Published by CCDCOE

2018

10th International Conference on Cyber Conflict

CyCon X: Maximising Effects

T. Minárik, R. Jakschis, L. Lindström (Eds.)

CYCON

IEEE
Advancing Technology for Humanity

30 May - 01 June 2018, Tallinn, Estonia

2018
10TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT
CYCON X: MAXIMISING EFFECTS

IEEE Catalog Number: CFP1826N-PRT
ISBN (print): 978-9949-9904-2-9
ISBN (pdf): 978-9949-9904-3-6

COPYRIGHT AND REPRINT PERMISSIONS

NATO CCD COE Publications
Filtri tee 12, 10132 Tallinn, Estonia
Phone: +372 717 6800
Fax: +372 717 6308
E-mail: publications@ccdcoe.org
Web: www.ccdcoe.org
Head of publishing: Jaanika Rannu
Layout: Jaakko Matsalu
Cover design: AKU

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

The NATO Cooperative Cyber Defence Centre of Excellence (CCD COE) is a NATO-accredited cyber defence hub focusing on research, training and exercises. It represents a community of currently 21 nations providing a 360-degree look at cyber defence, with expertise in the areas of technology, strategy, operations and law. The heart of the Centre is a diverse group of international experts from military, government and industry backgrounds.

The CCD COE is home to the *Tallinn Manual 2.0*, the most comprehensive guide on how International Law applies to cyber operations. The Centre also organises the world's largest and most complex international live-fire cyber defence exercise, Locked Shields. Another highlight of the Centre is the International Conference on Cyber Conflict, CyCon, a unique event joining key experts and decision-makers of the global cyber defence community in Tallinn every spring. As of January 2018 CCD COE is responsible for identifying and coordinating education and training solutions in the field of cyber defence operations for all NATO bodies across the Alliance.

The Centre is staffed and financed by its member nations - Austria, Belgium, the Czech Republic, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Slovakia, Spain, Sweden, Turkey, the United Kingdom and the United States. NATO-accredited centres of excellence are not part of the NATO Command Structure.

CYCON 2018 SPONSORS

TECHNICAL SPONSOR

DIAMOND SPONSORS

GOLD SPONSOR

TABLE OF CONTENTS

INTRODUCTION

CyCon X is the tenth iteration of the annual International Conference on Cyber Conflict, organised by the NATO Cooperative Cyber Defence Centre of Excellence and taking place in Tallinn from 29 May to 1 June 2018. Over the years, CyCon has become a world-recognised conference addressing cyber conflict and security from the perspectives of technology, strategy, operations, law, and policy. We are always glad to see our friends in Tallinn again – a number of them have been involved with CyCon since its origins a decade ago – and we also welcome newcomers, who can discover the cyber debates and 'white night' walks in Tallinn's Old Town. We are proud to offer them all the opportunity to meet and learn something new from each other. If CyCon has been able to contribute to interdisciplinary understanding of cyber conflict and security throughout the years, then it has achieved its main goal.

CyCon X's core topic is 'Maximising Effects'. Since the very beginning, cyberspace has provided unparalleled opportunities to achieve effects in new and novel ways. Today, cyberspace provides a technological platform and an environment for diverse actors, with both good and bad motivations, to influence everyone and everything. Maximising effects in the cyber realm is important for business, media, governments and military, and even private users. However, how will this be achieved and what will the consequences be? How will AI, machine learning and big data help to maximise effects in cyberspace? How will international law develop in light of the serious effects of state-sponsored operations that may or may not be hard to attribute? The effects generated through cyberspace, including new instabilities and vulnerabilities, will require new policies, legal frameworks and technological solutions to maximise security.

In response to the Call for Papers in June 2017, almost 200 abstracts were submitted in October. After a careful selection and peer review by the Academic Review Committee, this book contains 22 articles whose authors were invited to present at the conference.

Christopher Whyte, Brandon Valeriano, Benjamin Jensen, and **Ryan Maness** describe the prospects for open-source, public data collection for cyber security events and present an initial data collection and analysis of interstate cyber conflict incidents involving the United States. **Aaron F. Brantly** examines the applicability of deterrence in the digital age and for digital tools, based on examples from both within and beyond cyberspace. **Max Smeets** and **Herbert S. Lin** aim to explain if (and how) offensive cyber capabilities have the potential to change the role of military power and argue that these capabilities can alter the manner in which states use their military power strategically. **Quentin E. Hodgson** seeks to develop an understanding of how

states use cyber capabilities to coerce others for political objectives and examines the use of cyber operations by North Korea and Russia in recent years as part of their broader strategies. **Daniel Moore** argues that military offensive network operations can be usefully cast into a two-part taxonomy: event-based attacks and presence-based attacks – these two types offer different solutions, encompass varying risks, and may require different resources to accomplish.

Martin C. Libicki shows how cyber espionage between state adversaries can 'alter the balance of a confrontation' and 'shape the inferences that the other side draws about one's intentions' in cyberspace. **Brad Bigelow** suggests that 'cyberspace' as a label for a domain should not be confused with the individual networks – some interconnected ('open') and some relatively isolated ('closed') – involved in military operations; and illustrates the importance of precision in describing the composition of cyberspace. **Kim Hartmann** and **Keir Giles** investigate the potential opportunities and challenges of an adjustment to the principle of net neutrality to facilitate defensive action by legitimate actors; how this adjustment could contribute to regaining control in congested cyber domains in the case of national or international cyber incidents; and the associated risks. **Robert Koch** and **Mario Golling** analyse the development of both cyber threats and defence capabilities during the past 10 years, evaluate the current situation and give recommendations for improvements, including an overview of upcoming technologies that will be critical for cyber security. **Kārlis Podiņš** and **Kenneth Geers** describe the technical aspects of malware re-weaponisation and the implications and ramifications of this phenomenon for a range of strategic concerns, including weapons proliferation and attack attribution.

Turning to the legal perspective, **Asaf Lubin** provides his view of how low-intensity cyber operations and peacetime espionage operations should be subjected to a single regulatory framework: that cyber law and espionage law should be viewed as 'communicating vessels'. **Krisztina Huszti-Orban** explores the division of responsibilities between the public and private spheres in countering terrorism and violent extremism, focusing on ways to ensure that Internet intermediaries follow international human rights standards in the process. **Peter Z. Stockburger** examines the precautionary principle in international law and argues that its application can help crystallise the due diligence principle in cyberspace. **Cedric Sabbah** suggests a shift in the approach to cyber norms development: due to the lack of consensus in the UN GGE process, the international community should support the discussions that are already occurring between cybersecurity regulators and authorities. Finally, **Jeff Kosseff** proposes and elaborates on four goals of common international principles for cybersecurity law: modernisation of cybersecurity laws; uniformity of legal requirements; coordination of cooperative incentives and coercive regulations; and supply chain security.

There are seven articles with a technological viewpoint, the first being a case study authored by **Martin Strohmeier et al.** exploring the collection of air traffic communication data via open source intelligence methods, for tracking mission critical military and governmental movements. Next, **Roland Meier et al.** present a threat-intelligent feed that exhibits a robust resistance to tampering attempts in order to provide organisations and individuals with the most original, most valuable and newest feed entries. In their article, **F. Jesús Rubio Melón** and **Artūrs Lavrenovs** provide an examination of HTTP security headers of one million most popular websites to assess web security policy implementation rates compared to its HTTP equivalents. **Giovanni Apruzzese et al.** present an in-depth analysis of adopted machine and deep learning algorithms and their usability for intrusion detection, malware analysis, and spam detection. Regarding insider threat and malicious agents, **David Gugelmann** and **David Sommer et al.** explore a novel hidden screen watermarking technique for infiltrated organisations to more rapidly identify and reduce threats after document leaks have occurred. **Roman Graf** and **Ross King**'s contribution explores an automated approach for incident reports management, using neural networks and smart contracts. Finally, **Steven Noel et al.** highlight a prototype tool aimed at improving network security while simultaneously supporting the protection of mission-critical assets in enterprise or military environments.

All the articles in this book have gone through a double-blind peer review by, at minimum, two members of CyCon's Academic Review Committee. We greatly commend the members of the Committee for guaranteeing the academic quality of the book by reviewing and selecting the submitted papers.

Academic Review Committee Members for CyCon 2018:

- Siim Alatalu, NATO CCD COE
- Dr Elie Alhajjar, Army Cyber Institute, United States
- Prof Robert E. Barnsby, Army Cyber Institute; United States Military Academy
- Prof Col Daniel Bennett, Army Cyber Institute, United States
- Prof Giuseppe Bianchi, University of Rome Tor Vergata, Italy
- Bernhards Blumbergs, CERT Latvia
- Václav Borovička, National Cyber and Information Security Agency, Czech Republic
- Maj Pascal Brangetto, French Ministry of Defence
- Dr Russell Buchan, University of Sheffield, United Kingdom
- LtCol Joshua Bundt, Army Cyber Institute, United States
- Dr Joe Burton, University of Waikato, New Zealand
- Dr Steve Chan, Massachusetts Institute of Technology, United States

