 benign). The result of clean samples (CS 1 & 2) produces all 0, which indicates that there is no trace of differences in the element checked. While for the malicious samples (MS1, MS2, MS3 & MS4) give a positive response as there is some differences towards the element checked and assign as value 1. Next, the Weightage formula is used to calculate the total for each sample. Minimum total value is set as a benchmark, which is 9, thus if there is another sample tested and resulting total value lower than benchmark consider as clean sample and vice versa.

Table 1: Signature for each sample

DBE	CS 1^2	MS1	MS2	MS3	MS4	W
1. Suspicious/hidden injected code	0	0	1	1	1	**3**
2. Additional/Hidden service	0	0	1	1	1	**3**
3. Service state changing	0	1	1	1	1	**4**
4. Binary path unknown/hidden	0	0	1	1	0	**1**
5. Suspicious process	0	1	0	0	1	**2**
6. Unknown driver attach to device	0	1	0	0	1	**3**
7. Unknown module set timer	0	0	1	0	0	**1**
8. Callbacks by unknown module	0	0	1	1	0	**2**
Total weightage	**0**	**9**	**14**	**13**	**15**	**19**

Indicator CS: Clean (benign) sample MS: Malicious (infected) Sample W: Weightage

4.1 Test for false positive

In this section, a false-positive test is conducted to ensure that the analysis can be applied to other samples. Two malicious samples being chosen to test whether the result may lead to positive or negative feedback. Table 2 shows the result of the unknown sample. As shown in the Total Weightage row, both unknown samples produce a value that higher than the benchmark set. This result is positively shown that both samples indicate they are infected samples. Based on percentage calculation using the aggregator tool develop in Stage 2, unknown sample 1 leads to 72.22%, while unknown sample 2 produces 66.67% rootkits present.

Table 2: Test Result for Unknown Sample

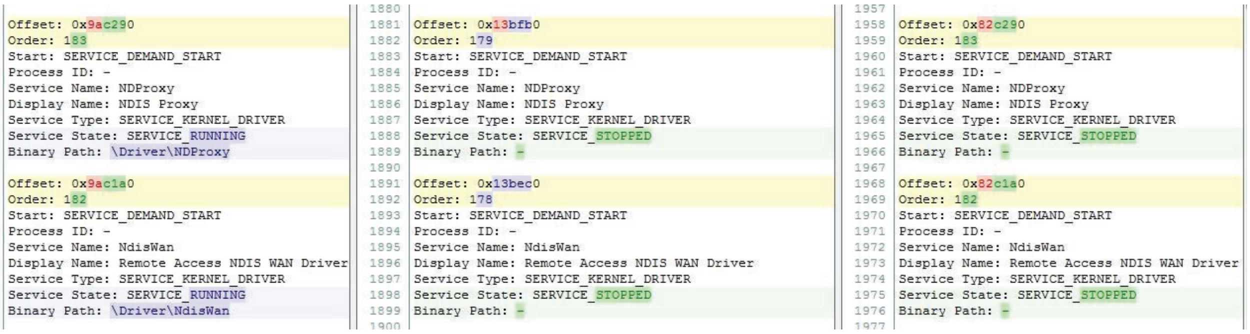
DBE	Unknown Sample 1	Unknown Sample 2
1. Suspicious/hidden injected code	1	0
2. Additional/hidden service	1	1
3. Service state changing	1	1
4. The binary path is hidden	1	0
5. Suspicious process	1	1
6. Unknown driver attach to a device	0	1
7. Unknown module set timer	0	0
8. Callbacks by unknown module	0	0
Total Weightage	**13**	**12**

4.2 Example of Analysis result

The result of the first phase for the first stage is a dataset of DBP that consists of 8 Data Behavior Element. As shown previously, the List of Data Behaviour Element (DBE) is being chosen based on valid differences of benign and malicious samples. Each output is being extracted using the plugin, as mention in Section 3.1 (B). Below are a few examples of analysis results with an explanation.

4.2.1 Service state changing

This is the output of svcscan plugin. Some service state differs from a benign and malicious sample. All four malicious samples show the same behavior. In both Figure 3 and 4, all malicious sample shows they are changing the service state from running to stop. The most left panel is the example of a benign sample that displays the running state of the service. There are many traces of this attribute in the output. Based on the frequency and the consistency of this suspicious trace, it may cause by rootkits infection.

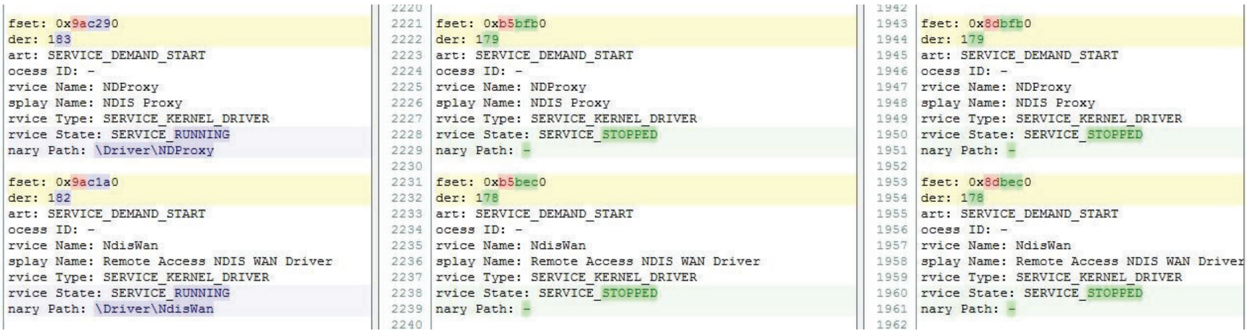


```
Offset: 0x9ac290
Order: 183
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NDProxy
Display Name: NDIS Proxy
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\NDProxy

Offset: 0x9ac1a0
Order: 182
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NdisWan
Display Name: Remote Access NDIS WAN Driver
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\NdisWan
```

```
Offset: 0x13bfb0
Order: 179
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NDProxy
Display Name: NDIS Proxy
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_STOPPED
Binary Path: -

Offset: 0x13bec0
Order: 178
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NdisWan
Display Name: Remote Access NDIS WAN Driver
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_STOPPED
Binary Path: -
```

```
Offset: 0x82c290
Order: 183
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NDProxy
Display Name: NDIS Proxy
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_STOPPED
Binary Path: -

Offset: 0x82c1a0
Order: 182
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: NdisWan
Display Name: Remote Access NDIS WAN Driver
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_STOPPED
Binary Path: -
```

Figure 3: Snapshot of changing service state of Malicious Sample 1 & 2



```
fset: 0x9ac290
der: 183
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NDProxy
splay Name: NDIS Proxy
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_RUNNING
nary Path: \Driver\NDProxy

fset: 0x9ac1a0
der: 182
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NdisWan
splay Name: Remote Access NDIS WAN Driver
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_RUNNING
nary Path: \Driver\NdisWan
```

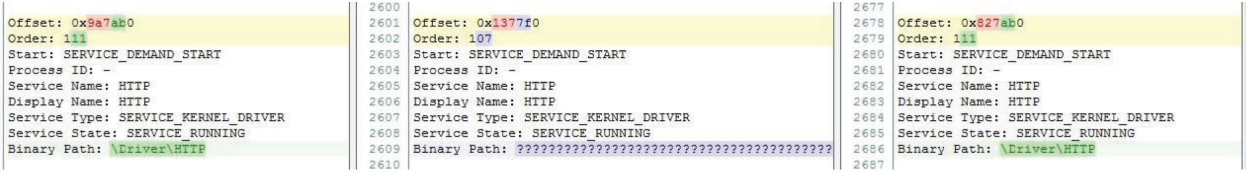
```
fset: 0xb5bfb0
der: 179
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NDProxy
splay Name: NDIS Proxy
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_STOPPED
nary Path: -

fset: 0xb5bec0
der: 178
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NdisWan
splay Name: Remote Access NDIS WAN Driver
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_STOPPED
nary Path: -
```

```
fset: 0x8dbfb0
der: 179
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NDProxy
splay Name: NDIS Proxy
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_STOPPED
nary Path: -

fset: 0x8dbec0
der: 178
art: SERVICE_DEMAND_START
ocess ID: -
rvice Name: NdisWan
splay Name: Remote Access NDIS WAN Driver
rvice Type: SERVICE_KERNEL_DRIVER
rvice State: SERVICE_STOPPED
nary Path: -
```

Figure 4: Snapshot of changing service state of Malicious Sample 3 & 4

4.2.2 Binary path suspicious/hidden

This attribute is from svcscan plugin. Normally the binary path in the service detail is being displayed and normally a valid path. Malicious Sample 3 (MS3) shows this kind of trace in the output, and Malicious Sample 2 shows the point of invalid binary path in one of its service detail.

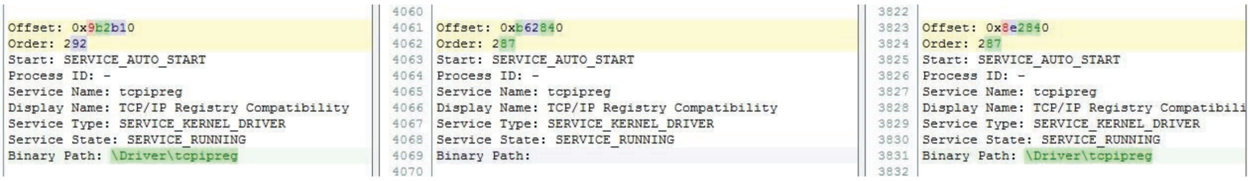


```
Offset: 0x9a7ab0
Order: 111
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: HTTP
Display Name: HTTP
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\HTTP
```

```
Offset: 0x1377f0
Order: 107
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: HTTP
Display Name: HTTP
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: ????????????????????????????????????????
```

```
Offset: 0x827ab0
Order: 111
Start: SERVICE_DEMAND_START
Process ID: -
Service Name: HTTP
Display Name: HTTP
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\HTTP
```

Figure 5: Snapshot of the suspicious binary path of Malicious Sample 2 (MS2)



```
Offset: 0x9b2b10
Order: 292
Start: SERVICE_AUTO_START
Process ID: -
Service Name: tcpipreg
Display Name: TCP/IP Registry Compatibility
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\tcpipreg
```

```
Offset: 0xb62840
Order: 287
Start: SERVICE_AUTO_START
Process ID: -
Service Name: tcpipreg
Display Name: TCP/IP Registry Compatibility
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path:
```

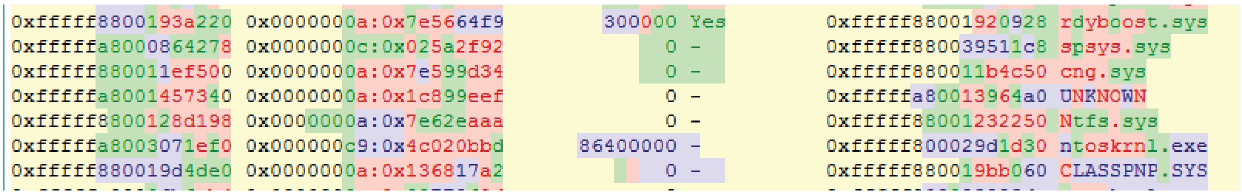
```
Offset: 0x8e2840
Order: 287
Start: SERVICE_AUTO_START
Process ID: -
Service Name: tcpipreg
Display Name: TCP/IP Registry Compatibili
Service Type: SERVICE_KERNEL_DRIVER
Service State: SERVICE_RUNNING
Binary Path: \Driver\tcpipreg
```

Figure 6: Snapshot of the hidden binary path of Malicious Sample 3 (MS3)

In Figure 5, the middle panel is from Malicious sample 2 (MS2), where the binary path displayed is invalid. While in Figure 6 is from Malicious Sample 3 (MS3) in the middle panel that shows the binary path is being hidden. The most left is from benign sample display, the binary path, same as the most right panel, which is the output of malicious sample 4 (MS4), also displays the binary path. Display invalid and hide the binary path is one of a sign of rootkits infection. This sign is consistent and valid.

4.2.3 Unknown module set a timer

The timer plugin can print kernel timers and associated module DPCs. In some cases, a rootkit may use an unknown module to set a timer. This timer is for rootkits to receive a notification that triggered once desired events occur. This can be easily being trace when the modules assign with the timer name as Unknown. Normally, the module should be named with a legal or valid module. One sample from the malicious sample shows the sign for this attribute.



```
0xfffff8800193a220 0x0000000a:0x7e5664f9        300000 Yes        0xfffff88001920928 rdyboost.sys
0xfffffa8000864278 0x0000000c:0x025a2f92             0 -          0xfffff880039511c8 spsys.sys
0xfffff880011ef500 0x0000000a:0x7e599d34             0 -          0xfffff880011b4c50 cng.sys
0xfffffa8001457340 0x0000000a:0x1c899eef             0 -          0xfffffa80013964a0 UNKNOWN
0xfffff880128d198 0x0000000a:0x7e62eaaa             0 -          0xfffff88001232250 Ntfs.sys
0xfffffa8003071ef0 0x000000c9:0x4c020bbd      86400000 -          0xfffff800029d1d30 ntoskrnl.exe
0xfffff880019d4de0 0x0000000a:0x136817a2             0 -          0xfffff880019bb060 CLASSPNP.SYS
```

Figure 7: Snapshot of Unknown module of Malicious Sample 2 (MS2)

Figure 7 above shows the unknown module that creates a timer. A module that only able to create a timer should be one of the valid modules.

5. Challenge and Conclusion

This paper can be used as proof of concept where a data-centric approach is possible to generate a signature and detect malware present, especially rootkits. This is because, to create a dataset in this study, a constant and controlled environment is used as a medium. Besides, this study uses a constant workload that is being executed for each sample to achieve consistent similarities and differences when conducting a comparison between benign and malicious samples. A constant environment is impossible in a real case scenario where different users may execute different workloads.

Overall, we conclude that the data-centric approach can use to detect OS kernel malware. By characterizing the OS kernel malware using Data Behaviour Profile (DBP), two unknown samples were chosen to test the ability of this approach. As a result, the Data Aggregator tool is developed to analyze and calculate the percentage of rootkits presence. The result given is positive. Both unknown samples are proven to have the criteria of rootkits presence based on the data obtained. The objective of this project, which aims to detect OS kernel malware based on trace patterns found in memory dump by using a data-centric approach, is successfully achieved.

6. Future Enhancement

As for future enhancement, it is better to use more samples (benign and malicious) to get more elements for Data Behaviour Element (DBE). With more DBE used in the analysis, the signature created is more valid and accurate. Besides, as for the testing phase, it could be better if the various unknown sample that can be either clean or malicious sample is being used. This is to test the probability of false-positive and true positive that may be triggered. Hence, with more test results, many things can be investigated and improved. Besides, there should be better to provide an appropriate way to determine the malicious sample used in exactly being infected. This is to avoid an issue of samples integrity from arising. Lastly, real-time analysis is being used in implementing the framework. Thus, if this framework can be integrated with machine learning technology, it will make a perfect combo. This is because the real-time analysis can ensure the quality of the framework, and machine learning makes it more proficient.

References

Amari, K. (2009) Techniques and Tools for Recovering and Analyzing Data from Volatile Memory, SANS Institute

Case, A. and Iii, G. (2015) Advancing Mac OS X rootkit detection, Digital Investigation, Vol 14, S25–S33.

Case, A. and Richard, G. (2016) Detecting objective-C malware through memory forensics, Digital Investigation, 18, S3–S10

Davide, B. Marco, C. Christoph, K. Christopher K, Engin, K. and Giovanni, V. (2010) Efficient Detection of Split Personalities in Malware. In Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS'10)

Ford, E. and Howard, M. (2007), How Not to Be Seen, pp 67–69.

Guri, M. and Poliak, Y. (2015) JoKER : Trusted Detection of Kernel Rootkits in Android Devices via JTAG Interface

Halleigh, M. Case, A. Levy, J. and Walters, A. (2014) The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux and Mac Memory, John Wiley & Sons, Inc.

Hili, G. Mayes, K. Markantonakis, K. (2013) The BIOS and Rootkits, pp 369–381.

Lin, Z. Rhee, J. Zhang, X. Xu, D. and Jiang, X. (2011) SigGraph : Brute Force Scanning of Kernel Data Structure Instances Using Graph-based Signatures.

Musavi, S. and Kharrazi, M. (2014) Back to Static Analysis for Kernel-Level Rootkit Detection, Vol 9(9), pp 1465–1476.

Prakash, A. Venkataramani, E. Yin, H. and Lin, Z. (2013) Manipulating Semantic Values in Kernel Data Structures: Attack Assessments and Implications.

Rhee, J. Riley, R. & Lin, Z. (2014) Data-Centric OS Kernel Malware Characterization, 9(1), pp 72–87.

Shahzad, F. Shahzad, M. and Farooq, M. (2013) In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS. Information Sciences, Vol 231, pp 45–63.

Sharif, M. Lanzi, A. Giffin, J. and Lee, W. (2008) Impeding Malware Analysis using conditional code obfuscation, in Proc. 15th Annu. NDSS, pp 1-30

Non Academic Paper

Mitigating Insider Threats: A Case Study of Data Leak Prevention

Nurul Mohd and Zahri Yunos
CyberSecurity Malaysia, Selangor, Malaysia
nurul@cybersecurity.my
zahri@cybersecurity.my

DOI: 10.34190/EWS.20.004

Abstract: Insider threats are considered as one of the most formidable threats that are an ongoing major problem for organizations. An insider threat is generally defined as a threat to a network or computer system that originates from a person with authorized access who misuses that access to compromise information, disrupt operations or cause physical harm, thus negatively impacting an organization. Insider threats are influenced by a combination of technical, behavioural and organizational issues. Information loss on account of insiders remains among the greatest threats to organizations, since data is now viewed as an important corporate asset. Insider threats not adequately addressed can cause harm to organizations. In this paper, a Data Leak Prevention (DLP) management workflow for mitigating insider threats is proposed. The workflow, derived from an internal CSIRT, describes practices and measures that are currently being implemented by CyberSecurity Malaysia to prevent the illicit transfer of confidential or sensitive information or data outside the corporate network. This paper provides a case study of how CyberSecurity Malaysia identifies and detects either intentional or unintentional insider data leakage incidents with focus on e-mail messages and social media applications, and how these incidents are managed according to the defined incident management process. This paper also highlights the ways in which the Information Security Management System (ISMS) has assisted CyberSecurity Malaysia with implementing the DLP workflow by incorporating the ISMS requirements. Comprehensive DLP policies and procedures ensure that all employees conform, thus speeding up the implementation process effectively.

Keywords: Data Leak Prevention, Data Breach, Incident Management, Insider Threat

1. Introduction

Insider threats are regarded as one of the most remarkable threats posing a major problem for organizations. Everyday data leaks are making headlines more often than the nefarious attack scenarios around which organizations plan most, if not all, of their data leakage prevention methods. Data volumes are also growing exponentially, dramatically increasing opportunities for theft and accidental disclosure of sensitive information. Information loss attributable to insiders is still one of the greatest threats to organizations, wherein the significance of data as a corporate asset is now recognized. The current use of many additional forms of communication within organizations further contributes to the emergence of more avenues for data leakage.

According to a survey conducted for the 2019 Insider Threat Report, 63% of the respondents answered that customer data is the most vulnerable data for an organization; 55% said that intellectual property is the most vulnerable and the remaining 45% opted for financial data (*Why organizations feel vulnerable to insider attacks - TechRepublic*, 2019). Insider threats that are not addressed properly be detrimental to an entire organization. Some examples of data leakage trends are weak and stolen credentials, application vulnerabilities, malware, insider threats and physical attacks.

2. Literature Review

2.1 Various Case Studies

Malindo Air, a low-cost Malaysian airline, reported a significant data breach that affected millions of passengers. Passenger information, including names, home addresses, phone numbers and passport numbers were leaked on public forums. Those affected have consequently been more likely to be at risk of identity theft and fraud. Malindo Air also admitted that two former employees of a contractor firm inappropriately accessed and stole information (*All Data Breaches in 2019 - An Alarming Timeline - SelfKey*, 2020).

Another incident happened at Desjardins Group, a Canadian bank. The respective data breach affected around 2.7 million people and roughly 173,000 companies. According to Desjardins, it was an employee who gathered information about the bank’s customers and sold it to a third party. The CEO mentioned that no individual in the bank has access to its members’ information. The employee must have used his privileged access along with other employees’ privileged access to pull off this crime (*A bank, an insider, and 2.7 million people’s data stolen*, 2019).

*Lowyat.*net, a famous local portal, reported that the personal data of 46.2 million mobile subscribers was compromised and was being sold online. The leaked data includes mobile numbers, unique serial phone numbers and home addresses. The portal was also informed that the huge database comprised personal data of at least 12 million Malaysian mobile users. The information was to be sold for an undisclosed amount of Bitcoin (*46.2 Million Malaysian Mobile Phone Numbers Leaked from 2014 Data Breach, 2017).*

A different data breach incident involved leaked data of 1,164,540 students and alumni enrolled between the years 2000 and 2018. The leaked data contained students' names, MyKad numbers, home addresses, e-mail addresses, campus codes and names, program codes, course levels, student IDs and mobile numbers (*Anonymous individual threatens to leak UiTM student data if demand not met | The Star Online*, 2019).

2.2 Understanding Insider Threats

In general, insiders are authorized users who have legitimate access to sensitive/confidential material and who may know the vulnerabilities of the systems and business processes deployed (Ivan Homoliak et al., 2019). An insider can be thought of as an individual who is an employee (past or present), contractor or other trusted third party with privileged access to the networks, systems or data of an organization (Silowash and King, 2013).

The CERT Insider Threat Center (Cappelli, Moore and Trzeciak, 2012) defines insider threats as follows:

> *"A malicious insider threat to an organization is a current or former employee, contractor, or other business partner who has or had authorized access to an organization's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems."*

Motivations behind malicious insiders could be related to monetary gain, a disgruntled employee, entitlement, ideology, or outside influence with the consequences of fraud, sabotage, espionage and theft or loss of confidential information (Guerrino Mazzarolo and Anca Delia Jurcut, 2019).

Based on an article by the cybersecurity defense team of KPMG (*What's next: Trends in insider threat*, 2019), the following can be influences for insider threats:

1. Foreign intelligence agencies
2. Political or social involvement
3. Personal financial issues
4. Unsatisfied employees
5. Fear of being sacked

2.3 Data leak prevention

Numerous issues that arise trigger organizations to implement data leak prevention mechanisms. The issues are more alarming when internal threat actors are involved, who most of the time have the least control but the most accessibility to sensitive information. Some common problems that lead to data leakage through insider threats are as follows:

1. Removable media such as universal serial bus (USB) flash drives may present an enterprise with unique problems. Insiders can use such media to remove proprietary information from company systems.
2. Traditional controls like information security policies and procedures as well as conventional security mechanisms such as firewall, VPN and IDS might lack user proactivity. In most cases, such controls require predefined rules according to which protection actions are to be taken. This can result in serious consequences, since confidential data can appear in different forms on various leaking channels.

Organizations have diverse objectives for implementing DLP. Many deploy DLP technologies to counter the risk of both deliberate and accidental disclosure of information. Data leakage prevention systems have also been introduced as dedicated mechanisms to detect and prevent the leakage of confidential data in use, in transit and at rest. DLP involves different techniques of analysing the content and context of confidential data to detect and prevent leakage. Besides, many mechanisms are implemented to prevent certain data from leaving the corporate perimeter based on predefined rules.

3. Data Leakage Prevention Implementation: A Case Study

DLP is implemented by CyberSecurity Malaysia in 2 phases: the deployment of hardware and software in the organization infrastructure, and the development of policies that are to become input for the system.

Figure 1 illustrates DLP placement and implementation considerations.

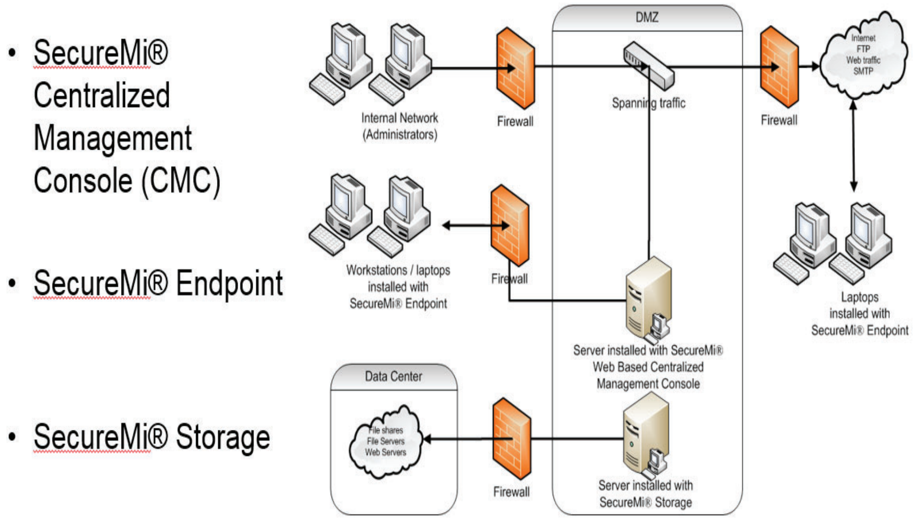

• SecureMi® Centralized Management Console (CMC)
• SecureMi® Endpoint
• SecureMi® Storage
DMZ
Internal Network (Administrators)
Firewall
Spanning traffic
Firewall
Internet
FTP
Web traffic
SMTP
Workstations / laptops installed with SecureMi® Endpoint
Firewall
Laptops installed with SecureMi® Endpoint
Server installed with SecureMi® Web Based Centralized Management Console
Data Center
File shares
File Servers
Web Servers
Firewall
Server installed with SecureMi® Storage

Figure 1: DLP architecture

The following 3 processes are executed in the first phase:

1. Installation by end users - Each employee is responsible for ensuring that CyberSecurity Malaysia's data and information are protected from loss or misuse. Thus, each employee is required to install SecureMi® Endpoint software on their notebook and computer.
2. Registration of portable storage devices - Any files transferred between end users' notebooks/computers and portable storage devices are monitored by the DLP system and any violation of DLP rules will lead to blocking. Only registered portable storage devices can be used to transfer data. End users are not allowed to copy or transfer data through unregistered portable storage devices in the CyberSecurity Malaysia environment. As such, all end users are required to register their portable storage devices with the Secure Technology System Department. Devices are limited to maximum 5 units.
3. Registration of a personal printer – End users may use personal printers to print work documents that contain classified information. Each printer needs to be registered with the STS department as well before it may be used for printing.

In the second phase, the policy intended for becoming input to the system is developed. This is the most crucial phase, as the system will only act as the executer and function only based on predefined rules.

The roles designated for DLP implementation are a steering committee, project manager, DLP policy advisor and incident handler. The DLP steering committee is a taskforce established mainly to discuss and decide any matters regarding the operation and strategic aspect of DLP at CyberSecurity Malaysia. The committee consists of a manager, project manager, DLP policy advisor and DLP incident handler.

For the policy development stage, the steering committee is in a position to decide which data classifications need to be safeguarded. The committee has decided on 3 classifications to be included in the scope, which are Top Secret, Secret and Confidential.

The ISMS driver is appointed DLP policy advisor, whose main role is to assist with identifying the organization's documents that need to be safeguarded according to their respective classifications. As the one overseeing and coordinating the organization's ISMS, this person is also responsible for ensuring that DLP implementation is in line with the organization's existing policies and procedures. The DLP policy advisor is additionally responsible for developing DLP guidelines and working with the DLP incident handler on DLP rule fine tuning and other process-related matters. For the DLP policy development process, the project manager compiles all documents as well as the key words predefined by the department's document controller that meet the classification requirements.

Once the policy development process is completed, a procedure on incident handling needs to be established. Figure 2 shows the high-level incident handling process workflow that enables taking appropriate corrective action when data leakage is detected.