- Prof Thomas Chen, City, University of London, United Kingdom
- Prof Michele Colajanni, University of Modena and Reggio Emilia, Italy
- Torsten Corall, NATO CCD COE
- Dr LtCol Christian Czosseck, NATO CCD COE Ambassador; CERTBw
- Prof Dorothy E. Denning, Naval Postgraduate School, United States
- Dr Kenneth Geers, NATO CCD COE Ambassador; Comodo
- Keir Giles, Chatham House, Conflict Studies Research Centre, United Kingdom
- Rudi Gouweleeuw, Netherlands Organisation for Applied Scientific Research (TNO)
- Prof Michael Grimaila, Air Force Institute of Technology, United States
- Dr Jonas Hallberg, Swedish Defence Research Agency (FOI)
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Jason Healey, Columbia University, School of International and Public Affairs, United States
- Prof David Hutchison, Lancaster University, United Kingdom
- LtCol Daniel Huynh, Army Cyber Institute, United States
- Prof Gabriel Jakobson, Altusys Corp; CyberGem Consulting
- Cpt Raik Jakschis, NATO CCD COE
- Taťána Jančárková, National Cyber and Information Security Agency, Czech Republic
- Prof Eric Talbot Jensen, Brigham Young University Law School, United States
- Dr Jan Kallberg, Army Cyber Institute, United States
- Maj Harry Kantola, Finnish Defence Forces
- Prof Sokratis K. Katsikas, Norwegian University of Science & Technology
- Dr Panagiotis Kikiras, European Defence Agency
- Markus Kont, NATO CCD COE
- Jarkko Kuusijärvi, VTT Technical Research Centre of Finland
- Clare Lain, NATO CCD COE
- LtCol Franz Lantenhammer, NATO CCD COE
- Dr Scott Lathrop, Soar Technology, Inc
- Artūrs Lavrenovs, NATO CCD COE
- Dr Sean Lawson, University of Utah
- Dr Corrado Leita, Lastline Inc.
- Dr Lauri Lindström, NATO CCD COE
- Dr Kubo Mačák, University of Exeter, United Kingdom
- Prof Olaf Manuel Maennel, Tallinn University of Technology, Estonia
- Dr Matti Mantere, Nordea Bank AB
- Prof Evangelos Markatos, University of Crete, Institute of Computer Science, Greece

- Dr Paul Maxwell, Army Cyber Institute, United States
- Maj Markus Maybaum, Bundeswehr Cyber Security Centre; NATO CCD COE Ambassador; Fraunhofer FKIE
- Roy Mente, Netherlands Organisation for Applied Scientific Research (TNO)
- Tomáš Minárik, NATO CCD COE
- Maarja Naagel, NATO CCD COE
- Dr Jose Nazario, Fastly Inc.
- Dr Lars Nicander, Swedish National Defence College
- Maj Erwin Orye, NATO CCD COE
- Dr Anna-Maria Osula, NATO CCD COE
- Dr Nikolas Ott, Organization for Security and Co-operation in Europe
- Dr Rain Ottis, Tallinn University of Technology, Estonia
- Prof Stephanie Pell, Army Cyber Institute, United States
- Piret Pernik, International Centre for Defence and Security, Estonia
- Mauno Pihelgas, NATO CCD COE
- Cpt Roy Ragsdale, Army Cyber Institute, United States
- Tarmo Randel, NATO CCD COE
- LtCol Glenn Robertson, Army Cyber Institute, United States
- Prof Gabi Dreo Rodosek, Bundeswehr University Munich, Germany
- Henry Rõigas, Guardtime
- Prof Juha Röning, University of Oulu, Finland
- Ragnhild Siedler, Norwegian Defence Research Establishment
- Dr Max Smeets, Stanford University, Center for International Security and Cooperation (CISAC), United States
- Dr Edward Sobiesk, Army Cyber Institute, United States
- Dr Daniel Spiekermann, FernUni Hagen/German Police Forces
- Dr Tim Stevens, King's College London, United Kingdom
- Dr Kris Stoddart, Aberystwyth University, United Kingdom
- Morta Strazdaitė, Paris School of International Affairs, France
- Dr Michail Sulmeyer, Harvard Kennedy School, United States
- Prof Bradley Thayer, Tallinn University, Estonia
- Dr Jens Tölle, Fraunhofer FKIE, Germany
- Lorena Trinberg, German Armed Forces
- Krista Jean Tuthill, Booz Allen Hamilton
- Prof Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCD COE
- Matthijs Veenendaal, Ministry of Defence (Defence Cyber Command), the Netherlands
- Prof Ari Visa, Tampere University of Technology, Finland
- Prof Col David Wallace, United States Military Academy
- Prof Bruce Watson, Stellenbosch University, South Africa

- Prof Sean Watts, Creighton University Law School, United States
- Cdr Mike Widmann, NATO CCD COE
- Prof Stefano Zanero, Polytechnic University of Milan, Italy

We would like to thank the Institute of Electrical and Electronic Engineers (IEEE) and its Estonia Section for again serving as a technical co-sponsor for CyCon and this publication. In addition, we highly appreciate the NATO CCD COE staff for their excellent organising skills and assistance during the publication process.

Special thanks are due to Dr Joe Burton from the University of Waikato, New Zealand, (16846.304 km away from Tallinn, according to Google) for his contribution to the work of the CyCon 2018 Programme Committee and to the Conference Proceedings.

Finally, we thank all the authors of the papers in this publication for their outstanding submissions, their friendly cooperation, and their efforts in advancing research on cyber security.

The CyCon X Programme Committee

Lauri Lindström
Raik Jakschis
Tomáš Minárik
Ann Väljataga

NATO Cooperative Cyber
Defence Centre of Excellence
Tallinn, Estonia, May 2018

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Rethinking the Data Wheel: Automating Open-Access, Public Data on Cyber Conflict

Christopher Whyte
Assistant Professor
Virginia Commonwealth University

Brandon Valeriano
Donald Bren Chair of Armed Politics
Marine Corps University

Benjamin Jensen
Associate Professor
Marine Corps University

Ryan Maness
Assistant Professor
Naval Postgraduate School

Abstract: To date, researchers studying cyber conflict through publicly available information sources have either selected on the actor or selected on the intrusion method when coding events. Both approaches lead to distinct challenges when it comes to result validation and the avoidance of selection bias. This article describes prospects for open-source, public data collection for cyber security events. We present an initial data collection and analysis effort of interstate cyber conflict incidents involving the United States as a pilot study. Using a tailored collection of more than 155,000 documents from print-only media sources, we describe a method to process data, parse document elements, and populate an event dataset. Human coders are then tasked with validation of incident information, after which the search code is updated to ensure greater accuracy in subsequent runs. In the study, the data produced are compared with previously available data on cyber conflict involving the United States. We demonstrate that the method can effectively capture and describe cyber conflict incidents for researchers to study in a broad range of research efforts. Moreover, this method captures greater granularity within cyber conflict episodes, which are inherently multi-faceted. This approach to cyber conflict analysis carries with it several distinct advantages over alternative research designs, in that it promises to produce significantly larger amounts of pertinent metadata than might otherwise be possible.

1. INTRODUCTION

Researchers analyzing the scope and scale of global cyber conflict face significant data collection challenges. In particular, the process of determining who is responsible for observed cyber incidents that are often covert by design produces research constraints for researchers seeking to describe modern competition, conflict and confrontation empirically (Gartzke and Lindsay, 2015; Rid and Buchanan, 2015). How can researchers systematically study cyber incidents globally and document recurrent patterns and trends, given inherent restrictions on coding what are essentially covert operations?

Such challenges are pressing for scholars and practitioners alike insofar as both aim to develop a sophisticated body of knowledge regarding the drivers, determinants, and effects of conflict waged via networked information and communications technologies (ICT). To date, the cyber security field tends to rely on thin case study descriptions of cyber incidents, using crucial cases to make inferences about actor motivation and the larger context of the cyber conflict, as well as using deductive reasoning to produce a foundation of theoretical knowledge regarding cyber conflict. For example, major work on the Stuxnet attack tends to take this form, with scholars debating the efficacy and larger implications of the series of espionage and degradation intrusions launched by multiple states against Iranian targets (Lindsay, 2013; Slayton, 2017; Kello, 2017). With respect to deductive reasoning, major studies use a series of anecdotal examples to work through a series of logical claims about cyber deterrence and crisis escalation in cyberspace, even including paralyzing cyber first-strikes and offensive action (Libicki, 2012; Gompert and Libicki, 2014; Whyte, 2016; Nye, 2017). Despite its classified nature, most intelligence analysis of cyber events likely replicates these methods. Faced with a poverty of data, analysts and scholars alike use individual incidents and deductive reasoning to illuminate emerging threats and opportunities in cyberspace.

To date, research that systematically collects data on cyber incidents is scarce. Outside of work on cyber rivalry and limited studies of denial of service attacks within a conflict setting, both of which limit the sample under investigation, most of the cyber security literature lacks large databases and robust samples (Valeriano and Maness, 2014; Valeriano and Maness, 2015; Kostyuk and Zhukov, 2017; Whyte, 2017; Valeriano, Jensen, and Maness, 2018). The absence of large datasets limits the development of inductive meta-theories about cyber conflict. Policy makers and scholars cannot determine whether an intrusion event is an isolated and insignificant incident, or consistent with a larger correlate of cyber conflict, without understanding the true scope of cyber interactions.

For scholars interested in the cyber domain, assessment of information derived from publicly-available outlets is an option that is as attractive as it is problematic. The capture and treatment of massive amounts of published data pertaining to cyber conflict promises a unique resource for those seeking to assess the context of cyber security engagements. Nevertheless, such approaches often garner broad criticisms pertaining to generalizability and methodology. If much of what constitutes cyber conflict is covert, how can data produced from information found in the public sphere offer researchers the opportunity to generalize? Even if that hurdle were to be cleared, how can researchers reconcile attribution challenges in determining the sources, targets and technical shape of varied cyber interactions? Without some notion of reliability as a measurement of the value of such information, open source data efforts are likely to run into serious problems.