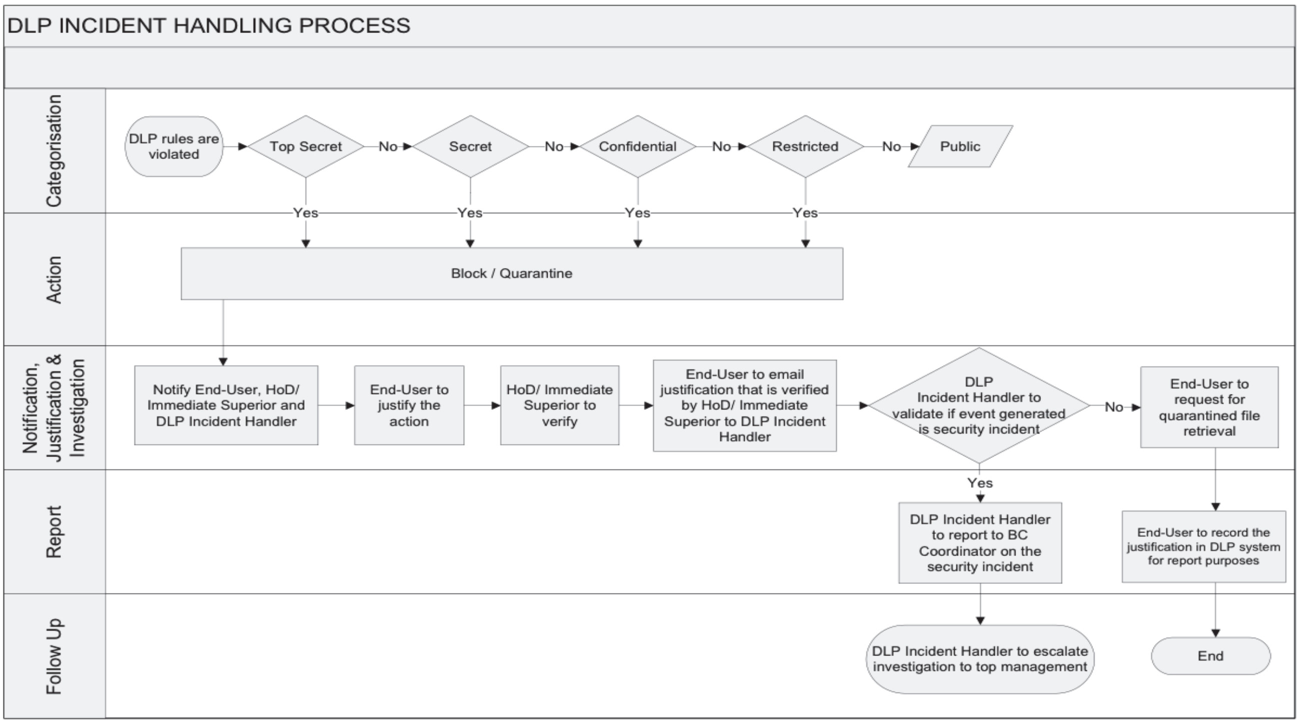
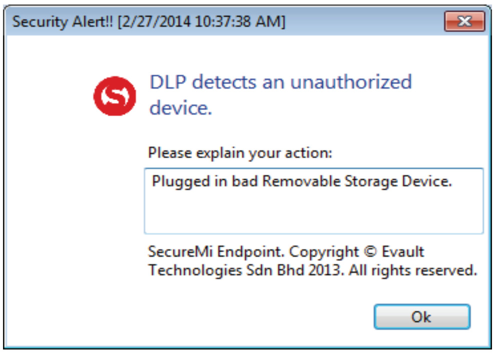

DLP INCIDENT HANDLING PROCESS
Categorisation
Action
Notification, Justification & Investigation
Report
Follow Up
DLP rules are violated
Top Secret
No
Secret
No
Confidential
No
Restricted
No
Public
Yes
Yes
Yes
Yes
Block / Quarantine
Notify End-User, HoD/ Immediate Superior and DLP Incident Handler
End-User to justify the action
HoD/ Immediate Superior to verify
End-User to email justification that is verified by HoD/ Immediate Superior to DLP Incident Handler
DLP Incident Handler to validate if event generated is security incident
No
End-User to request for quarantined file retrieval
Yes
DLP Incident Handler to report to BC Coordinator on the security incident
End-User to record the justification in DLP system for report purposes
DLP Incident Handler to escalate investigation to top management
End

Figure 2: DLP Incident handling process

Details of the activities related to the DLP incident handling processes are given in Table 1.

Table 1: DLP incident handling process activities

Steps	Activities
1) Categorization	• The DLP system monitors end-user actions and analyses the content to determine if any DLP rule violation has occurred. Once a data protection violation is detected, an event is generated that describes the violation. The DLP system categorizes the DLP incident based on specified data classification (top secret, secret, confidential, or restricted).
2) Action	• The DLP system will deliver the action based on the predefined rules according to the data classification. An end-user action that violates the rules will be blocked. If a file is violated on a portable storage device, it is blocked/quarantined and removed from the respective device.
3) Notification & investigation	• Apart from notifying the end user who violated the DLP rules, the DLP system also sends a notification of the event to the end user's head of department (HoD) or immediate superior. • The end user needs to justify the action, which the HoD/immediate superior will verify. • After verification, the end user e-mails the justification to the DLP incident handler.

Steps	Activities
	• If according to verification the action is determined to have been legitimate, the end user can request an OTP (one-time password) from the DLP administrator to recover the quarantined files. Please refer to section 6.1: Quarantine File Retrieval for details.
4) Report	• If the event notification is verified as a valid data leakage incident, the DLP incident handler reports it to the BC coordinator. The BC coordinator will address the incident in accordance with the Incident and Crisis Management Procedure. • A pop-up message is prompted whenever the DLP system detects any violation of the DLP rules. The end user must provide an explanation of the action performed for records.  Security Alert!! [2/27/2014 10:37:38 AM] DLP detects an unauthorized device. Please explain your action: Plugged in bad Removable Storage Device. SecureMi Endpoint. Copyright © Evault Technologies Sdn Bhd 2013. All rights reserved. • The DLP System also sends a notification to the DLP incident handler to validate the event generated for clarification. The DLP incident handler will verify and confirm whether the incident is a false negative or false positive. If the event is identified as false positive, the DLP incident handler needs to liaise with the DLP policy advisor to review and fine tune the DLP rules if necessary.
5) Follow-up	• If the reported event is verified as a valid data leakage incident, the DLP incident handler will escalate the report to the top management for security incident handling.

4. Data Classification in DLP Implementation

CyberSecurity Malaysia is a certified ISO 27001:2005 (Information Security Management System) organization since 2008. This achievement has contributed to the successful DLP implementation at CyberSecurity Malaysia. The Standard defines 114 requirements to which the organization must comply. Most clauses pertain to controls that should be implemented to safeguard the confidentiality, integrity and availability of information.

Clause A.8.2 ISO 27001:2005 states:

> *"Data shall be handled according to its classification when created, during transmission and while in storage."*

Data classification is the process of organizing data into categories or levels for effective and efficient use. This definition implies that a DLP system relies entirely on well-defined data classification that enables the system to differentiate confidential from normal data. The main purpose of classifying data is to determine the baseline security controls to be used for safeguarding data. CyberSecurity Malaysia employs 5 data classifications levels: Top Secret, Secret, Confidential, Restricted, and Public. This classification eases identifying specifically confidential data, thereby equipping the system adequately to protect the respective confidential data.

Classification criteria are guided by the ISMS Information Classification Policy & Procedure. ISMS has been of significant assistance to CyberSecurity Malaysia in the successful implementation of the DLP workflow by incorporating the ISMS requirements including information classification and handling in the DLP policies and procedures. A comprehensive set of DLP policies and procedures ensures that all employees conform, thus expediting the implementation process effectively.

However, the problem with data classification is in determining the level of secrecy of sensitive data. To achieve proper classification of secrecy levels, the owner of the data to be protected should be the one responsible for this classification process. However, most of the time the classification process is left to people who lack sufficient knowledge about all data. This creates a vulnerability in the DLP system, as those who are not permitted to see certain information are now equipped with access to confidential information. It is therefore important to have good classification in place, because the lack thereof renders the DLP system ineffective.

5. Conclusion

Organizations are now facing more complex, evolving security threats like data leakage. The insider threat is a persistent security thorn that puts firms at risk. Unless adequately addressed, insider threats can cause significant harm to organizations. The extensive range of communication means used within organizations is one factor that has led to the increasing emergence of additional data leakage possibilities, which need to be managed prudently.

The case studies presented in this paper highlight DLP implementation as an effective mechanism for organizations to classify data and information. It is evident that if data leakage is not managed properly, catastrophic situations may ensue. Effective management is attainable with an efficient DLP mechanism and incident handling procedure. Incorporating ISMS requirements such as information classification and handling helps reduce the implementation time and makes the policy more comprehensive and effective.

The successful implementation of DLP by organizations must involve the participation of the stakeholders. The stakeholders' role implies their responsibility to manage DLP implementation and incidents related to DLP. Appropriate causal analysis and post-incident analysis are important for an organization in determining the cause of incidents and applying necessary measures to not undergo new or repeated incidents or mistakes.

References

A bank, an insider, and 2.7 million people's data stolen (July 2019). Available at: https://www.pandasecurity.com/mediacenter/news/dejsardins-insider-data-breach/ (Accessed: 8 January 2020)

All Data Breaches in 2019 - An Alarming Timeline - SelfKey (January 2020). Available at: https://selfkey.org/data-breaches-in-2019/ (Accessed: 8 January 2020)

Alneyadi, S., Sithirasenan, E. and Muthukkumarasamy, V. (2016) 'A survey on data leakage prevention systems', *Journal of Network and Computer Applications*. Academic Press, 62, pp. 137–152. doi: 10.1016/j.jnca.2016.01.008

Anonymous individual threatens to leak UiTM student data if demand not met | The Star Online (Jan 2019). Available at: https://www.thestar.com.my/tech/tech-news/2019/01/30/uitm-threatened-to-improve-websites-or-face-leak (Accessed: 13 January 2020)

Cappelli, D., Moore, A. and Trzeciak, R. (2012) *The CERT Guide to Insider threats*, *Books.Google.Com*. Available at: http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract (Accessed: 7 January 2020).

Cybersecurity Challenges and Solutions | Toptal (no date). Available at: https://www.toptal.com/finance/finance-directors/cyber-security (Accessed: 7 January 2020).

Guerrino Mazzarolo, Anca Delia Jurcut (2019) 'Insider threats in Cyber Security: The enemy within the gates', *European CyberSecurity Journal*

Ivan Homiliak et al (April 2019) ' Insight Into Insiders and IT: A Survey of Insider Threat Taxonomies, Analysis, Modelling, and Countermeasures' *ACM Computing Surveys (CSUR)* Article No.: 30 https://doi.org/10.1145/3303771

Silowash, G. and King, C. (2013) 'Insider threat control: Understanding data loss prevention (DLP) and detection by correlating events from multiple sources', (January), p. 30. Available at: http://repository.cmu.edu/sei/708/

Stamati-Koromina, V. *et al.* (2012) 'Insider threats in corporate environments: A case study for data leakage prevention', *ACM International Conference Proceeding Series*, (December 2014), pp. 271–274. doi: 10.1145/2371316.2371374

What's next: trends in insider threat (2019). Available at: https://advisory.kpmg.us/articles/2019/trends-in-insider-threat.html (Accessed: 9 January 2020)

Why organizations feel vulnerable to insider attacks - TechRepublic (October 2019). Available at: https://www.techrepublic.com/article/why-organizations-feel-vulnerable-to-insider-attacks/ (Accessed: 8 January 2020)

46.2 Million Malaysian Mobile Phone Numbers Leaked From 2014 Data Breach (October 2017). Available at: https://www.lowyat.net/2017/146339/46-2-million-mobile-phone-numbers-leaked-from-2014-data-breach/ (Accessed: 8 January 2020)

Work in Progress Papers

Modern Techniques for Discovering Digital Steganography

Michael T. Hegarty and Anthony J. Keane
TU-Dublin, Blanchardstown, Ireland
Michael.hegarty@tudublin.ie
Anthony.keane@tudublin.ie

DOI: 10.34190/EWS.20.504

Abstract: Digital steganography can be difficult to detect and as such is an ideal way of engaging in covert communications across the Internet. This research paper is a work-in-progress report on instances of steganography that were identified on websites on the Internet including some from the DarkWeb using the application of new methods of deep learning algorithms. This approach to the identification of Least Significant Bit (LSB) Steganography using Convolutional Neural Networks (CNN) has demonstrated some efficiency for image classification. The CNN algorithm was trained using datasets of images with known steganography and then applied to datasets with images to identify concealed data. The algorithm was trained using 5000 clean images and 5000 Steganography images. With the correct configurations made to the deep learning algorithms, positive results were obtained demonstrating a greater speed, accuracy and fewer false positives than the current steganalysis tools.

Keywords: Steganography, Deep Learning, JPEG, LSB, Darknet, Openpuff

1. Introduction

Steganography can be explained as the art and science of writing hidden messages. Using this method no one apart from the sender and intended recipient suspects the existence of the message. In the modern digital world, Steganography methods are many as information can be hidden not only within graph images but also within other digital carriers such as video and audio (Kessler, 2015).

According to (Zielińska *et al,* 2014), technology makes it easy to perform digitalized methods of concealing information with many of the applications available for free from internet portals. Currently, there are approximately 1200 tools available to create different forms of steganography (Hegarty, 2018). These tools can be easily used by criminals to communicate undetected and challenges the cyber security specialists to establish if covert communication has occurred. Digital evidence of communication may never be detected on a suspect's machine due to it being concealed within a carrier file (Hunt, 2012). Richer (2015) suggests that Steganography can be used to communicate with complete freedom under conditions that are monitored.

While there is the suspicion that terror groups are using technologies like Steganography to help conceal their communications on the Internet, there is little solid evidence to support that claim. In May 2011 a 22-year-old Austrian citizen Maqsood Lodin was stopped and questioned in Germany. He was traveling from Pakistan to Berlin via Hungary. During a search, a memory stick was found in his underpants. The memory stick contained two pornographic videos. Over 100 documents were concealed within data on the storage device. They contained secret documents about operational details and training manuals for al-Qaeda. Its source came from a German Magazine "Die Zeit". The journalist and author of the article Yassin Musharbash has published an investigative report on his personal blog that includes a summary of the findings (Musharbash, 2012).

There are examples of the use of Steganography in industrial espionage when in 2018, Zheng Xiaoqing and Zhang Zhaoxi were both indicted for their role in stealing proprietary trade secret information related to GE's turbine technology. The indictment identified highly sophisticated steganography measures deployed to hide the stolen GE trade secrets in JPeg images files and sending them to a personal Hotmail account (Lynch and Shepardson, 2019).

2. Digital Steganalysis

Steganalysis involves examining several parameters of media type for Steganography to check for hidden data. Statistical Steganalysis is when alterations are applied to a carrier file, there is a change in the statistical information such as changes made to the LSB of the original image file in order to hide information (Ker *et al*., 2013).

The detection of files with hidden data required that the servers/applications that store the file do not change the files or the secret payload can be destroyed. Ebay does this to files to fit its requirements and renders the files useless for retrieving hidden information. Web-portals on the clear and darknet were identified as places that steganography could survive as the files are stored in original formats. Some of these web-portals that were researched are:

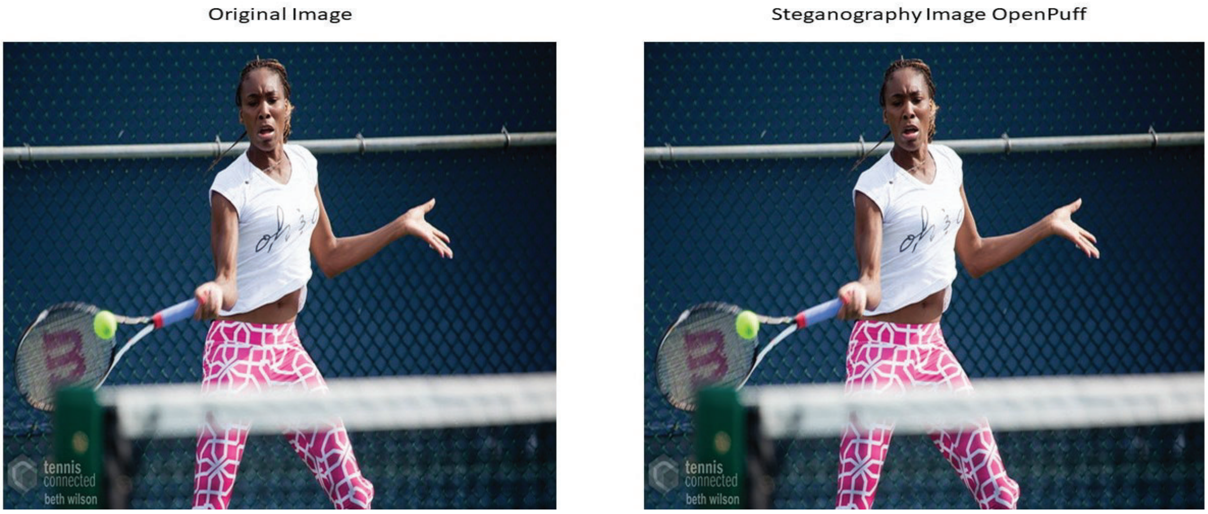
- Draugiem: The main social media platform in Latvia with 2.6 million users
- VKontakte (VK): Russia's largest social media platform claiming to have 500 million registered accounts
- ISIS Darknet: wpcxzq4ykmsxpacm.onion a propaganda outlet for Islamic state on the darknet

A dataset of 3 million images was harvested from clear-net and dark-net webportals. This data was analysed using existing steganalysis tools such as Stegdetect, Stegexpose, Outguess, and Virtual Steganographic Laboratory (VSL). Steganalysis tools will return a probability and categorise findings. Highly probable steganography images were then further analysed. Of the 3 million images analysed 0.02% (600 images) returned as suspicious of concealing data.

3. Machine Learning for Steganalysis

According to (Giarimpampa, 2018), *"Neural Networks have proven to be a useful tool in image Steganalysis".* The use of both Support Vector Machine (SVM) and Artificial Neural Networks (ANN) algorithms have become the predominant modern Steganalysis trend.

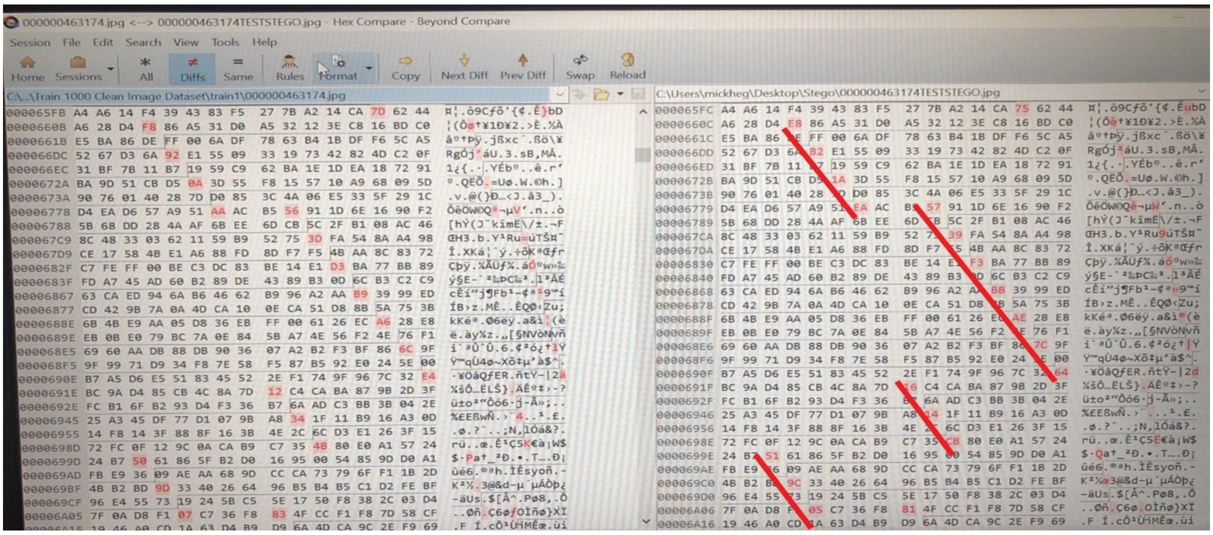
Using Python, Anaconda, on the Spyder IDE, 5000 clean images with no concealed data was developed. 5000 steganography images were created using OpenPuff steganography tool. OpenPuff is the most downloaded steganography tool according to Softpedia with 60,352 downloads as of September 2019. Figure 1 displays an original image on the left and the carrier image (with a secret massage) on the right, this secret message was created using OpenPuff steganography software. Figure 2 presents a sample of the HEX code and the changes OpenPuff has created to hide the secret message in the LSB.



Original Image

Steganography Image OpenPuff

Figure 1: Sample original image and same image with hidden message


000000463174.jpg <--> 000000463174TESTSTEGO.jpg - Hex Compare - Beyond Compare
Session File Edit Search View Tools Help
Home Sessions All Diffs Same Rules Format Copy Next Diff Prev Diff Swap Reload
C:\...\Train 1000 Clean Image Dataset\train1\000000463174.jpg
C:\Users\mickheg\Desktop\Stego\000000463174TESTSTEGO.jpg

```
000065FB A4 A6 14 F4 39 43 83 F5  27 7B A2 14 CA 7D 62 44  ¤¦.ô9Cƒõ'{¢.Ê}bD
0000660B A6 28 D4 F8 86 A5 31 D0  A5 32 12 3E C8 16 BD C0  ¦(Ôø†¥1Ð¥2.>È.½À
0000661B E5 BA 86 DE FF 00 6A DF  78 63 B4 1B DF F6 5C A5  åº†Þÿ.jßxc´.ßö\¥
000066DC 52 67 D3 6A 92 E1 55 09  33 19 73 42 82 4D C2 0F  RgÓj'áU.3.sB,MÂ.
000066EC 31 BF 7B 11 B7 19 59 C9  62 BA 1E 1D EA 18 72 91  1¿{.·.YÉbº..ê.r'
0000672A BA 9D 51 CB D5 0A 3D 55  F8 15 57 10 A9 68 09 5D  º.QËÕ.=Uø.W.©h.]
0000673A 90 76 01 40 28 7D D0 85  3C 4A 06 E5 33 5F 29 1C  .v.@(}Ð…<J.å3_).
00006778 D4 EA D6 57 A9 51 AA AC  B5 56 91 1D 6E 16 90 F2  ÔêÖW©Qª¬µV'.n..ò
00006788 5B 68 DD 28 4A AF 6B EE  6D CB 5C 2F B1 08 AC 46  [hÝ(J¯kîmË\/±.¬F
000067C9 8C 48 33 03 62 11 59 B9  52 75 3D FA 54 8A A4 98  ŒH3.b.Y¹Ru=úTŠ¤˜
000067D9 CE 17 58 4B E1 A6 88 FD  8D F7 F5 4B AA 8C 83 72  Î.XKá¦ˆý.÷õKªŒƒr
0000682F C7 FE FF 00 BE C3 DC 83  BE 14 E1 D3 BA 77 BB 89  Çþÿ.¾ÃÜƒ¾.áÓºw»‰
0000683F FD A7 45 AD 60 B2 89 DE  43 89 B3 0D 6C B3 C2 C9  ý§E-`²‰ÞC‰³.l³ÂÉ
00006867 63 CA ED 94 6A B6 46 62  B9 96 A2 AA B9 39 99 ED  cÊí"j¶Fb¹–¢ª¹9™í
00006877 CD 42 9B 7A 0A 4D CA 10  0E CA 51 D8 8B 5A 75 3B  ÍB›z.MÊ..ÊQØ‹Zu;
0000688E 6B 4B E9 AA 05 D8 36 EB  FF 00 61 26 EC A6 28 E8  kKéª.Ø6ëÿ.a&ì¦(è
0000689E EB 0B E0 79 BC 7A 0E 84  5B A7 4E 56 F2 4E 76 F1  ë.ày¼z.„[§NVòNvñ
000068E5 69 60 AA DB 88 DB 90 36  07 A2 B2 F3 BF 86 6C 9F  i`ªÛˆÛ.6.¢²ó¿†lŸ
000068F5 9F 99 71 D9 34 F8 7E 58  F5 87 B5 92 E0 24 5E 00  Ÿ™qÙ4ø~Xõ‡µ'à$^.
0000690E B7 A5 D6 E5 51 83 45 52  2E F1 74 9F 96 7C 32 E4  ·¥ÖåQƒER.ñtŸ–|2ä
0000691E BC 9A D4 85 CB 4C 8A 7D  12 C4 CA BA 87 9B 2D 3F  ¼šÔ…ËLŠ}.ÄÊº‡›-?
0000692E FC B1 6F B2 93 D4 F3 36  B7 6A AD C3 BB 3B 04 2E  ü±o²"Ôó6·j-Ã»;..
00006945 25 A3 45 DF 77 D1 07 9B  A8 34 1F 11 B9 16 A3 0D  %£EßwÑ.›¨4..¹.£.
00006955 14 F8 14 3F 88 8F 16 3B  4E 2C 6C D3 E1 26 3F 15  .ø.?ˆ..;N,lÓá&?.
0000698D 72 FC 0F 12 9C 0A CA B9  C7 35 4B 80 E0 A1 57 24  rü..œ.Ê¹Ç5K€à¡W$
0000699D 24 B7 50 61 86 5F B2 D0  16 95 00 54 85 9D D0 A1  $·Pa†_²Ð.•.T….Ð¡
000069AD FB E9 36 09 AE AA 68 9D  CC CA 73 79 6F F1 1B 2D  ûé6.®ªh.ÌÊsyoñ.-
000069BF 4B B2 BD 9D 33 40 26 64  96 B5 B4 B5 C1 D2 FE BF  K²½.3@&d–µ´µÁÒþ¿
000069CF 96 E4 55 73 19 24 5B C5  5E 17 50 F8 38 2C 03 D4  –äUs.$[Å^.Pø8,.Ô
00006A05 7F 0A D8 F1 07 C7 36 F8  83 4F CC F1 F8 7D 58 CF  ..Øñ.Ç6øƒOÌñø}XÏ
00006A15 19 46 A0 CD 1A 63 D4 B9  D9 6A 4D CA 9C 2E F9 69  .F Í.cÔ¹Ù¡MÊœ.ùi
```

```
000065FC A4 A6 14 F4 39 43 83 F5  27 7B A2 14 CA 75 62 44  ¤¦.ô9Cƒõ'{¢.ÊubD
0000660C A6 28 D4 E8 86 A5 31 D0  A5 32 12 3E C8 16 BD C0  ¦(Ôè†¥1Ð¥2.>È.½À
0000661C E5 BA 86 DE FF 00 6A DF  78 63 B4 1B DF F6 5C A5  åº†Þÿ.jßxc´.ßö\¥
000066DD 52 67 D3 6A B2 E1 55 09  33 19 73 42 82 4D C2 0F  RgÓj²áU.3.sB,MÂ.
000066ED 31 BF 7B 11 B7 19 59 C9  62 BA 1E 1D EA 18 72 91  1¿{.·.YÉbº..ê.r'
0000672B BA 9D 51 CB D5 1A 3D 55  F8 15 57 10 A9 68 09 5D  º.QËÕ.=Uø.W.©h.]
0000673B 90 76 01 40 28 7D D0 85  3C 4A 06 E5 33 5F 29 1C  .v.@(}Ð…<J.å3_).
00006779 D4 EA D6 57 A9 51 EA AC  B5 57 91 1D 6E 16 90 F2  ÔêÖW©Qê¬µW'.n..ò
00006789 5B 68 DD 28 4A AF 6B EE  6D CB 5C 2F B1 08 AC 46  [hÝ(J¯kîmË\/±.¬F
000067CA 8C 48 33 03 62 11 59 B9  52 75 39 FA 54 8A A4 98  ŒH3.b.Y¹Ru9úTŠ¤˜
000067DA CE 17 58 4B E1 A6 88 FD  8D F7 F5 4B AA 8C 83 72  Î.XKá¦ˆý.÷õKªŒƒr
00006830 C7 FE FF 00 BE C3 DC 83  BE 14 E1 F3 BA 77 BB 89  Çþÿ.¾ÃÜƒ¾.áóºw»‰
00006840 FD A7 45 AD 60 B2 89 DE  43 89 B3 0D 6C B3 C2 C9  ý§E-`²‰ÞC‰³.l³ÂÉ
00006868 63 CA ED 94 6A B6 46 62  B9 96 A2 AA BB 39 99 ED  cÊí"j¶Fb¹–¢ª»9™í
00006878 CD 42 9B 7A 0A 4D CA 10  0E CA 51 D8 8B 5A 75 3B  ÍB›z.MÊ..ÊQØ‹Zu;
0000688F 6B 4B E9 AA 05 D8 36 EB  FF 00 61 26 EC AE 28 E8  kKéª.Ø6ëÿ.a&ì®(è
0000689F EB 0B E0 79 BC 7A 0E 84  5B A7 4E 56 F2 4E 76 F1  ë.ày¼z.„[§NVòNvñ
000068E6 69 60 AA DB 88 DB 90 36  07 A2 B2 F3 BF 86 7C 9F  i`ªÛˆÛ.6.¢²ó¿†|Ÿ
000068F6 9F 99 71 D9 34 F8 7E 58  F5 87 B5 92 E0 24 5E 00  Ÿ™qÙ4ø~Xõ‡µ'à$^.
0000690F B7 A5 D6 E5 51 83 45 52  2E F1 74 9F 96 7C 32 64  ·¥ÖåQƒER.ñtŸ–|2d
0000691F BC 9A D4 85 CB 4C 8A 7D  16 C4 CA BA 87 9B 2D 3F  ¼šÔ…ËLŠ}.ÄÊº‡›-?
0000692F FC B1 6F B2 93 D4 F3 36  B7 6A AD C3 BB 3B 04 2E  ü±o²"Ôó6·j-Ã»;..
00006946 25 A3 45 DF 77 D1 07 9B  A8 14 1F 11 B9 16 A3 0D  %£EßwÑ.›¨...¹.£.
00006956 14 F8 14 3F 88 8F 16 3B  4E 2C 6C D3 E1 26 3F 15  .ø.?ˆ..;N,lÓá&?.
0000698E 72 FC 0F 12 9C 0A CA B9  C7 35 CB 80 E0 A1 57 24  rü..œ.Ê¹Ç5Ë€à¡W$
0000699E 24 B7 51 61 86 5F B2 D0  16 95 00 54 85 9D D0 A1  $·Qa†_²Ð.•.T….Ð¡
000069AE FB E9 36 09 AE AA 68 9D  CC CA 73 79 6F F1 1B 2D  ûé6.®ªh.ÌÊsyoñ.-
000069C0 4B B2 BD 9C 33 40 26 64  96 B5 B4 B5 C1 D2 FE BF  K²½œ3@&d–µ´µÁÒþ¿
000069D0 96 E4 55 73 19 24 5B C5  5E 17 50 F8 38 2C 03 D4  –äUs.$[Å^.Pø8,.Ô
00006A06 7F 0A D8 F1 05 C7 36 F8  81 4F CC F1 F8 7D 58 CF  ..Øñ.Ç6ø.OÌñø}XÏ
00006A16 19 46 A0 CD 1A 63 D4 B9  D9 6A 4D CA 9C 2E F9 69  .F Í.cÔ¹Ù¡MÊœ.ùi
```

Figure 2: HEX code of original image and HEX code of image with hidden message

The Deep Learning algorithm Convolutional Neural Networks (CNN) was deployed to train and test the image datasets. The algorithm has been used for steganalysis purposes because of efficiency in image classification (Liu., Yang, and Kang, 2017).