This article addresses the data challenge at the core of cyber security. First, we address the utility of open-source data collection on cyber conflict processes for scholars and practitioners alike. In addition to being the most promising route available for academic researchers to develop a robust knowledge foundation from which to undertake sophisticated analyses, assessing open access materials both allows researchers to look at the context of cyber conflict and provides opportunities for use of advanced analytic methods that can parse signal from immense noise. Second, we describe an approach – commonly found in research on political violence, and in recent efforts to build comprehensive conflict event data – for producing cyber conflict data that draws from public-facing information sources and allows the researcher to address validation shortcomings inherent to such an approach. Then, we demonstrate the value of this approach by employing a tailored collection of more than 155,000 documents from print media sources in the United States, in order to produce data on interstate cyber interactions across a two-year period. This approach performs on par with data previously produced via traditional collection approaches and, insofar as different elements of episodes are captured, produces a more granular picture than has been produced in prior large-N work on cyber conflict. Likewise, opportunities to enrich such data via additional treatment of surrounding text and linked documentation promise further value to researchers seeking to understand the sociopolitical context of cyber conflict (Schrodt, Beieler, and Idris, 2014).

The article proceeds in five sections. First, it considers the state of cyber conflict data production, describes the few attempts that have been made to date to produce systematic accounts of warfare conducted online, and outlines enduring challenges. Then, we make a case for the clear utility of data produced from publicly-available information sources. Third, we describe the requirements for robust, replicable efforts to develop such data resources for scholarly use, before demonstrating this via the presentation of two years' worth of event data on interstate cyber conflict involving

the United States. We conclude with a discussion of the implications of our arguments, and a demonstration for both researchers and policymakers as well as practitioners.

2. CYBER CONFLICT DATA: PRIOR EFFORTS AND ENDURING CHALLENGES

The incidence of cyber conflict dates back to the early 1980s with episodes such as the Farewell incident, in which the CIA targeted KGB technology transfer programs, and the Cuckoo's Egg hack-back, in which private network operators identified Soviet operatives (Stoll, 1988; Healey and Grindal, 2013). In spite of this, systematic and comprehensive resources describing cyber conflict incidents are virtually non-existent. Major political science efforts to catalogue different forms of interstate conflict and political violence fail to include cyber actions, either owing to their ambiguous origins or to difficulty attributing the incident. Stuxnet, for example, although a crucial case in descriptive treatments, is often not represented in major databases due to attribution issues, difficulty dating the start and end of the incident, and the question of whether it was the United States or Israel that launched the action (Radford, 2016).

This general lack of focus on cyber conflict issues in the context of broader efforts to record and problematize international security dynamics is troubling for a number of reasons. Foremost among these is the fact that there is arguably a consensus among political scientists that cyber instruments work as adjunct modifiers – essentially force multipliers – of conventional and asymmetric warfare (Gartzke, 2013, Valeriano et al., 2018). This suggests that cataloguing cyber incidents is useful not only as a means of assessing conflict restricted to that domain, but also as a means of understanding a critical variable in broader conflict processes. Without better understanding of the nature of cyber conflict, scholars and security practitioners of all stripes are (and will be) hard pressed to describe accurately how digital actions and possibilities intersect with existing mechanisms of human interaction. Indeed, without such a development, it is likely that we inject bias – from using data obtained only from select stakeholders or employing methods that misunderstand the significance of different actors – into our continued efforts to construct knowledge of macro global security processes.

The main reason that no comprehensive data resource to describe cyber conflict exists is that the attribution of cyber incidents is not always feasible (Rid and Buchanan, 2015; Lindsay, 2015). This is true on two fronts. Firstly, the method is covert: while there are often observable outputs of cyber conflicts, where victims (or, in rare instances, observers) report on incidents or attackers broadcast their involvement, this is not always true. Indeed, anecdotal evidence and simple recognition of the scope of the domain to be canvassed by researchers suggests that this is true only infrequently.

Bound up in this problem is the manner in which the digital world operates. Whereas with other forms of conflict – terrorist attacks, for instance – it may be possible to adjudicate reasonably on the frequency of otherwise invisible attacks based on knowledge of past actions, analytic breakdown of capabilities, or journalistic efforts to validate rumor, the same is not generally accepted in cyberspace. Even where indicators suggest the existence of incidents to researchers, validation usually requires the cooperation of victims or infrastructural stakeholders (i.e. backbone operators or non-backbone ISPs). Thus, particularly where relevant actors are motivated by the possibility of reputational, financial or political costs, confirmation of the full scope of cyber conflict is difficult for those operating in the public domain.

Added to these challenges are the dual problems of bounding scale and controlling for negative cases. With respect to scale, a successful cyber operation might involve thousands of individual intrusion incidents. For example, spear phishing campaigns that resulted in the compromising of the German Bundestag and, more recently, the U.S. Senate, involved hundreds of e-mails sent to unsuspecting elected officials and staffers.[1] Does each e-mail constitute an individual cyber intrusion, or can researchers include them all as one campaign? Regarding negative cases, researchers must acknowledge the fact that cyber security firms, journalists, and governments tend to report only successful intrusions or attempts that nevertheless cause at least some measure of disruption (Brodsky, 2008). Unsuccessful intrusions, which likely are significantly larger by count, are thus under-reported.

Similarly, the second facet of the reporting problem lies with the value of information that can be obtained. Though such challenges are often surmountable, as we describe below, it is certainly true that gathering enough detail on a given incident to allow sociopolitical attribution is possible but difficult. Despite the clear imperative social scientists have to use any and all information available in attempting to understand the world around them, efforts to understand cyber phenomena better regularly run into criticism, as operating in a covert domain will generate no observable data (Lewis, 2002). This point fundamentally misunderstands the meaning of covert action, however, which implies a difficulty in determining responsibility, but not whether or not the event occurred.

Datasets are routinely released in the broad international relations field cataloguing all manner of security phenomena.[2] Among these, a small number are broadly focused on conflict with a relatively unlimited remit. Rather than focus solely on the efforts of terrorist non-state actors, insurgent movements, social activists or state militaries, such data collection efforts aim to catalogue the full spectrum of conflictual incidents

1 See *inter alia* http://www.zeit.de/digital/2017-05/cyberattack-bundestag-angela-merkel-fancy-bear-hacker-russia and http://thehill.com/policy/cybersecurity/368671-russia-linked-hackers-targeting-us-senate.

2 See, for instance, the Militarized Interstate Dispute dataset (http://cow.dss.ucdavis.edu/data-sets/MIDs) at the Correlates of War project, the International Crisis Behavior project (https://sites.duke.edu/icbdata/) and the Uppsala Conflict Data Program (http://ucdp.uu.se).

around the world. Over the past few years, such efforts have rapidly become more sophisticated. Efforts like Phoenix[3] and the Integrated Conflict Early Warning System (ICEWS)[4] provided extremely granular information on the nature of security events worldwide using a series of automated data scraping, parsing and treating methods, often in tandem with human validation inputs. Such approaches constitute the new normal for political science researchers in terms of the resources being made available to study international conflict. And yet, these macro efforts to describe global security matters do not systematically aim to capture all manner of cyber incidents (though they may include individual, prominent events) as part of their approach. This is possibly because the various attack chain elements that constitute the wide array of techniques of interest to cyber conflict scholars are not obviously conflictual in nature, and thus present a challenge when determining inclusion.

To date, there is only one dataset that accounts for all actors, states, and regions in the world available to scholars interested in the contours of global cyber conflict. The Dyadic Cyber Incident and Dispute dataset (DCID) describes interstate cyber conflict over more than fifteen years and employs a Correlates of War (CoW)-style coding scheme to describe the character of cyber warfare campaigns among rival states. The authors of DCID, Valeriano and Maness (2014, 2015), include a range of information on the type of instruments involved in observed cyber events, the impact of such events, and more. The data collected originates from publicly-available descriptions of cyber conflict incidents, including news stories, industry and government reporting, and expert testimony. Nevertheless, as the authors freely admit and others note (Radford, 2016), DCID was designed as an initial effort to scope the cyber conflict domain by selecting on rival states most likely to engage in cyber conflict. It is not aimed at the production of cross-domain conflict data, and does not draw from the universe of possible information on cyber incidents in a comprehensive sense. While outputs of the project might describe contours of cyber conflict between rival actors, any comprehensive effort to produce cyber conflict data must inevitably drop such selection parameters in order to ensure generalizability. Thus, the need to address the role of future open source data collection on cyber conflict is twofold, insofar as researchers must grapple with *both* absent resources and limited foundational efforts from which to begin their investigations.

Briefly, the data collection approach we describe below addresses these dual needs and goes a step further than previous social science projects. We rethink prior approaches to data collection in line with work undertaken in political violence and terrorism research programs, and expand beyond a limited focus rival states. In doing so, we provide for reliability checks that have been absent – or hard to effect – in past efforts, and argue that sophisticated data collection in this vein must turn to human reliability checkers for all machine learning processes. The result would be a dataset

3 See http://openeventdata.org/datasets/OEDA.datasets.php.
4 See https://dataverse.harvard.edu/dataverse/icews.

both large and relatively free of the errors common to other large event databases, such as ICEWS (Boschee et al., 2015) or IDEA (King and Lowe, 2003). Part of the reason we argue that this will be the case is the fact that projects like ICEWS and IDEA aim to capture all events between all actors annually. A cyber conflict effort would include a significantly reduced scope of inquiry, and would make the parsing of signal from noise a more feasible task. In short, though we cede the point that there are limitations to any open-source data collection effort on cyber conflict patterns in the form of lagged information about cyber threats that occur in clandestine settings, such an effort would regardless lead to a useful resource useful to cyber-security scholars across a range of disciplines, upon which others can build in the future.

3. THE UTILITY OF OPEN SOURCE DATA COLLECTION

Open-source collection of information on cyber conflict processes represents the future of data generation in the field, but also presents many challenges. Whereas most open source data collection seeks to parse signal from noise, a cyber conflict effort will miss things simply because not all of the signals are observable from the public sphere.

Why should researchers even attempt to undertake open-source collection of information on cyber conflict trends, given the inherent problems in doing so? We argue that there are three reasons. First, social science research on cyber conflict requires a foundation of knowledge from which to build and infer. Second, assessing open-access description of cyber conflict allows researchers to look at both the content and context of cyber interactions. Third, there are distinct benefits to a scaled-up approach to studying cyber conflict over traditional small-n approaches, as there is additional clarity and opportunity to use advanced analytic methods to parse observable relationships.