The CNN algorithm Python code is designed to be able to perform image classification on the datasets being trained and tested. The convolution code will extract the features and learn the image feature matrix also known as the feature map. The Pooling code will reduce the dimensions of each feature map but keep the most important data. The Flattening code will convert the matrix into a linear array and Fully Connecting code will connect the convolutional network to the neural network and then compile the network. The Epoch code is set to 100 which means the dataset will pass through the neural network 100 times which will result in an increase in accuracy and a decrease in loss .

The following metrics were used:

- The **Training Accuracy** represents the accuracy that is applied to the CNN model on the training data function as a percentage and how well the CNN model was trained.
- The **Training Loss** represents the error on the training set of that data. A decrease in the Training Loss means the CNN model is improving.
- The **Validation Accuracy** is used after the CNN has been trained for adjusting the networks hyper-parameters and comparing how the changes to them affect the predictive accuracy of the CNN model.
- The **Validation Los**s is used after the CNN has been trained and represents the errors discovered after the test was used to test the predictive accuracy of the trained CNN model.

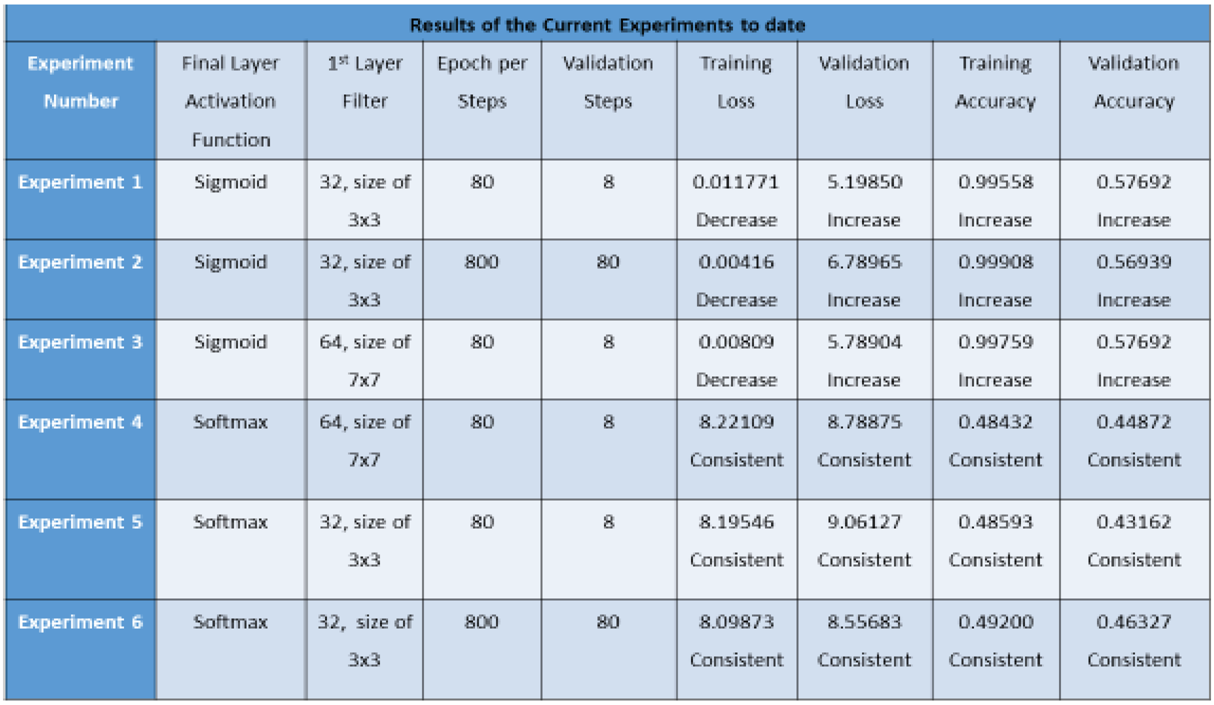
4. Initial Experiments

The CNN utilised the algorithm to train 5000 clean images and test 5000 stego-images testing different configurations to decrease the loss and increase validation

Experiments 1-3 were tested using the Final Layer Activation Function Sigmoid which is used for binary classification in logistical regression modelling.

Experiments 4-6 were tested using the Final Layer Activation Function Softmax which is used for multiple classifications in logistical regression modelling.

Table 1: Results of Experiments



Results of the Current Experiments to date								
Experiment Number	Final Layer Activation Function	1st Layer Filter	Epoch per Steps	Validation Steps	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
Experiment 1	Sigmoid	32, size of 3x3	80	8	0.011771 Decrease	5.19850 Increase	0.99558 Increase	0.57692 Increase
Experiment 2	Sigmoid	32, size of 3x3	800	80	0.00416 Decrease	6.78965 Increase	0.99908 Increase	0.56939 Increase
Experiment 3	Sigmoid	64, size of 7x7	80	8	0.00809 Decrease	5.78904 Increase	0.99759 Increase	0.57692 Increase
Experiment 4	Softmax	64, size of 7x7	80	8	8.22109 Consistent	8.78875 Consistent	0.48432 Consistent	0.44872 Consistent
Experiment 5	Softmax	32, size of 3x3	80	8	8.19546 Consistent	9.06127 Consistent	0.48593 Consistent	0.43162 Consistent
Experiment 6	Softmax	32, size of 3x3	800	80	8.09873 Consistent	8.55683 Consistent	0.49200 Consistent	0.46327 Consistent

4.1 Comparison of Experiment 1 & 3: Sigmoid

Results showed a decrease in the Training Loss of 0.003681 but an increase of 0.59054 in the Validation Loss. The Training Accuracy increased by 0.00201 and the Validation Accuracy was identical 0.57692. This comparison of results indicates when configuring the 1st Layer Filter of “32, size of 3x3” to “64, size of 7x7” it can improve the CNN model by decreasing the Training Loss and increasing the Training Accuracy. The experiment had a Validation prediction loss increase of 0.59054% but showed improvement when compared to the 1st Comparison as the Validation Accuracy remained balanced.

4.2 Comparison of Experiment 5 & 6: Softmax

Examining and comparing experiments 5 and 6, results showed a decrease in the Training Loss of 0.9673 and a decrease in the Validation Loss of 0.50444. The Training Accuracy increased by 0.00607 and the Validation Accuracy increased by 0.03165. This comparison of results indicates when configuring settings for the Epoch per steps from 80 to 800 and the Validation steps from 8 to 80 it can improve the CNN model results as both the training and validation Losses decreased. Also both the training and validation Accuracy results were increased.

4.3 Comparison of Experiment 2 & 6

Comparing experiment 2 with Sigmoid configurations and experiment 6 with Softmax configurations. Using the configurations for the Epoch per steps 800 and the Validation steps 80 with the 1st Layer Filter of “32, size of 3x3” being set. The results showed the Training Loss increased by 8.09873 from Sigmoid to Softmax and the Validation Loss increased by 1.76718. The results also showed a decrease in Training Accuracy by 0.50708 and a decrease in the Validation Accuracy by 0.10612. The results in this comparison are significantly larger than the other experiments as the Sigmoid is used for binary classification and the Softmax is used for multiple classifications.

5. Conclusions

As the configuring the Epoch per steps to 800 plus Validation steps to 80, and configuring 1st Layer Filter of “32, size of 3x3” to “64, size of 7x7” have showed that they can improve the performance of the CNN model for both Accuracy and Loss.

These low results for Softmax for the Validation Accuracy, which show a decrease of 0.1% and a Validation Loss increase of 2 on average when all 6 experiments were compared may indicate detection of steganalysis when compared to the Sigmoid results.

When implementing the deep learning CNN algorithm, results showed that increasing configurations such as the 1st Layer Filter, size, the epoch steps and the validation steps in the CNN code. The CNN model can improve the training and validation accuracy, also decreasing the loss in the training and validation of the CNN model. This leads to improved accuracy in the detection of OpenPuff steganography in a large dataset

This work in progress will now examine increased secret message size within the carrier image and various settings in the CNN model to improve accuracy.

References

Giarimpampa, D., 2018. Blind Image Steganalytic Optimization by using Machine Learning.

Types of Machine Learning. (2018). https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab [Accessed 14 Aug. 2019]

Hegarty, M. (2018), “Steganography, The World of Secret Communications”. CreateSpace Publishing Platform. ISBN-10: 1986125424

Hunt, R., 2012, December. New developments in network forensics—Tools and techniques. In 2012 18th IEEE International Conference on Networks (ICON) (pp. 376-381). IEEE.

Kessler, Gary C. "An Overview of Steganography for the Computer Forensics Examiner (Updated Version, February 2015)." (2015).

Ker, A.D., Bas, P., Böhme, R., Cogranne, R., Craver, S., Filler, T., Fridrich, J. and Pevný, T., 2013, June. Moving steganography and steganalysis from the laboratory into the real world. In Proceedings of the first ACM workshop on Information hiding and multimedia security (pp. 45-58). ACM.

Liu, K., Yang, J. and Kang, X., 2017, May. Ensemble of CNN and rich model for steganalysis. In 2017 International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 1-5). IEEE.

Lynch, S. Shepardson, D. “U.S. accuses pair of stealing secrets, spying on GE to aid China” Reuters News Agency, April 23, 2019. https://www.reuters.com/article/us-usa-justice-ge/us-accuses-pair-of-stealing-secrets-spying-on-ge-to-aid-china-idUSKCN1RZ24O [Accessed: 15th August 2019]

Morrison, B. (2004) “Ex-USA TODAY reporter faked major stories” USA Today. 3/19/2004 4:13 AM http://usatoday30.usatoday.com/news/2004-03-18-2004-03-18_kelleymain_x.htm [Accessed:10th September 2019]

Provos N and Honeyman P. 2001. Detecting Steganographic Content on the Internet: Center for Information Technology Integration University of Michigan (August 2001).

Richer, P., (Updated Version, February 2015). Steganalysis: Detecting hidden information with computer forensic analysis. SANS/GIAC Practical Assignment for GSEC Certification, SANS Institute, 6.

Van Heerden, H. "A First Course in Ethical Hacking" Second Edition, 07 November 2014

SHAPES Secure Cloud Platform for HealthCare Solutions and Services

Jyri Rajamäki[1, 2] and Aarne Hummelholm[2]
[1] Laurea University of Applied Sciences, Espoo, Finland
[2] University of Jyväskylä, Finland
jyri.rajamaki@laurea.fi
aarne.hummelholm@elisanet.fi

DOI: 10.34190/EWS.20.047

Abstract: The SHAPES project is an ambitious endeavour that gathers stakeholders from across Europe to create, deploy and pilot at large-scale a EU-standardised open platform incorporating and integrating a broad range of solutions, including technological, organisational, clinical, educational and societal, to enable the ageing population of Europe to remain healthy, active and productive, as well as to maintain a high quality of life and sense of wellbeing for the longest time possible. Not only each digital solution will be ethical, legal and appropriate for users, but also the results will align with the full and ethically responsible end-to-end exploitation of the new functionalities empowered by the secure cloud platform. Relevant EU and national legal or regulatory frames, ethical principles and fundamental rights will be considered, not only to guarantee due alignment with the development and use of the secure cloud platform but also to determine how to contribute to the reinforcement of the legal and policy frames to strengthen the deployment of large-scale cyber-secure digital solutions and the sustainability of the digital transformation of health and care delivery in Europe. This work in progress paper presents the project's findings so far with regard to privacy and cybersecurity issues of the SHAPES secure cloud platform.

Keywords: cloud services, cross-border healthcare, eHealth, healthy aging, SHAPES project, well-being

1. Introduction

Digital transformation and ecosystem thinking steer the Smart and Healthy Ageing through People Engaging in Supportive Systems (SHAPES) project (European Commission, 2019) that supports the well-being of the elderly at home. This work in progress paper discusses the project's finding during first three months with regard to privacy and cybersecurity issues.

2. From Hospitals to Home

Europe is ageing. Good health is not only of value to the individual as a major determinant of quality of life, well-being and social participation; it contributes to general social and economic growth (Eurostat, 2018). Accessibility to healthcare is thus one of the key priorities of the EU health policy. Integrated care focuses on the needs of the recipient on coordination between diagnosis and treatment and between primary care and secondary care, and between different therapeutic areas and specialties. Benefits of integrated care models are clear; still, the complexity of healthcare systems in individual countries and regions adds to the challenge. The redesign requires shifting care from hospitals to home. While individuals with chronic conditions need regular care and/or support that can often be delivered at home by community nurses, or others, potentially using information and communication technologies. Regular, reliable homecare ensures changes in the condition and treatment, resulting in better-managed conditions and fewer hospitalizations. New digital solutions include assistive robots, eHealth sensors and wearables, Internet of Things (IoT)-enabled devices and mobile applications. Cyber security is a prerequisite for the launch of these services.