The Need for a Knowledge Foundation

Most simply, there is a clear need for foundational knowledge about cyber conflict. At present, there is a relative lack of empirical work in the domain that presents a comprehensive and systematic description of the global impact of the information revolution. One clear argument in favor of scholarly attempts to build a representation of such processes via collection of public-facing information is quite simply that scholars are duty-bound to utilize any resource available in trying to contribute to the condition of knowledge on a given topic.

More pressing than the duty of social scientists, however, is the need to develop knowledge foundations in order to spur the development of a robust research

program. The nature of the development of research programs is a source of hot debate among both classical and current philosophers of the social science enterprise. It is generally acknowledged, however, that research programs are layered bases of theoretical knowledge where peripheral hypotheses linked to core suppositions are appraised with the aim of advancing the state of a given field (Jackson, 2008). Often, hypothesis testing results in rapid rethinking of specific assumptions such that there is a revolution in macro knowledge. In the debate about progress in the field of International Relations, Lakatos is often invoked as the exemplar for establishing which theoretical ideas are of value over others (Vasquez, 1997). This view requires the development of a theoretical and empirical core, which then is investigated with the purpose of seeking advances over prior investigations. Advances can be examined in the context of providing more theoretical and empirical context over past efforts (Lakatos, 1970).

At present, the research program on cyber conflict is still in its infancy. The condition of general core knowledge at the heart of the research program is remarkably unclear, which suggests that there is a strong imperative to articulate macro-theoretical perspectives. Given this, the need for projects that aim for comprehensive modeling of the scope of global cyber conflict is particularly pronounced.

The Context of Cyber Conflict

Building from the perspective that meaning emerges from the interaction of empirical dynamics and the human treatment thereof, researchers should attempt to undertake open-source collection of cyber conflict trends. Such an approach will inevitably capture more than just the actuarial detail of cyber incidents offered by thick case descriptions; specifically, open-source data collection allows researchers the opportunity to understand the context and content of cyber conflict dynamics more fully. Via the capture of textual metadata, cataloguing of adjacent conflict phenomena, and more (Hopkins and King, 2007), open source data modeling of cyber conflict trends (given relevant controls for duplication of information) offers the ability to understand the nature of information about cyber conflict that exists in the public sphere. Social science scholars of cyber conflict are, for instance, naturally interested in how framing of conflict influences the discourse and deliberation of policymakers, practitioners, and the general public. Is a particular cyber event over-reported in news media? What kinds of information are used in public discourse to construct attribution cases, and do these assessments vary given the context of, say, ongoing foreign policy spats with particular foreign countries? Do certain kinds of attacks receive more negative coverage, and how are relevant stakeholders discussed in such coverage? Understanding such dynamics is critical to efforts that aim to gain a systematic understanding of public reactions to cyber threats, the manner in which the citizenry ascribes responsibility for cyber security to public or private sector actors,

and more. Public-facing data promises an ability to answer fundamental questions about the relationship between cyber conflict and the sociopolitical environment in which foreign policymaking and strategy development take place. Answering such questions should be of paramount importance to scholars.

The Benefits of Scale

Finally, efforts to scale up data collection using computer coding, web scrapping, and machine learning exponentially increase the available data. This universe of big data provides an empirical foundation from which to sort signal from noise in a way that is difficult to do where less input data is involved. This effort requires narrowing search terms based on automated construction of parameters and machine learning, followed by subsequent Bayesian updating of the process based on human review and validation of subsets of input data (as described in Hopkins and King, 2010; Ward, Beger, Cutler, Dickinson, Dorff, and Radford, 2013). At the level of the research project, the benefits of such an approach are obvious. With ICEWS, researchers reported a 50% increase in accuracy with semi-supervised approaches using large amounts of input information over those that had previously attempted only to have machines sort raw data. In essence, sophisticated application of an ontological understanding of conflict processes in coding massive amounts of data allows dissection of information in a way that is not possible with small samples.

At the level of the research program on cyber conflict processes itself, the clear benefit of scale is clarity. Given inherent attribution issues associated with cyber incidents, researchers need to cast their net as wide as possible to include not just major media outlets, but also government documents and cyber security reporting. Cyber security firms in particular are a critical source of reporting. These third-party firms have a financial and reputational incentive to report on the nefarious acts of government operatives online. They are constantly monitoring and looking to expose major intrusions (see, for instance, Kaspersky, 2015). Shifting to a machine-coding scheme that collects disparate sources brings these perspectives together in building a cyber security incident database. The combined observations, even if still imperfect, are orders of magnitude better than any one reporting line.

Put together, each argument for the construction of a larger-event based dataset of cyber interactions is not only needed, but prudent and responsible. The production of knowledge is a process fraught with friction, but we can reduce the hindrances common at the start of such enterprises by seeking to establish an empirical baseline early in the lifespan of a research program. Now we move to a formal description of how such a process of data collection takes place, and observe our results in the pilot study.

4. BUILDING A LARGE-SCALE DATA COLLECTION AND TREATMENT PIPELINE

Machine-coded event datasets such as Phoenix or ICEWS are developed using publicly-available resources.[5] To date, most efforts in political science have used news stories scraped from RSS feeds, repositories like Factiva, and outlet websites. It is, however, possible to draw information from any text resource. Although researchers are likely to favor news stories of various kinds for event data production, it is possible to utilize social media data feeds and information like industry reports.

The production of event data from large corpora is relatively straightforward. Unstructured information is taken from feeds and repositories using the researcher's favored method of text crawling and fed into a database program. From there, information can be sent in a specified format to a program that produces structured, usable event data. A number of such programs exist, but the most well known are TABARI/PETRARCH/PETRARCH2, a series of Python-based programs that treat text and produce data. The function of these programs is also relatively straightforward. Text inputs are broken down to the level of individual sentences and are parsed to produce an XML input that includes both the original text and a language element breakdown. From there, files are passed through the main program, which references a series of preset dictionaries to produce structured data. The dictionary inputs represent the expected vocabulary pertaining to a given topic and are designed by the researcher.[6] The resultant structured data are then usable by researchers or are available for further enrichment. Up to the point described here, data output by a program like TABARI would include event description, source and target information, and metadata (date, source of information, etc.). Further enrichment of this data for the purposes of understanding the context or surrounding content can then be achieved via further application of a range of text modeling, entity extraction, and topic modeling tools, with human interaction only required when specifying input text or when making a particular effort to enrich descriptive event data.

5 The same is true for both data based on the Conflict and Mediation Event Observations (CAMEO) framework (Gerner, Schrodt, Yilmaz, Abu-Jabr, 2002) and the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013). These efforts, and earlier ones like the Conflict and Peace Data Bank (COPDAB) and the World/Event Interaction Survey (WEIS) (Azar, 1980), provide granular information on human behavior drawn from an immense collection of available public sources of input data. CAMEO and other frameworks are employed for the purposes of structuring and making sense of the resultant information for analytic purposes.

6 Recently, some advances have been made in automatically generating dictionaries based on the input text (Radford, 2016) specifically in the context of cyber security.

5. THE UNITED STATES' EXPERIENCE WITH INTERSTATE CYBER CONFLICT, 2013-14

In order to demonstrate the utility of such a machine-coded event data production approach to comprehensively scoping the cyber domain, we supplement our arguments here with an application of PETRARCH2[7] to a limited corpus of news stories pertaining to cyberspace and information security issues published in the United States. After discussing our data production effort, we present data below on incidents involving the United States and other countries, and compare our results to those of the only existing cyber conflict data resource (DCID). Though this demonstration is a limited, proof-of-concept effort that focuses on two years and one country's relationships with other countries, we note that results match and arguably outperform those of DCID. Given that this data emerges from a relatively small scrape of available information on national cyber security events, the opportunity for expanded efforts seems clear.

Constructing a Demonstration Dataset Using Machine-Coding

The foundation of our demonstration dataset is a corpus of documents downloaded from LexisNexis. The documents that make up our corpus were selected based on two sets of criteria. First, we select on only United States-based print and wire publications so that we can effectively gauge the viability of a machine coding approach to event data production at the level of an individual country. Second, we collate all news articles that correspond to an extensive formula of keyword collocations that aim to capture all coverage of cyber security issues. The result is an extensive corpus of more than 155,000 news stories across more than thirty years. For purposes of matching outputs to DCID and assessing the viability of a machine-coding approach in the context of the contemporary landscape of cyber conflict, our construction of the demonstration dataset presented below focuses on a two-year period between 2013 and 2014. Specifically, data is drawn from 859,423 input text files at the level of individual statements (sentences).

Raw text taken from LexisNexis is passed through several stages of treatment prior to the output of structured event data. First, text is parsed using the Stanford Core Natural Language Processing (NLP) suite of available programs, which tag named entities and parts of speech (i.e. nouns, adjectives, verbs, etc.) found in the text (Manning, Surdeanu, Bauer, Finkel, Bethard, and McClosky, 2014). The parsing process outputs an XML file that details a breakdown of different language elements. This provides the constituency tree parse necessary for event coding using PETRARCH2. Then, a glue program is used to format raw text chunks and the parsed language information into a file format specified by the authors of PETRARCH2 (see *inter alia* Beieler, 2016). Finally, these files are passed to PETRARCH2 for analysis. Analysis of text fragments at the level of sentences works via reference to a series of dictionaries to which the

7 See https://github.com/openeventdata/petrarch2.

program refers. These dictionaries contain vocabulary for types of conflict actions to be coded, agent types to be considered, and actors that might specifically be identified; the dictionaries can be automatically generated (Schrodt, Beieler, and Idris, 2014; Radford, 2016) but are generally updated manually by the researchers, as was the case here. The resultant data output includes information on the type of conflict action recorded, the source of that action, the target of that action and metadata pertaining to the incident (date, type of agent in the context of a particular actor, etc.).

Resultant Data on U.S. Experience with Interstate Cyber Conflict, 2013-14

Our demonstration set of incident records includes 512 distinct events for the two-year period between January 1, 2013 and December 31, 2014. Of those events, 279 events pertain directly to the United States insofar as the machine-coding process identifies either the originator or target as being American. This is not to say that the United States government or a particular federal entity is linked with every event; rather, this number refers to any actor (often named but sometimes an unknown hacker) that is identified as having a relationship with the United States (i.e. an American firm, individual or domestic person, for instance). Of events that link an incident directly to the United States (as a discrete entity) or the U.S. government, the U.S. is coded as the originator of a cyber conflict incident in 151 instances, and as the target in 91 instances.

FIGURE 1. NUMBER OF CYBER CONFLICT INCIDENTS INVOLVING THE U.S. (TOTAL), 2013-14.

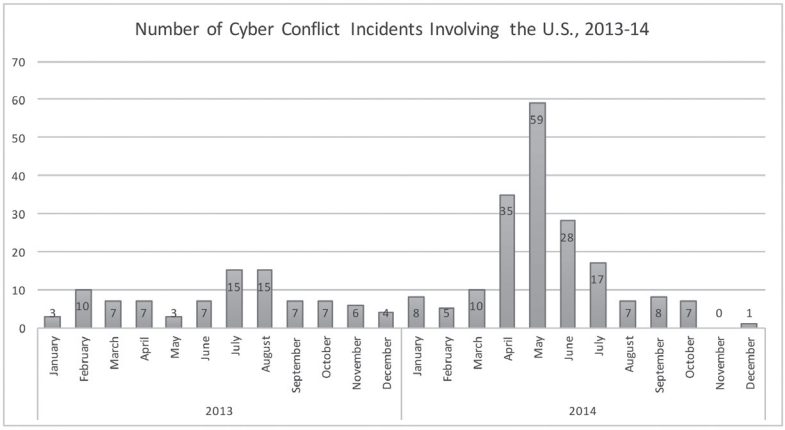
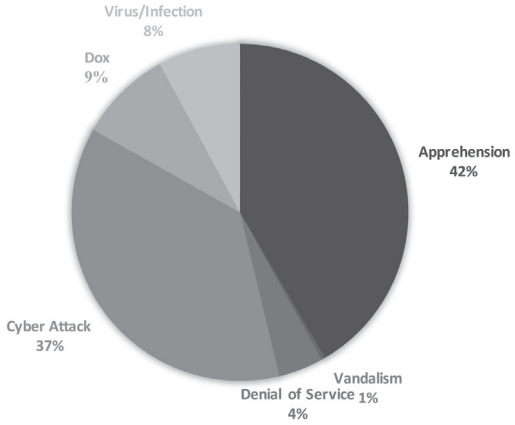


Figure 1 presents the raw count of incidents involving the United States (total attribution, not only government or national attribution) captured in our demonstration machine-coding effort for the years 2013 and 2014. Of these 279, the bulk are identified from March through July of 2014. This is perhaps unsurprising, as this constitutes

the period of time immediately following data breaches at Target, Inc. The Target hacking episode stands as one of the first major instances of a major private firm in the United States going public with the theft of information pertaining to millions of consumers. This period also follows the release of information by Edward Snowden at the end of 2013 pertaining to U.S. cyber operations and electronic surveillance programs, as well as intrusions at the Office of Personnel and Management (OPM) which would stay secret until early 2015. It is worth noting, however, that this data includes both government and non-government activity as captured in open-source reporting, potentially including criminal actions and espionage.

FIGURE 2. TYPES OF CYBER CONFLICT EVENTS CAPTURED INVOLVING THE UNITED STATES, 2013-14.

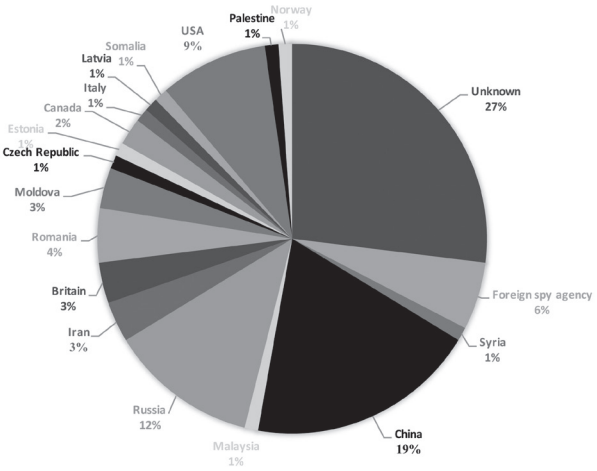


Our test dataset also captures information about the nature of different conflict actions. At the highest level, our approach presents the researcher with six categories of cyber actions – denial of service, vandalism, generic cyber intrusion, malware usage/infection, information doxing, and the apprehension of an involved actor. The denial of service and vandalism categories capture events that specifically reference the terminology of defacement and DDoS. The infection category captures incidents that reference the discovery or presence of a piece of malware based on a set of preset terms and specific malware instances (added to the program dictionary). Cyber intrusions generically refer to cyber actions linked with terminology indicating use of force (‘attacked,’ ‘hacked,’ ‘breached,’ ‘infiltrated,’ etc.) and can therefore cover a wide array of incident types. Apprehension events include instances where perpetrators of an act are caught, arrested or identified. Doxing events include those wherein information is intentionally leaked or released.

Figure 2 breaks down the set of incidents we found involving U.S. actors (as either originators or targets) in 2013 and 2014. By far, the most common incidents recorded

are the apprehension of actors and generic cyber intrusions. Apprehension incidents are coded in a relatively straightforward fashion in that PETRARCH2 identifies language elements pertaining to the arrest and capture of people. Again, cyber intrusions are coded in such a way that a broad number of methods and techniques can produce a cyber intrusion event (such as hacking, intruding, gaining access, injecting code, etc). By contrast, denial of service attacks and digital vandalism are rare in this data set, whilst the leaking of information and incidence of malware (wherein input text does not suggest an attacking action) are uncommon.

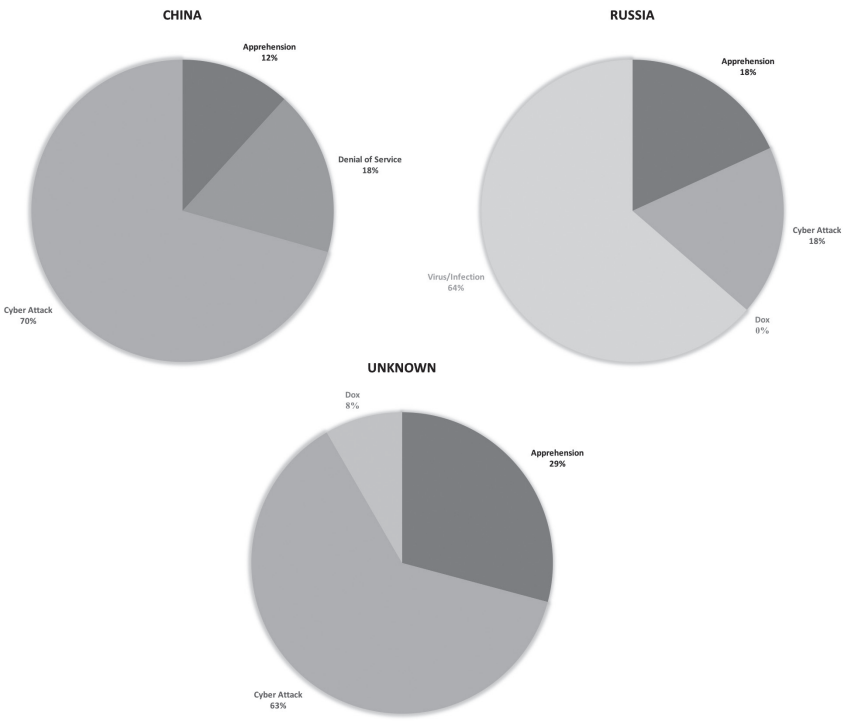
FIGURE 3. SOURCE COUNTRIES FOR CYBER OPERATIONS TARGETING THE UNITED STATES (GOVERNMENT/MILITARY TARGETS).



By means of demonstrating the manner in which machine-coding approaches are useful for capturing attribution within dyads (i.e. where one actor can be seen to have engaged with another), *Figure 3* outlines originator countries for all actions on targets coded as either U.S. government/military targets or 'the United States.' As above, these originator countries are not necessarily identified as government/military/intelligence targets, although many are. It is worth noting that the largest category is 'unknown', where the program is unable to identify a country with which to link a cyber conflict action; this result in itself highlights the attribution challenge faced by researchers in this vein. In almost no instances does this mean that there is no information on the originators of actions; rather, source information is most often tagged at the level of agent types, meaning that no country or specific threat actor can be identified, but the program identifies the originator as a foreign individual or criminal organization. Following this category, the next categories of action are linked with the Russian Federation, the People's Republic of China, and countries linked in analytic work on global cyber conflict with both these countries, such as Moldova and Malaysia. A relatively high percentage of attacks attributed to the U.S. were incidents

where U.S. individuals or groups were involved in cyber conflict actions (mostly being apprehended by authorities).[8]

FIGURE 4. TYPES OF CYBER CONFLICT ACTIONS FOR ATTACKS ON THE UNITED STATES FOR CHINA, RUSSIA AND UNATTRIBUTED INCIDENTS.