Figure 1 shows the process of Intelligent Information Management for eHealth environments and it gives possibilities of how to follow and verify that patient sensor and IoT device data are going to the right place and are accessible only by authorized healthcare individuals. A system called 'IIMP' monitors patients' IoT-devices and sensors' information flow to hospital healthcare systems to provide healthcare personnel with a rapid analysis of the information. IIMP ensures that medical staff can find information about patients even in critical situations. This type of system can also help to find anomalies or data changes, or identify if someone has attempted to penetrate or use data in an undesirable way. The user can create specific action groups for this system. Members of the user group can follow the progress and changes of situations, data resources and reports in real time.

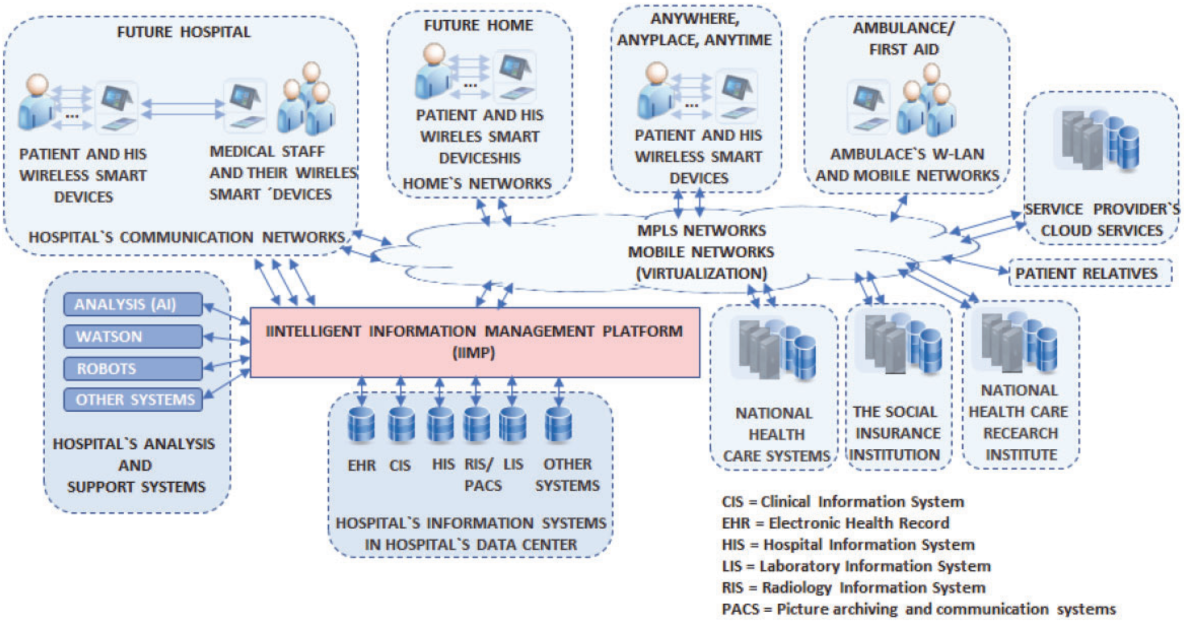

FUTURE HOSPITAL
PATIENT AND HIS WIRELESS SMART DEVICES
MEDICAL STAFF AND THEIR WIRELES SMART ´DEVICES
HOSPITAL`S COMMUNICATION NETWORKS
FUTURE HOME
PATIENT AND HIS WIRELES SMART DEVICESHIS
HOME`S NETWORKS
ANYWHERE, ANYPLACE, ANYTIME
PATIENT AND HIS WIRELESS SMART DEVICES
AMBULANCE/ FIRST AID
AMBULACE`S W-LAN AND MOBILE NETWORKS
SERVICE PROVIDER`S CLOUD SERVICES
MPLS NETWORKS
MOBILE NETWORKS
(VIRTUALIZATION)
PATIENT RELATIVES
ANALYSIS (AI)
WATSON
ROBOTS
OTHER SYSTEMS
HOSPITAL`S ANALYSIS AND SUPPORT SYSTEMS
IINTELLIGENT INFORMATION MANAGEMENT PLATFORM (IIMP)
EHR
CIS
HIS
RIS/ PACS
LIS
OTHER SYSTEMS
HOSPITAL`S INFORMATION SYSTEMS IN HOSPITAL`S DATA CENTER
NATIONAL HEALTH CARE SYSTEMS
THE SOCIAL INSURANCE INSTITUTION
NATIONAL HEALTH CARE RECEARCH INSTITUTE
CIS = Clinical Information System
EHR = Electronic Health Record
HIS = Hospital Information System
LIS = Laboratory Information System
RIS = Radiology Information System
PACS = Picture archiving and communication systems

Figure 1: eHealth operating environment (Hummelholm, 2019)

SHAPES leverages the outcomes of an open interoperability framework (symbIoT, 2018) that empowers the interconnection of different IoT-based platforms, devices, digital solutions and services. symbIoT middleware supports controlled and secure exchange of information, sharing of resources and best practices and delivering of services to third-parties, thus enabling the building of a true ecosystem of digital solutions and services that may be selected and tailored to the specific individuals' needs, interests and contextual environment, fostering SHAPES' large-scale pan-European deployment and market scale-up. symbIoTe including open software development kits and application programming interfaces provides an interoperable mediation framework that enables the discovery and sharing of connected devices across existing and future IoT-platforms for development of cross-platform IoT-applications.

3. Authentication and Security Assessment as a Service in SHAPES

SHAPES manages a high degree of sensitive information pertaining to older individuals thus it is critical that high standards for security and privacy (fully adopting GDPR) are implemented, resulting in a trusted platform among its users and stakeholders. User authentication considers mechanisms that are seamless, noticing that seniors may lack technological skills or have unreliable memory to recall strong passwords. So, SHAPES implements user authentication mechanisms based on Multimodal Biometrics using several physiological or behavioural human characteristic for enrolment, verification, authentication or identification. SHAPES demonstrates capabilities (face recognition and fingerprint) applying multi-factor authentication in a seamless way. Concerning device and component authentication, SHAPES brings a twofold approach:

1. SHAPES implements an authentication mechanism that is able to take into consideration state-of-the-art protocols (end-to-end encryption, PKI, web-based using secure API call and token-based authentication) and IoT-friendly lightweight protocols, such as the Constrained Application Protocol (CoAP) for message exchanges. Most IoT-devices allow re-use of existing Enrolment over Secure Transport (EST) functionality for energy-efficient certification management and secure bootstrapping operations.
2. SHAPES implements a state-of-the-art authentication mechanism based on secure stateless tokens 'Paseto' (Platform-Agnostic SEcurity Tokens) that enables a distributed mechanism for token-based authentication achieving inter-module authentication.

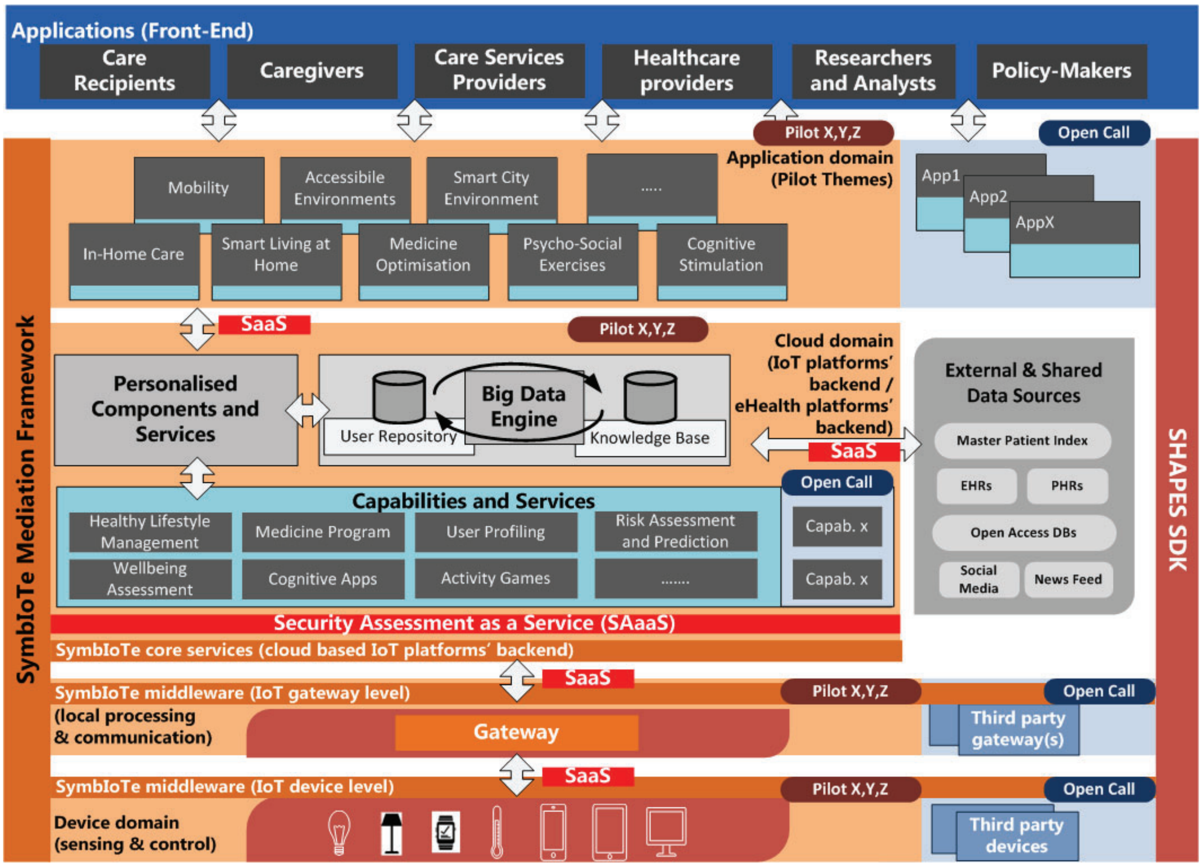


Figure 2: SHAPES high level architecture (SHAPES, 2019)

Considering the likelihood of operating in a not secured environment, SHAPES creates over-the-top secure environment (Virtual Local Area Network) enforcing security best practices. Moreover, it integrates a Security Assessment as a Service (SAaaS) cross-layered system to dynamically detect any existing and newly introduced network device, perform vulnerability assessments, certify the device against a standardised CVSS, assign it to a connectivity-appropriate VLAN and authenticate the service or device. Figure 2 presents SHAPES' high-level diagram.

4. Empowering

Health literacy, technology and individual involvement in care make healthcare more user-friendly and empowering, meaning that citizens (including seniors) must be seen as custodians of their own health (International Alliance of Patients' Organisations, 2014). Citizens continue to take a more central role in decisions about their own healthcare, and new technologies enable and facilitate this trend. New patients are evolving, similar to retail consumers. These former patients – new healthcare consumers – are driven by desire to take control over own health records and want to take active part in choosing healthcare providers and services. They are driven by the desire for more trustworthy, secure and timely healthcare information. Due to this changing role of patients, their empowerment has become a key priority for policy makers, professionals and service providers. Citizens' role is transforming from passive receivers of healthcare to active decision-makers; and managing own health data. Security is important aspect to empowering and creating the trust (Lettieri, et al., 2015).

As an example, activity-trackers enable to collect data from individuals' physical activities and health. A qualitative study (Lehto & Lehto, 2017) regarding the user perception on the privacy and sensitivity of health information collected with activity-tracker explored the user perception on health information sensitivity in general, and their willing, ness to share such information to other parties. Privacy calculus model (Laufer & Wolfe, 1977; Dinev & Hart, 2006) was the theory throughout the study. Table 1 summarize the results. Individuals do not perceive the information collected by activity-trackers as private or sensitive. But, information in medical records is considered to be very sensitive and private, as they include text written by doctors about procedures and discussions that are considered to be the most sensitive type of information.

Table 1: Research findings from activity tracker users' study (Lehto & Lehto, 2017)

Perceived privacy risks	**Perceived privacy concerns**	**Willingness to provide personal health data**
Information stolen	Being tracked or followed	Not on social media
Information lost	Banks deny loan applications	For medical research
Information misused	Employers won't hire	For healthcare purposes
Information goes to third parties	Insurance companies deny payments	For improving wearable products and services
Unauthorized access to the information	Information used for marketing	For occupational health services

The study extends earlier study information by presenting a new context to utilize the Privacy Calculus theory. The study's benefit is knowledge that activity-trackers users are ready to share their information that can be utilized in research and development of public services.

5. Cross-border Healthcare

The EU Directive on the Application of Patients' Rights in Cross-Border Healthcare is a starting point, delivering a legal framework for individuals willing to gain greater access to information related to healthcare available across Europe. However, in order to secure above-mentioned rights and unleash the potential of cross-border healthcare exchange, new solutions are needed to secure the storage and cross-border exchange of health data.

SHAPES cross-border implementation will be in line with modular strategy and related specifications and initiatives of epSOS (Smart Open Services for European Patients) project. In addition to the epSOS, the related and furthered projects are:

1. OpenNCP (Open Source Components for National Contact Points) as base realization starting point for SHAPES which includes a novel framework to foster cross-border eHealth services and which is base into the epSOS specifications;
2. current DECIPHER project which creates a mobile health care solutions and enables secure cross-border access to existing patient healthcare portals; and
3. STORK (Secure idenTity acrOss borders linked) which establishes a European eID Interoperability Platform that will enable citizens and businesses to use their national electronic identities in any participant Member State for public eGovernment services.

The designed contribution is that SHAPES can improve security and resiliency elements to related mechanisms and transactions, which are already established in the related projects.

Based on technological, integration and system readiness levels (Pirinen, 2015), SHAPES should develop new security readiness level (SecRL) metrics that supports the development of European operational standards for secure cross-border data exchange and patient privacy protection. Based on these metrics and prior open-source solutions (such as the OpenNCP suite (OpenNCP, 2015)), SHAPES could realise secure node platforms and components that enables the secure sharing and exchange of eHealth related data among countries.

6. Discussion

The rights of ageing individuals and their ability to live a good life at home or in a home-like environment are at the heart of the services designed in the SHAPES project. Privacy and security competence play a key role in the project, from planning to implementation and assessment. However, according to an ongoing Horizon 2020 cybersecurity project, health care sector can be identified as the most far from the ideal cybersecurity situation (ECHO, 2019).

The future complex environments present many challenges because the standards are not yet set at the international level. IoT products and sensors are mainly used at proprietary-based standards and getting them work at the same platforms in the smart devices will be a really big challenge.

Acknowledgements

This work was supported by the SHAPES project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 857159.

References

Dinev, T. & Hart, P., 2006. An extended privacy calculus model for e-commerce transactions. Information Systems Research, 17(1), pp. 61-80.