Among the three largest originators of conflict actions targeting the United States and its government or military-intelligence apparatus, cyber intrusions are the most common type of event for both China and nationally un-attributable actors. Intrusions might include a wide range of possible techniques, but generally refer to a forceful infiltration without permission, as exemplified in incidents like the OPM hack. With Russia, however, although a substantial percentage of actions linked with the country are generically coded as cyber intrusions undertaken against the U.S., the bulk of coded cyber conflict actions are coded as malware infections. Though such a conclusion is purely speculative, this trend does fit with the narrative of existing research on the nature of global malware distribution, the role of Eurasian organized criminal enterprises in underwriting major ransomware, denial of service and phishing attacks,

8 Regarding the methodological challenges facing the researcher in assessing cyber conflict processes, another point worthy of note off this finding is the degree to which offensive deception is not only possible, but normal. Operators may take steps to mask their point of origin when launching offensive or exploitative actions. See, for instance, http://www.star-telegram.com/news/nation-world/national/article96062667.html.

and on the unique character of the Russian cyber ecosystem that leverages third-party criminal enterprises (Valeriano, Jensen, and Maness, 2017).

Capturing Major Events

The data described above represent only a limited demonstration of how a machine-coding approach to open-source data collection can furnish scholars with unique information about the scope of global cyber conflict. But does the method of approach really function better than traditional human equivalents? Can automated coding of event data match or outperform the research skills of human coders wading through similar information in order to parse signal from noise?

Here, we briefly consider these questions by comparing the results of our demonstration dataset to the preceding DCID cyber conflict data collection effort. Specifically, we ask if incidents involving the United States during 2013 and 2014 that are catalogued in DCID were captured by our initial coding of cyber conflict incidents using an input set of information drawn from all U.S. newspaper sources. Given that our selected input source is news reports, the band of incidents we are most interested in assessing here is those cyber conflict interactions that begin within the period covered (i.e. on or later than January 1, 2013). DCID contains 21 such incidents, which are detailed in *Table 1*.

TABLE 1. CYBER CONFLICT INTERACTIONS (DCID) BEGINNING AFTER JANUARY 1, 2013.

Incident	Start Date	Description
Iron Tiger	1/15/13	Sophisticated APT information theft on US military
Black Coffee	4/1/13	APT17 group hacks Microsoft Tech Net forum
NMCI Hack	9/23/13	Navy and Marine Corps unclassified intranet briefly breached
UConn Hack	9/24/13	Chinese hack steals University of Connecticut data
Operation Pawn Storm	9/30/13	Trojan campaign against Blackwater, State Dep't, and SAIC
Saffron Rose	10/23/13	Information theft campaign on US Aerospace industry
Operation SnowMan	2/1/14	Chinese hackers infiltrate VVW to access military personnel info
Register.com	3/1/14	Register.com, which manages more than 1.4 million websites for businesses world wide, steals network and employee passwords
OPM Hack	3/15/14	OPM hack, personal information of 20 million people stolen
CyberBerkut	3/15/14	Signal NATO members to avoid intervention in Ukraine
Premera Blue Cross	5/5/14	State-sponsored Chinese data breach group steals personal information of 11 million Premera customers
Operation Pawn Storm #1	6/2/14	Backdoor intrusion in to military networks via spear phishing
Operation Pawn Storm #2	6/3/14	Backdoor intrusion in to several commercial networks via spear phishing
US Banks Hack	6/4/14	Retaliation on US targeted sanctions on Russia
UCLA Health Breach	9/1/14	State-sponsored Chinese group, 4.5 million records stolen
White House Hack	10/26/14	White House email server compromised,
DHS Hack	11/6/14	25,000 DHS employees' information stolen from OPM
USPS breach	11/8/14	Personal information of 800,000 USPS employees compromised
State Dep't hack	11/15/14	State Dep't unclassified email system breached and contained
Sony Hack	11/24/14	Sony Pictures is breached and secretive information leaked
Anthem Breach	12/10/14	Black Vine hacker group (China-sponsored) steals sensitive information from health insurance giant Anthem

Our demonstration dataset produced and presented here records events pertaining to 14 of the 21 cyber conflict interactions beginning after January 1, 2013 in the DCID dataset (see *Table 2*). Importantly, incidents not captured by the machine-coding treatment of news stories from the United States largely fall at the end of the period covered. This implies that non-capture is the result of a delay in reporting cyber incidents, and that this issue will be alleviated by a larger time span examining disclosures that happen at a later date (as with the OPM hack, which was revealed in 2015). Moreover, the demonstration dataset contains 1.301 events for each interaction described in DCID, meaning that the average incident described there is matched by more than one reported interaction (even after controlling for duplicates) in the machine-coded version. For instance, the University of Connecticut hack in 2013 was caught twice, with one event annotation describing the infection of computers at the institution, and a later report describing a purposive cyber intrusion aimed at stealing user information.

TABLE 2. CYBER CONFLICT INTERACTIONS IN DCID
(BEGINNING AFTER JANUARY 1, 2013) CAPTURED BY DEMONSTRATION SET.

Incident	Start Date	Description	Recorded?
Iron Tiger	1/15/13	Sophisticated APT information theft on US military	Y
Black Coffee	4/1/13	APT17 group hacks Microsoft Tech Net forum	Y
NMCI Hack	9/23/13	Navy and Marine Corps unclassified intranet briefly breached	N
UConn Hack	9/24/13	Chinese hack steals University of Connecticut data	Y
Operation Pawn Storm	9/30/13	Trojan campaign against Blackwater, State Dep't, and SAIC	Y
Saffron Rose	10/23/13	Information theft campaign on US Aerospace industry	Y
Operation SnowMan	2/1/14	Chinese hackers infiltrate VVW to access military personnel info	Y
Register.com	3/1/14	Register.com, which manages more than 1.4 million websites for businesses world wide, steals network and employee passwords	Y
OPM Hack	3/15/14	OPM hack, personal information of 20 million people stolen	N
CyberBerkut	3/15/14	Signal NATO members to avoid intervention in Ukraine	N
Premera Blue Cross	5/5/14	State-sponsored Chinese data breach group steals personal information of 11 million Premera customers	Y
Operation Pawn Storm #1	6/2/14	Backdoor intrusion in to military networks via spear phishing	N
Operation Pawn Storm #2	6/3/14	Backdoor intrusion in to several commercial networks via spear phishing	N
US Banks Hack	6/4/14	Retaliation on US targeted sanctions on Russia	Y
UCLA Health Breach	9/1/14	State-sponsored Chinese group, 4.5 million records stolen	Y
White House Hack	10/26/14	White House email server compromised,	Y
DHS Hack	11/6/14	25,000 DHS employees' information stolen from OPM	Y
USPS breach	11/8/14	Personal information of 800,000 USPS employees compromised	Y
State Dep't hack	11/15/14	State Dep't unclassified email system breached and contained	Y
Sony Hack	11/24/14	Sony Pictures is breached and secretive information leaked	N
Anthem Breach	12/10/14	Black Vine hacker group (China-sponsored) steals sensitive information from health insurance giant Anthem	N

Given these basic results, we argue that it is reasonable to expect that machine-coding of cyber conflict information can at least match human coder efforts. Indeed, since automated coding of cyber conflict incidents invariably captures the detail of particular actions, it seems reasonable to say that event data production using programs like PETRARCH2 quite clearly outperforms all prior traditional efforts because the scope

is much more comprehensive than a selection on rivals. Specifically, the capture of unique features of different elements of a cyber conflict campaign is a natural byproduct of the heuristic-style approach taken by such programs to describing conflict.

Moreover, machine coding of large quantities of publicly-available and publicly-produced textual information stands to help researchers significantly in addressing attribution challenges with cyber conflict research. Though *political* attribution of cyber attacks is not always feasible and technical attribution is enduringly challenging – if not actually impossible – the use of open source documentation offers researchers advantages on two fronts. First, scale brings with it options for verifying the existence of a particular event (and agency therein) in the form of replicable coding rules that, for instance, only report an incident feature that appears in multiple independent reports. Second, open source data collection generates information that is contextually defined. Regardless of whether or not one considers an effort along these lines to be 100% accurate or not, it is indisputably the case that data collected will reflect the state of public knowledge on a given incident. This is significant because much of what social scientists aim to study with cyber conflict patterns is based on context and perception.

Next Steps

No data collection program of approach is perfect. Both this research team and others attempting to produce a reasonably comprehensive data on global cyber conflict using machine-coding of open source information must grapple with distinct methodological issues over and above the macro challenges of such an approach, as described in the sections above. In addition to this challenge, we must also grapple with the construction of additional independent variables in the composition of cyber security data such as indicators of severity, effects, efficacy, actors, cascades, malware tools, and other associated variables.

From our experience in producing the demonstration dataset employed in this section, we argue that two specific methodological challenges in particular are worthy of attention. First, any major effort to leverage state-of-the-art event data production approaches in this vein must consider the fact that available tools remain relatively dumb. That is to say that tools like TABARI and PETRARCH are entirely focused on extracting meaning from a relatively simple understanding of how language works at the level of the statement. This inevitably leads to errors that need to be checked by human coders when, for instance, the program fails to recognize that a particular event is being offered as a hypothetical.

Correcting such errors might take one of several forms. Simply put, however, the idea for researchers moving forward – the gold standard approach – should be a

hybrid approach consisting of what has been presented here alongside relevant human reliability coding for the purposes of more effectively training algorithms for automated coding. Far from suggesting that researchers use preset understandings of cyber conflict ontologies expressed in dictionaries set by scholarly panels, future work should construct and continually reconstruct the tools of event detection from the collections of information being processed. Doing so will allow researchers to control for several things, not least potential problems with the irrelevance of robustness checks as work is scaled upwards, and the shifting terminology – and even the changing nature – of cyber conflict.

Of course, this first challenge leads to additional work for the researcher that might, in the future, be remedied with increased reliance on machine learning augmentations of current approaches. The second (related) major challenge is that researchers aiming to produce event data must recognize that incident capture is often only meaningful alongside the relevant capture of contextual metadata. Enrichment of event data with information about its construction, framing and more stands to benefit researchers from many disciplines and provides deep detail that compensates for the necessary position researchers must take in producing data that will – at least in terms of how much of cyber conflict can truly be observed – be good, but perfect. Moreover, in the research program on cyber conflict, addressing the attribution problem effectively means providing for uncertainty in empirical investigations. Without appropriate efforts to ensure that quality and certitude metrics are provided by researchers alongside a host of metadata on the presentation of raw information pertaining to cyber conflict, efforts to produce comprehensive resources for the research program will be enduringly limited.

6. CONCLUSION

Though the scope and scale of cyber conflict has grown exponentially over the past four decades, scholarly efforts to examine the domain in a comprehensive fashion remain lacking. To some degree, as we have outlined above, this makes sense as there are real challenges for researchers in the form of attribution difficulties, timing of disclosures, and self-interested gatekeepers of useful data. Given these barriers, lack of enthusiasm for and interest in setting up open source efforts to produce cyber conflict event data is understandable.

We have argued, however, that there is both a clear need and a compelling set of reasons for the development of machine-aided, large-scale data production efforts that utilize public-facing information. Though some argue that open source coding of cyber conflict incidents is impossible due to the covert nature of many acts in

the domain, we both argue and demonstrate that this misstates the issue for security researchers. Data coding carried out in this way both (1) parallels the contours of previous data produced on the subject and (2) additionally provides information on the sociopolitical context of cyber operations. In short, not only does the scope of such an approach to data collection promise an ability for researchers to generalize and cross-validate; it also provides the tools to study cyber conflict in its proper international context, examine the tools utilized in each attack, and understand the nested socio-political dynamics at work during cyber conflicts.

Over and above other factors, an effort to provide comprehensive data on the scope of global cyber conflict as it presents in public-facing information sources stands to give researchers the tools needed to build a robust knowledge foundation. At present, the research program on cyberspace and international security lacks an extensive set of core theses and assumptions that can be challenged. Part of the reason that such a core has been slow to develop is that building bridges between otherwise disparate efforts to flesh out specific topics within the research program is extremely difficult without such a comprehensive data foundation. Even if such a foundation were to contain flaws, it would still function as a common platform upon which researchers could situate meaningful research questions and assumptions, contextualize small-n research, and critique methodological approaches. Naturally, this kind of methodological approach will not include – but rather will stand to augment understanding of – the 'thick' context of cyber conflict, from strategic and institutional cultures to cognitive processes. As projects from Correlates of War to those of the Political Instability Task Force have demonstrated on numerous fronts, however, event data and inferences made from them are necessary elements of field-defining research.

Finally, such an effort to build open source data resources also directly stands to benefit policymakers and practitioners. In addition to the clear added value that comes with improved scholarly knowledge of a given topic, academic data resources might be used by both public and private sector actors as a reference to help excise conjecture from the discourse. An academic basis of knowledge on cyber conflict, founded on a common data resource, affords practitioners the opportunity to involve themselves in scholarly and public debate on issues that can be corroborated without surrendering private information advantages.

7. REFERENCES

"Equation Group: The Crown Creator of Cyber-Espionage," Kaspersky Labs, (February 16, 2015).

Azar, Edward E. "The conflict and peace data bank (COPDAB) project." *Journal of Conflict Resolution* 24, no. 1 (1980): 143-152.

Beieler, John. “Generating politically-relevant event data.” *arXiv preprint arXiv:1609.06239* (2016).

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O’Brien, Steve Shellman, James Starz, and Michael Ward, “ICEWS Coded Event Data”, doi:10.7910/DVN/28075, Harvard Dataverse, (2015), V22.

Brodsky, A. E. “Negative case analysis.” *The SAGE encyclopedia of qualitative research methods* (2008): 553.

Gartzke, Erik. “The myth of cyberwar: bringing war in cyberspace back down to earth.” *International Security* 38, no. 2 (2013): 41-73.

Gartzke, Erik, and Jon R. Lindsay. “Weaving tangled webs: offense, defense, and deception in cyberspace.” *Security Studies* 24, no. 2 (2015): 316-348.

Gerner, Deborah J., Philip A. Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. “The creation of CAMEO (Conflict and Mediation Event Observations): An event data framework for a post cold war world.” In *annual meeting of the American Political Science Association*, vol. 29. 2002.

Gompert, David C., and Martin Libicki. “Cyber warfare and Sino-American crisis instability.” *Survival* 56, no. 4 (2014): 7-22.

Hopkins, Daniel, and Gary King. “Extracting systematic social science meaning from text.” *Manuscript available at https://gking.harvard.edu/files/words.pdf* 20, no. 07 (2007).

Hopkins, Daniel J., and Gary King. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54, no. 1 (2010): 229-247.

Healey, Jason, and Karl Grindal, eds. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*. Washington, DC: Cyber Conflict Studies Association, 2013.

Jackson, Patrick Thaddeus. “Foregrounding ontology: dualism, monism, and IR theory.” *Review of International Studies* 34, no. 1 (2008): 129-153.

Kello, Lucas. *The Virtual Weapon and International Order*. Yale University Press, 2017.

King, Gary, and Will Lowe, “An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design” *International Organization*, Vol. 57, No. 03, (2002): 617-642.

Kostyuk, Nadiya, and Yuri M. Zhukov. “Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events?” *Journal of Conflict Resolution* (2017).

Lakatos, Imre, “Falsification and the Methodology of Scientific Research Programmes,” in Imre Lakatos and Alan Musgrave, eds., Criticism and the Growth of Knowledge (New York: Cambridge University Press, 1970), pp. 91-197.

Leetaru, Kalev, and Philip A. Schrodt. “GDELT: Global data on events, location, and tone.” In *ISA Annual Convention*. 2013.

Lewis, James Andrew. *Assessing the Risks of Cyber Terrorism, Cyber War and Other Cyber Threats*. Washington, DC: Center for Strategic & International Studies, 2002.

Libicki, Martin C. *Crisis and Escalation in Cyberspace*. Rand Corporation, 2012.

Lindsay, Jon R. “Stuxnet and the limits of cyber warfare.” *Security Studies* 22, no. 3 (2013): 365-404.

Lindsay, Jon R. “Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack.” *Journal of Cybersecurity* 1, no. 1 (2015): 53-67.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit" In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2014), pp. 55-60.

Nye Jr, Joseph S. "Deterrence and Dissuasion in Cyberspace." *International Security* 41, no. 3 (2017): 44-71.

Radford, Benjamin James. "Automated Learning of Event Coding Dictionaries for Novel Domains with an Application to Cyberspace." PhD diss., Duke University, 2016.

Rid, Thomas, and Ben Buchanan. "Attributing cyber attacks." *Journal of Strategic Studies* 38, no. 1-2 (2015): 4-37.

Schrodt, Philip A., John Beieler, and Muhammed Idris. "Three's a Charm?: Open Event Data Coding with EL: DIABLO, PETRARCH, and the Open Event Data Alliance." In *ISA Annual Convention*. 2014.

Slayton, Rebecca. "What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment." *International Security* 41, no. 3 (2017): 72-109.

Stoll, Clifford. "Stalking the wily hacker." *Communications of the ACM* 31, no. 5 (1988): 484-497.

Valeriano, Brandon, and Ryan C. Maness. "The dynamics of cyber conflict between rival antagonists, 2001–11." *Journal of Peace Research* 51, no. 3 (2014): 347-360.

Valeriano, Brandon, and Ryan C. Maness. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. Oxford University Press, USA, 2015.

Valeriano, Brandon, Benjamin Jensen and Ryan C. Maness. *Cyber Coercion: The Evolving Character of Cyber Power and Strategy*. Oxford University Press, USA, 2018.

Vasquez, John A. "The realist paradigm and degenerative versus progressive research programs: An appraisal of neotraditional research on Waltz's balancing proposition." *American Political Science Review* 91.4 (1997): 899-912.

Ward, Michael D., Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. "Comparing GDELT and ICEWS event data." *Analysis* 21 (2013): 267-297.

Whyte, Christopher. "Ending cyber coercion: Computer network attack, exploitation and the case of North Korea." *Comparative Strategy* (2016).

Whyte, Christopher, "Out of the Shadows: Subversion and Counterculture in the Digital Age," (PhD diss., George Mason University, 2017).

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

The Cyber Deterrence Problem

Aaron F. Brantly
Assistant Professor, Department of Political Science
Virginia Polytechnic and State University
United States
abrantly@vt.edu

Abstract: What is the role of deterrence in an age where adept hackers can credibly hold strategic assets at risk? Do conventional frameworks of deterrence maintain their applicability and meaning against state actors in cyberspace? Is it possible to demonstrate credibility with either in-domain or cross-domain signaling or is cyberspace fundamentally ill-suited to the application of deterrence frameworks? Building on concepts from both rational deterrence theory and cognitive theories of deterrence this work attempts to leverage relevant examples from both within and beyond cyberspace to examine applicability of deterrence in the digital age and for digital tools in an effort to shift the conversation from Atoms to Bits and Bytes.

Keywords: *cyber, deterrence, denial, punishment*

1. INTRODUCTION

The challenge of the digital era is not to define deterrence. Deterrence is a well-defined concept that has been studied and practiced throughout history and to an even greater depth following the advent of nuclear weapons. The present challenge it is to understand the role digital technologies play in the broader scope of interstate deterrence. Deterrence in one domain rarely if ever operates independently of other domains. Much of the literature on cyber deterrence focuses on within domain deterrence. Yet, this is a dangerous constraint that elevates risks and minimizes the probability of success. This paper seeks to draw out the literature on deterrence and identify its applicability within a newly delineated domain of interactions, cyberspace. The resultant analysis strives to encompass the complexity of deterrence and advance an argument beyond within domain modeling.