ECHO, 2019. Health Care Sector. [Online] Available at https://www.youtube.com/watch?v=X9ViO1w_9ko

European Commission, 2019. Smart and healthy ageing through people engaging in supportive systems. [Online] Available at: https://cordis.europa.eu/project/id/857159

Eurostat, 2018. Sustainable development in the European Union: Monitoring report on progress towards the SDGs in an EU Context, European Union.

Hummelholm, A., 2019. E-health systems in digital environments. 18th European Conference on Cyber Warfare and Security, pp. 641-649.

International Alliance of Patients' Organisations, 2014. Patient empowerment: for better quality, more sustainable health services globally, London: All Party Parliamentary Group.

Laufer, R. S. & Wolfe, M., 1977. Privacy as a concept and a social issue: A multidimensional developmental theory. Journal of social Issues, 33(3), pp. 22-42.

Lehto, Miikael & Lehto Martti, 2017. Health information privacy of activity trackers. 16th European Conference on Cyber Warfare and Security, pp. 243-251.

Lettieri, E. et al., 2015. Empowering patients through eHealth: a case report of a pan-European project. BMC Health Serv Res, 15(309).

OpenNCP, 2015. OpenNCP Installation. [Online] Available at: https://openncp.atlassian.net/wiki/display/ncp/OpenNCP+Installation

Pirinen, R., 2015. Towards common information sharing: Study of integration readiness levels. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 355-364.

SHAPES, 2019. Grant Agreement ID: 857159. European Commission.

symbIoT, 2018. symbIoTe project. [Online] Available at: https://www.symbiote-h2020.eu/

Late Submission

Artificial Intelligence in Megaprojects: The Next Frontier

Virginia A. Greiman
Boston University, Boston, MA, USA
ggreiman@bu.edu

DOI: 10.34190/EWS.20.123

Abstract:

Megaprojects are continuing to capture worldwide attention from India's Smart Cities, to the $4.75 billion Hadron Collider at CERN, to the US $150 billion International Space Station, and Europe's largest railway infrastructure project, Crossrail in London. The Organization for Economic Cooperation and Development (OECD) estimates global investment needs of $6.3 trillion per year from 2016-2030 (Mirabile, et al. 2017). To meet this growing demand, there has been a recent call, within the megaproject scholarship, for a better understanding of "what goes on in megaprojects – how they are managed and organized, from within, by the managers who are tasked with bringing them to fruition." (Söderlund, et al. 2017).

Studies about megaprojects have generally concentrated on the cost conflicts between the stakeholders and cost overrun issues (Adam et al., 2014; Flyvbjerg, 2017), however these have been superseded by more important issues in recent years such as project security, protecting the health and safety of its workers, project sustainability and value creation, and managing the impact of climate change and presently a global pandemic. All of these require a more agile approach to project management and a more sophisticated intelligence that can be generated through Algorithms that are used to generate artificially intelligent systems.

Through an analysis of the AI and Project Management literature, ethnographic studies, and semi-structured interviews with project management professionals, this paper explores the growing use of Artificial Intelligence to manage megaprojects including the obligations of private industry, and the government as the guardian of the public interest, while at the same time exploring the technical, managerial and ethical considerations in the deployment of AI.

Keywords: Megaprojects, Artificial Intelligence, Ethics, Machine Learning, Algorithms

Introduction to Megaprojects

Megaprojects are known for their vast size, long duration, and complexity due to the large number of stakeholders with diverse interests and expectations. These massive undertakings often defined as projects with budgets over $1 billion, are complex in design, and are known for their ambiguity, uncertainty, dynamic governance, financial challenges, environmental impact, and mobilization of social movement and citizen opposition (Clegg, et al. 2017; Cuganesan and Floris, 2020; Merrow, 2011; Pitsis et al. 2018). The complexity brings pressing challenges such as cost overruns, schedule delays, political risk, and safety issues. Megaprojects also create ecosystems that can last for decades even centuries as evidenced by the U.S. Interstate Highway System and High-Speed rail in Europe and Asia (Greiman and Sclar, 2020; EESI, 2018; Weingroff, 2017).

Megaprojects are not confined to large infrastructure or industrial projects but include research and development, cyber infrastructure, defense, food security, biotechnology, and many other fields. Given the increasing number of megaprojects and the movement towards the professionalization of the discipline, project management practice has struggled to develop workable solutions and practices to address complexities that exceed the technical concerns of engineering (Clegg, et al., 2017; van Marrewijk, et al. 2016). To overcome these challenges machine learning and intelligent design have become critical to every megaproject. This paper raises awareness about how governments and private industry face an unprecedented challenge in managing AI systems designed for megaprojects that include regulators, markets, and special interests that all play a role in influencing the development of AI in different contexts without a full appreciation of the impact of AI on people's livelihood and other consequences.

Project Management Research and Artificial Intelligence

One of the primary reasons for project management research is the development of a body of knowledge that can be applied to future projects across industry, culture, and continents. Megaprojects generate a tremendous amount of scrutiny and public concern, but also serve as generators of environmental, economic, institutional, and social sustainability. As noted by scholars in the field of project management research, the nature of megaprojects brings together significant tacit knowledge that is embedded within particular groups in the project requiring effective knowledge management both within and across projects (Olaniran, 2017; Walker, 2016). Large capital projects, especially in the energy and chemicals, and mining and metals markets, are incredibly complex with enormous amounts of data, people and moving parts that are constantly changing and need to be understood through social theory on the part of ethnographic researchers imbedded in the project (Greiman and Bernardin, 2018). As one scholar observes, "[the] notion of the researcher being a disinterested observer is a fallacy or perhaps even a fantasy….as the research process and the research findings are understood to be drawn on social awareness, activities and requirements in which the researcher will be highly involved" (Remenyi 2013, p. 145).

This paper focuses on the use of Artificial Intelligence in the world's megaprojects, an area of research that has not been widely covered in the engineering, technology, or the project management journals. In particular, AI has raised the concern of the potential impact of this technology on labor income and the global economy. The World Economic Forum's Future of Jobs Report indicates that from 2018 to 2022 "75 million jobs may be displaced by a shift in the division of labor between humans and machines" (WEF, 2018).

In addition to job loss, the research points out major risks in relation to AI and autonomous systems such as causation of damages, lack of transparency, loss of privacy and personal autonomy, potential information biases, and susceptibility to manipulation of AI and autonomous systems (Walz and Firth-Butterfield, 2019). Various approaches to AI regulation have been offered by governments and private industry and continued dialogue and collaboration is essential. Price Waterhouse Coopers (PwC) projects that artificial intelligence will contribute 15.7 trillion U.S. dollars to the global economy by 2030 (PwC, 2017).

The potential benefits of AI as demonstrated in the research are tremendous. AI could play a crucial role in many aspects of human problems including climate change, international conflict, and medical breakthroughs, and could contribute greatly in improving human exploration, scientific knowledge and even quality of life. However, the project management literature has raised a number of areas where AI deployment has created serious concerns about the impact on decisions and outcomes (Raso et al., 2018; Yu, 2019). The paper will explain how AI is currently being deployed in megaprojects, and the challenges and risks in AI deployment.

The focus of this research is to identify through an empirical investigation including the extant literature and semi-structured interviews the challenges of implementing artificially intelligent systems in large scale projects, and to stimulate ideas for further research (Dafoe, 2018). The research focuses on two primary areas: (1) How will AI help deliver "better" projects? (2) What are the risks, managerial implications, and challenges in designing and implementing AI in major projects?

This paper is structured in two parts. In Part I we discuss the evolution of AI, the literature, and the case studies on the use of artificial intelligence in megaproject decision making and for project oversight. In Part II we analyze the technical and managerial implications and offer recommendations for the strategic use of AI and the acceptance of AI by a project's sponsors and stakeholders.

1. Part I: The Evolution of AI

"Artificial intelligence (AI) refers to a set of computer science disciplines aimed at the scientific understanding of the mechanisms underlying thought and intelligent behavior and the embodiment of these principles in machines that can deliver value to people and society" (U.S. Senate, 2018). Since 1837 when the first design for a programmable machine was developed by Charles Babbage and Ada Lovelace in 1837, we have come a long way. Though by the mid-1970s computers flourished it was not until the1980s that deep learning techniques were explored in greater depth when Edward Feigenbaum introduced expert systems which mimicked the decision-making process of a human expert, and autonomous vehicles were introduced (Haenlein and Kaplan, 2019).

AI can be understood by contrasting it with its related components. Algorithms are the building blocks that make up machine learning and artificial intelligence, whereas machine learning sometimes called deep learning is a subset of AI composed of algorithms that allow computers to learn without being explicitly programmed (Linux, 2005). Moreover, artificial intelligence can be categorized into three basic stages of development. Basic AI or Artificial Narrow Intelligence (ANI) is limited in scope and restricted to just one functional area. Advanced AI or Artificial General Intelligence (AGI) usually covers more than one field, such as power of reasoning, abstract thinking, or problem solving on par with human adults. Autonomous AI or Artificial Super Intelligence (ASI) is the final stage of intelligence expansion in which AI surpasses human intelligence across all fields. This stage of AI is not expected to be fully developed for several decades (Mou, 2019).

A key consideration in the use of algorithms is not only whether safeguards are in place but also whether they are able to operate effectively (ACM, 2017; McGregor et al., 2019, p. 21).

1.1 How is AI currently being deployed in Megaprojects?

AI is currently being deployed by large engineering and construction firms in different ways. These include project and site selection, design optimization, risk management, cost estimation management, schedule performance, and quality assurance. One of the most important areas for the introduction of AI is the management of workers safety and health. Presently, AI is being used by megaproject managers to identify patterns of behavior that will minimize the threat of serious accidents or even worse catastrophic loss (Greiman, 2013). Discussed below are examples of deployment of AI in various megaprojects and for different purposes. These examples are not to be considered models for deployment since every megaproject has its own goals, mission, economic, social, financial technical, and legal environments to consider and the use of AI should be carefully examined on a case by case basis.

A study conducted by The Economist Intelligence Unit in 2018 and sponsored by Microsoft of more than 400 business executives and policymakers in both advanced and developing economies across all sectors revealed that the AI technologies most frequently used include image analysis (35%), virtual assistants (31%), predictive analytics (29%), machine learning (28%) and natural language processing (26%). The recent advances in AI come under the rubric of "machine learning," which involves programming computers to learn from example data or past experience (Agrawal, et al., 2017, p. 23). Environments with a high degree of complexity are where machine learning is most useful. Predicting the presence of a disease, a risk condition, or the need for repair of heavy equipment are all benefits of machine learning that are applicable to megaprojects. The examples of AI deployment in megaprojects are numerous but in this paper the discussion will be limited to the use of AI in intelligence analytics in the defense industry, occupational safety and health incident tracking in construction projects, and intelligence systems in risk analysis.

1.2 AI Application for Defense Megaprojects

The U.S. Department of Defense (DoD) has considered a number of diverse applications for AI in the megaproject domain (CRS, 2019). Other nations, particularly China and Russia, are making significant investments in AI for military purposes, including in applications that raise questions regarding international norms and human rights (DoD, 2018). The Office of the Under Secretary of Defense for Research and Engineering, which oversaw the development of DoD's A1 Strategy, continues to support AI development and delivery. A few of the areas where AI is already being employed include predictive aircraft maintenance; detection of anomalies in patterns of network activity thus creating a dynamic barrier to attack; force protection, recruiting, healthcare, multi-domain command and control, and the development of semiautonomous and autonomous vehicles, including fighter aircraft, drones, ground vehicles, and naval vessels (CRS, 2019; DoD, 2018).

Raytheon, one of America's largest defense contractors is using AI "[t]o improve the speed and power of computers used in intelligence analytics, by testing quantum processing technologies that may hold the keys to advanced encryption technologies, nanotechnology, artificial intelligence, and machine learning, and more accurate radars and sensor systems. To meet growing demand for deterrence, they are pioneering semi-autonomous warfare technologies that bring together warfighters, computers, and robots to speed decision-making and targeting (Raytheon, 2018).

1.3 Safety Incident Tracking

Project complexity and having a high number of workers increase the probability of encountering incidents. The safety incidents should be traced and documented rigorously. The AI research highlights the occupational safety problem in construction sites for mega-projects, which can also have highly significant cost overrun problems. Within the context of safety assessment, record keeping systems are gaining more importance in addressing safety problems in megaprojects (Vidal et al., 2011). Record keeping is the key element of preventative actions for megaprojects; and the use of AI-based systems can prevent incidents in megaprojects. One important research study by Ayhan and Tokdemira (2019) proposed an innovative safety assessment methodology to predict the possible scenarios and determine preventative actions. The study introduces predictive models based on factual information and shows how to deal with the safety data which has a significant level of heterogeneity. The models achieved results in 86% of the test cases with 18% error at most. Moreover, the overall prediction accuracy of fatal incidents was equal to 86.33%. As a final step, the study presented preventative actions to eliminate safety failures. Ultimately, the proposed methodology can assist construction industry professionals in examining future safety problems by utilizing the collected large-scale data. Furthermore, the study provides necessary preventative measures to be implemented before and during the construction stage (Ayhan and Tokdemira, 2019).

1.4 Intelligent Systems for Risk Analysis

One of the primary purposes of artificial intelligence in Megaprojects is for the determination of risk. The number one priority on all projects is on-the-job safety. Natural Language Processing (NLP) is an interdisciplinary field involving humanistic, statistical–mathematical, and computer skills that has been used to understand risk perception. The aim of NLP is to process languages using computers. The human language can be defined as natural because it is ambiguous and changeable. On the contrary, machine language is defined as formal because it is unambiguous and internationally recognized (Di Giuda, 2020). The NLP must deal optimally with the ambiguity, imprecision, and lack of data inherent in natural language.

This application of NLP would help an organization extract information regarding the perception of risk from hundreds of contracts so that policies and procedures could be put in place to mitigate these risks in the master contract documents before they became uncontrollable and cause significant cost increases which impact the project financing. For more easily identifiable risks the people would train the machines to recognize the common terms. For more complicated risks such as analysis of human behaviors and governance failures the people rather than the machines would analyze these risks for a deeper understanding of the psychological factors that impact behavior.

2. Part II: Technical and Managerial Considerations in the Implementation of AI

Project owners, sponsors, designers, and contractors face many obstacles and challenges in the development and implementation of AI. While cutting-edge technology and talent are needed, it’s equally important to align an organization’s culture, structure, and ways of working to support broad AI adoption which requires interdisciplinary collaboration, and agile, experimental, and adaptable data-driven decision making at the front line (Fountain et al., 2019). Implementing AI includes problem formulation, feasibility studies, strategic analysis as to the criticality of the AI, financial analysis of the project’s schedule and budgets, standards for an AI project selection process, testing, monitoring and evaluating the AI throughout the project life cycle, integrating stakeholders in the process, developing operational plans, and building team organization and change management that can adapt to the evolution that AI will introduce. Implementing AI is exacerbated in megaprojects due to the overwhelming number of stakeholders, documents, large scale policy-making and dynamic governance structures involved. In Part II we discuss three of the most important challenges facing AI implementation including data collection and analytics, the integration of machine and human intelligence, and ethical concerns in the deployment of AI.

2.1 Data Analytics

One of the major challenges for the implementation of AI is the collection of data and analysis of vast amounts of data. AI can only be as good as the data it is given. If the input data reflects biases, so too will the output (Kleinberg et al.,2017). The regularity of the data is dependent on a variety of factors including the variance in published information, the completion of the published information, the consistency of the data collected and any bias that may exist in the development of the data. The next step is the analysis of the data and the modeling of the data. This requires that the data be analyzed both on a qualitative and quantitative basis using algorithmic modeling (Kleinberg et al., 2017). Algorithmic models have been used for decades in the insurance industry to assess aggregate exposure and can be used as models for other disciplines. An interdisciplinary approach is essential to understanding the various impacts that data can have on human rights, privacy, and confidentiality (Raso, 2018). Expert analysis and input from various disciplines including risk management, geo-technical analysis, quality assurance, engineering, construction, actuarial science, law, and project management is employed to provide a broader perspective to the data analysis. The interaction between the AI analysis and the human analysis is critical to the development of predictive accuracy that can better support rational decision making (Agrawal, 2017). Some important questions for project management in data analysis include:

- Do you have a data analytics and machine learning strategy and policy in place?
- What are the benchmarks you will use to collect and structure the data?
- What are the sources of your data and will they be sufficient to address the problem identified as suitable for an AI analysis?
- How much data is required to use the algorithms and technologies in a particular quantitative or qualitative analysis?
- Who will collect and compile the data and ensure its legitimacy and authenticity?
- How will you decide how to organize your data?
- What data architecture and tools will you use?
- Is the quality of your data appropriate for the project?
- How will you test the quality of your data?
- Have you tested your data before building your machine learning algorithms?
- How will you perform data integration?

2.2 Machine Intelligence versus Human Intelligence

In his 2016 testimony before the Senate Armed Services Committee, Commander of U.S. Cyber Command Admiral Michael Rogers stated that relying on human intelligence alone in cyberspace is "a losing strategy (Rogers, 2016). "If you can't get some level of AI or machine learning with the volume of activity you're trying to understand when you're defending networks … you are always behind the power curve" (Lyle, 2016).

Though AI has been proven to assist humans in preventing attacks in cyber space and other significant activities, the reality of artificial intelligence is that machines cannot solve all problems. Because they lack the ability to reason there are limits to what AI can produce. A test-and-learn mentality will reframe mistakes as a source of discoveries, reducing the fear of failure (Fountaine, et al. 2019). Some researchers have found that while computers and AI can be superior to humans in some skill and rules based tasks, under situations that require judgment and knowledge, in the presence of significant uncertainty, humans are superior to computers (Cummings, 2017). From the military perspective, Dr. Arati Prabhakar, the former DARPA Director commented, "When we look at what's happening with AI, we see something that is very powerful, but we also see a technology that is still quite fundamentally limited … the problem is that when it's wrong, it's wrong in ways that no human would ever be wrong (Pomerlau, 2016).

A major deficiency across all AI algorithms is the lack of so-called explainability (Gunning, 2017). For example, even when an AI system correctly identifies an image, such as a cat, the system's developers may not be able to determine which traits of a cat the system was using in its identification process (Markoff, 2012). In heavy construction, machines may prevent or mitigate accidents, but humans will need to assess the root cause of a potential accident and adopt new rules and procedures for human conduct.

2.3 Ethical Concerns in the Deployment of AI

If cognitive bias leads humans to struggle to trust their AI-enabled smartphone GPS, how will it play out when commanders ask unit-level operators to rely on specific AI applications in life or death situations? Will it be able to distinguish the war fighter from the civilian? There are many ethical and human rights issues that arise in the implementation of AI to perform tasks normally performed by humans, or to assist humans in performing the task (Azoulay, 2019). In addition, to live and death issues these include usurpation of privacy, discrimination in decision making, and bias due to selective collection of data. For example, to the extent that the data used to train an AI system is biased, the resulting system will reflect or even strengthen those biases (Osoba and Welser, 2017). The investigation of human rights violations by the Berkman Klein Center for Internet and Society at Harvard University identified bias from AI decision-making in several different fields. These included criminal justice (risk assessments); finance (credit scores); healthcare (diagnostics); content moderation (standards enforcement); human resources (recruitment and hiring); and education (essay scoring) (Raso et al., 2018).

Decisions made by an AI system's human designers can have significant human rights consequences both positive and negative which are informed by the individual life experiences and biases of the designers. The one AI tool with the greatest threat for ethical considerations in megaprojects is the tool designed to help select managers, staff, and team members. It would be a simple task to develop algorithms for compliance with state and federal laws regarding citizenship, age, gender, or specific job requirement skills like educational and work experience. However, there is also a risk of bias being introduced into the algorithms depending on how they are written. For example, there may be gender-specific job sites or ethnic bias or directives to investigate a candidate's social media sites for political or personality traits deemed to be unacceptable by the algorithm developer. The only way to minimize this risk would be for all data and algorithms to be created through a diverse inter-disciplinary team with regular reviews and testing to ferret out bias.

One of the most important tools for avoiding or mitigating ethical concerns in hiring and promotion is to implement auditing systems that regularly test for bias and errors as well as intentionally designing algorithms to control for human biases. Just as we would test a product before it is released to the market we must also be constantly testing and evaluating the algorithms we are using to ensure that conflicts and bias are eliminated to the fullest extent possible.

3. Conclusion

A review of the literature and practice of AI in megaprojects has revealed a much more extensive use than one might have perceived considering that AI is still in the early stage of development in certain industries and organizations. The broad use of AI in megaprojects creates many opportunities for greater efficiency and value added, but also many challenges. One of the greatest challenges for researchers, policymakers and private industry in the building of large infrastructure is to ensure its purpose is being fulfilled while also creating AI systems that address ethical concerns in the collection and analysis of data and algorithmic accountability. AI presents multifaceted impacts on a project's workers, managers, and stakeholders that must be fully understood from a political, civil, economic, social, and cultural perspective. The adoption of AI must be fully investigated, and evaluated by internal and external experts, if projects are to move forward and reap the benefits of a successful AI deployment. The conclusions reached so far indicate a need for a better understanding of the existing frameworks now used to monitor artificial intelligence while being aware of its potential applications.

References

Adam, A., Josephson, P.E., Lindahl, G. (2014) "Implications of cost overruns and time delays on major public construction projects," Proceedings of the 19th International Symposium on the Advancement of Construction Management and Real Estate, Chongqing.

Agrawal, A., Gans, J.S., Goldfarb, A. (2017) "What to Expect from Artificial Intelligence," MIT Sloan Management Review, Vol. 58, No. 3, pp 22-26.

Association for Computing Machinery (ACM), (2017) Statement on Algorithmic Transparency and Accountability, US Public Policy Council (USACM), January 12.

Azoulay, A. (2019) "Towards an Ethics of Artificial Intelligence." UN Chronicle, Vol. 55, pp 3-4, January.

Ayhan, B.U. and Tokdemira, O.B. (2019) “Safety assessment in megaprojects using artificial intelligence,” Safety Science, Vol. 118, pp 273-287, October.

Banker, D., Mashruwala, R. and Tripathy, A. (2014) "Does a differentiation strategy lead to more sustainable financial performance than a cost leadership strategy?" Management Decision, Vol. 52, No. 5, pp 872-896.

Clegg, S., Shankar, S., Biesenthal, C., Pollack, J. (2017) “Power and Sensemaking in Megaprojects.” In the Oxford Handbook of Megaproject Management edited by Bent Flyvbjerg, Oxford University Press, pp. 238-258.

Congressional Research Service (CRS) (November 21, 2019) Artificial Intelligence and National Security, Washington, D.C.

Cuganesan, S. and Floris, M. (2020) “Investigating perspective taking when infrastructure megaproject teams engage local communities: Navigating tensions and balancing perspectives,” International Journal of Project Management, Vol. 38, pp. 153-164.

Cummings, M.L. (2017) Artificial Intelligence and the Future of Warfare, International Security Department and US and the Americas Programme, Chatham House, London.

Dafoe, A. (2018) “AI Governance: A Research Agenda,” Future of Humanity Institute, University of Oxford, August 27.

Department of Defense (DoD (2018) ‘Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance our Security and Prosperity,” U.S. Department of Defense, Washington, D.C.

Di Giuda G.M, Locatelli, M., Schievano, M., Pellegrini, L., Pattini, G., Giana, P.E., Seghezzi, E. (2020) “Natural Language Processing for Information and Project Management.” In: Daniotti B., Gianinetto M., Della Torre S. (eds) Digital Transformation of the Design, Construction and Management Processes of the Built Environment. Research for Development. Springer, Cham.

The Economist. (2018) “Intelligent Economies: AI's transformation of industries and society.” The Economist Intelligence Unit, July 29.

Environmental and Energy Study Institute (EESI) Fact Sheet: High Speed Rail Development Worldwide, July 19, 2018.

Fountaine, T., McCarthy, B., and Saleh, T. (2019) “Building the AI-Powered Organization,” Harvard Business Review, July–August.

Flyvbjerg, B (2017) “Megaprojects: Over Budget, Over Time, Over and Over,” Policy Report, Cato Institute, Washington, D.C., January/February.

Greiman, V. A. (2013) Megaproject Management: Lessons on Risk and Project Management from the Big Dig, John A. Wiley and Sons, Inc., London, Hoboken.

Greiman, V. and Bernardin E.M. (2018) “The Shaping of Megaprojects: An Autoethnographic Study.” Published in the Proceedings of the 2018 International Research Network on the Organizing of Projects (IRNOP), RMIT, Melbourne, Australia.

Greiman, V. and Sclar, E. (2020) “Mega-Infrastructure as a Dynamic Ecosystem: Lessons from America’s Interstate System and Boston’s Big Dig,” Journal of Mega Infrastructure and Sustainable Development, https://doi.org/10.1080/24724718.2020.1742624.

Gunning, D. (2017) “Explainable AI Program Description, XIA,” The Defense Advanced Research Projects Agency DARPA/I20), U.S. Department of Defense, November 4.

Haenlein, M. and Kaplan, A. (2019) “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence,” California Management Review, Vo. 64, No. 4, pp 5-14.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). “Human Decisions and Machine Predictions,” National Bureau of Economic Research, February.

Linux. (2005) “Algorithms a Very Brief Introduction.” The Linux Foundation, San Francisco, CA, July 29.

Lyle, A (2016) “National Security Experts Examine Intelligence Challenges at Summit,” U.S. Department of Defense, September 9.

Markoff, J. “How Many Computers to Identify a Cat? 16,000.” The New York Times, June 25, 2012.

McGregor, L., Murray, A., Ng, V. (2019) “International Human Rights Law as a framework for Algorithmic Accountability,” International and Comparative Law Quarterly, 68, pp. 309-343, April 1.

Merrow, E.W., (2011) Industrial Megaprojects: Concepts, Strategies and Practices for Business, John Wiley & Sons, Hoboken.

Mirabile, M., Marchal, V., and Baron, R. (2017) Technical notes on estimates of infrastructure investment needs: Background note to the report. Investing in Climate, Investing in Growth, Office of the Secretary General, OECD, Paris, France.

Mou, X. (2019) “Artificial Intelligence: Investment Trends and Selected Industry Uses,” The International Finance Corporation, The World Bank Group, Washington, D.C., September.

Olaniran, O.J. (2017) “Barriers to tacit knowledge sharing geographically dispersed project teams in oil and gas projects.” Project Management Journal, Vol. 48, No. 3, pp 41-57.

Osoba, O.A. and Welser, W. (2017) The Risks of Bias and Errors in Artificial Intelligence. Santa Monica: Rand Corporation.

Pitsis, A., Clegg, S. Freeder, D., Sankaran, S., Burdon, S. (2018) “Megaprojects redefined – complexity vs cost
and social imperatives,” International Journal of Managing Projects in Business, Vol. 11, No. 1, pp 7-34.

Price Waterhouse Coopers (PwC) (2017) “Sizing the prize What’s the real value of AI for your business and how can you capitalize?” Price Waterhouse Coopers, London.

Pomerlau, M. “DARPA Director Clear Eyed and Cautious on AI,” Government Computer News, May 10, 2016.

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., Kim, L. (2018) "Artificial Intelligence and Human Rights: Opportunities and Risks," Berkman Clein Center for Internet and Society, Harvard University, Cambridge, MA, September 25.

Raytheon (2018) Transforming Tomorrow: Raytheon 2018 Annual Report.

Remenyi, D. (2013) Grounded Theory: A Reader for Researchers, Students, Faculty and Others. Academic Conferences and Publishing International Limited, Reading, UK.

Söderlund, J., Sankaran, S., Biesenthal, C., (2017) "The past and present of megaprojects," Project Management Journal, Vol. 48, No. 6, pp 5–16.

Testimony of Michael Rogers, Senate Armed Service Committee, Hearing to Receive Testimony on Encryption and Cyber Matters, September 13, 2016.

U.S. Senate Committee on the Judiciary, 115th Cong. (2018) (written statement of Christopher Wylie to the United States Senate Committee on The Judiciary in the Matter of Cambridge Analytica and Other Related Issues), 16 May.

van Marrewijk, A., Ybema, S., Smits, K., Clegg, S., Pitsis, T. (2016), "Clash of the titans: Temporal organizing and collaborative dynamics in: the panama canal megaproject," Organizational Studies, Vol. 37, No. 12, pp 1745-1769.

Walker, D.H.T., Lloyd-Walker, B. (2016) "Rethinking project management: Its influence on papers published in the international journal of managing projects in business," International Journal of Managing Projects in Business, Vol. 9, No. 4, pp. 716-743.

Walz, A., and Firth-Butterfield, K. (2019) "Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to The Development of An AI Governance Regime," 18 Duke Law and Technology Review, Vol. 18, pp. 176-231, December 20.

Weingroff, R.F. (2017) The greatest decade 1956-1966: Celebrating the 50th anniversary of the Eisenhower interstate system, part 1 essential to the national interest, U.S. Department of Transportation, Federal Highway Administration, (FHWA), Washington, D.C.

World Economic Forum (WEF) (2018) Future of Job's Report. Centre for the New Economy and Society. Geneva, Sept. 17.