Classical deterrence centers on a potential adversary's cost-benefit calculus to dissuade specific actions and differs from compellence by focusing on ex-ante behavior manipulation through a priori uses of force or other tools of state power. Both compellence and deterrence are forms of coercion, however, the former employs both hard and soft power both in the present and future with continued or escalated actions, while the latter threatens use of force (power) absent their employment. The focus below is on ex-ante actions by states and sub-state entities that threaten, but that do not use the tools of state against an adversary to manipulate their decision-making calculus. Additionally, actions undertaken independent of threats that can, ex-ante, reduce the benefits associated with a given attack are examined.

Focusing on classical deterrence and deterrence by denial helps illustrate the similarities and differences between deterrence in the pre- and post-delineation of cyberspace as a domain of military operations. Deterrence in cyberspace has been addressed by a variety of scholars across the subfields of International Relations.[1] Many examinations of cyber deterrence rely on direct applications of IR theory absent robust technical understandings of how the domain functions. The development and application of classical deterrence theories to a domain necessarily requires an understanding of how state and non-state actors achieve, develop, and assess costs and benefits within this domain.

This work proceeds in three sections. First, it examines some of the relevant literature on deterrence and identifies some of the gaps within the field and provides a trajectory for the subsequent sections to examine a more dynamic theory of deterrence in cyberspace. The second section focuses on the technical, tactical, operational, and strategic aspects of the domain in an effort to identify those areas where deterrence can alter the costs-benefit analysis of adversaries. Third, the work concludes by providing a discussion on national strategy development for integrated cyber deterrence incorporating the lessons from the first two sections.

2. FROM ATOMS TO BITS AND BYTES

Deterrence is not a novel concept. The classical IR cannon on deterrence can be traced back to the Peloponnesian War and the threat of violence in response to adversary actions.[2] Yet, more modern formulations of deterrence are largely rooted in the nuclear world following World War 2. The most common form of deterrence known as conventional deterrence was established by Bernard Brodie, Thomas Schelling and

1 Mandel, Robert. 2017. *Optimizing Cyberdeterrence: A Comprehensive Strategy for Preventing Foreign Cyberattacks*. Georgetown University Press; Jasper, Scott. 2017. *Strategic Cyber Deterrence the Active Cyber Defense Option*. Lanham, MD: Rowman & Littlefield.

2 Thucydides and Rex Warner. 1968. "The Sixth Book, Chapter XVIII". In *History of the Peloponnesian War*. Baltimore, MD: Penguin Books.

others and focuses on the ex-ante dissuasion of adversaries through the threat of ex-post costs in response to potential adversary actions.

Robert Jervis identified three "waves" of deterrence theorizing to which a potential fourth wave has been added by Jeffery Knopf.[3] First wave deterrence theory rested on the rise and consequences of nuclear weapons. Bernard Brodie et al. asserted that the use of nuclear weapons had almost no innate strategic or tactical value outside of being a threat against an adversary.[4] The consequences of nuclear weapons use, even in limited strike situations, would quickly and dramatically escalate. This escalation made the limited use of such weapons untenable in all but the most extreme situations. Lawrence Freedman summarized the second wave as the realization that "total war could now only be threatened, but never fought".[5]

Second wave deterrence posited how nuclear weapons could be threatened and the dynamics of those threats.[6] Thomas Schelling and others posited a series of conditions in which states could develop deterrence in the nuclear era. As Jervis noted, second wave theorizing became extremely popular because of its abstraction and logical structuring.[7] Game theory and other rational models were used to illustrate rational costs and benefits, creating models suited to rigorous concepts of rationality.[8] The second wave arose under stable bi-polar conditions in which it was assumed states engaged in rational decision-making in matters of foreign policy and national security. Schelling found deterrence largely dependent upon credibility and rationality. He illustrated that signaling potential costs to an adversary absent credibility creates deterrence failure. By using divergent game-theoretic structures from prisoner's dilemma to chicken – theorists developed arguments about deterrence. Despite rigorous theory, this abstraction contained systemic flaws and gave rise to a third wave of deterrence.

The third wave of deterrence theory in the 1970s addressed challenges beyond game theoretic models, including the failing rationality. Irving Janis and Graham Allison, both, but with different perspectives, illustrated the weaknesses of rationality in decision-making.[9] The third wave led to extensions into cognitive psychology and behavioral studies. Robert Jervis, Richard Ned Lebow, and Janis Stein provided insight into the general problems associated with parsimonious use of rationality through case analyses. Specifically, Jervis et al. identified the potential for over-valuation of

3 Jervis, Robert. 1979. "Review: Deterrence Theory Revisited". *World Politics* 31(2): 289–324; Knopf, Jeffrey W. 2010. "The Fourth Wave in Deterrence Research". *Contemporary Security Policy* 31(1): 1–33.

4 Brodie, Bernard, Frederick Sherwood Dunn, Arnold Wolfers, Percy Ellwood Corbett, and William T. R. Fox. 1946. *The Absolute Weapon: Atomic Power and World Order.* New York: Harcourt, Brace and Co.

5 Freedman, Lawrence. 2004. *Deterrence*. Cambridge: Polity Press: 21.

6 Ibid: 22.

7 Jervis. Review: 291-292.

8 Schelling, Thomas C. 1966. *Arms and Influence*. New Haven: Yale University Press: 36-40.

9 Janis, Irving L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Boston: Houghton Mifflin; Allison, Graham T. 1971. Essence of Decision; Explaining the Cuban Missile Crisis. Boston: Little, Brown.

certain attributes of classic deterrence that might inadvertently make conflict more and not less likely.[10]

Jeffrey Berejikian incorporated Daniel Kahneman and Amos Tversky's analysis of prospect theory into the deterrence calculus and challenged parsimonious rational thought by illustrating cognitive dimensions associated with decision-making beyond groupthink and bureaucratic processes. His work highlighted issues related to risk in cognitive decision-making that undermine rationality. Concepts such as sunk costs or tying hands fit well within parsimonious deterrence theory, yet the mechanisms that made them effective were not well understood prior to the third wave.

Although modern deterrence theory encompasses a spectrum from pure rational modeling to cognitive models, the objective of deterrence as identified by John Mearsheimer remains the development of fear of the consequences (in particular of "military action") or a "function of costs and risks".[11] Developing shared knowledge about costs and risks for nuclear events differs from non-nuclear conflicts. Early deterrence models relied heavily on rationality and parsimony but did not underestimate the clarity provided by the use and subsequent impact of the weapons themselves. The generation of fear or knowledge of consequences to assess costs and risks loses clarity the as analyses shift away from nuclear weapons. Lawrence Freedman defines single weapon or type of warfare deterrence as "narrow deterrence".[12] Narrow deterrence is less effective when expanded beyond single weapon or type warfare.

General or broad deterrence covers a range threatened actions to dissuade an adversary. Freedman writes: "broad deterrence involves deterring all war".[13] Ted Hopf explains: within deterrence there is a need to expand deterrence beyond the scope of military tools to the entire range of options available to actors.[14] Extending analysis further, scholars also emphasize concepts of direct deterrence and extended deterrence. Direct deterrence is concerned with actions against "your" state and its immediate interests as opposed to extended deterrence – dissuasion of adversary actions against a third party or non-immediate interests. Delineating between these two types of deterrence in a globalized world is difficult. Cyberspace compounds the challenge of delineation because attacks on foreign infrastructure can and do have ramifications globally.

Concepts of the means to achieve deterrence or more simply how to deter are often contested. Threats can be narrowed to weapon type or category, or include

10 Berejikian, Jeffrey D. 2004. *"International Relations Under Risk: Framing State Choice"*. Albany: State University of New York Press; Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk". *Econometrica* 4(2); Jervis, Robert, Richard Ned Lebow, and Janice Gross Stein. 1985. *Psychology and Deterrence*. Baltimore, MD: Johns Hopkins University Press.

11 Mearsheimer, John J. 1990. *Conventional Deterrence*. Ithaca: Cornell University Press: p 23.

12 Freedman. 2004.

13 Ibid.

14 Hopf, Ted. 1994. *Peripheral Visions: Deterrence Theory and American Foreign Policy in the Third World*, 1965-1990. Ann Arbor: University of Michigan Press.

interdependent relationships such as diplomatic, informational, military and economic effects. Threats signaling a potential response to adversary action should provide clear, unambiguous consequences. The ex-ante threat should causally lead to an ex-post consequence; punishment.

Often left out of traditional international relations literature, deterrence by denial has seen a surge of interest in the years following the 9/11 terrorist attacks. Alex Wilner defines deterrence by denial as "reducing the perceived benefits an action is expected to provide a challenger".[15] Deterrence by denial in the physical world often includes hardening targets by building higher walls, adding security mechanisms, or other tactics to reduce the susceptibility of targets to attack. If the Strategic Defense Initiative (SDI – also known as Star Wars) had been successful, it would have been a deterrence by denial strategy to limit the effect of Soviet nuclear weapons. Commonly used forms of deterrence by denial in conflict zones include land mines, razor wire, surface to air missiles (SAMs) and fortifications.

Deterrence by punishment and denial are intended to manipulate the cost-benefit analysis of an adversary. To function they must both be credible. Credibility requires undertaking ex-ante costs by the deterrer. Threats absent ante impetum costs lack credibility. A state without nuclear weapons cannot credibly threaten nuclear retaliation. If a state wishes to deter it must provide demonstrable evidence that it is able to carry out its threat.

Likewise, deterrence by denial fails when it lacks the material capabilities to deny. The Maginot Line built by the French following World War I stands an example of failed deterrence by denial. The French system of fortifications on portions of their northern territory failed because the line itself only covered one vector of attack into France. The elevation of costs to a potential attacker must be complete and provide no reasonable alternatives to achieve the attacker's intended utility. Both strategies require ex-ante costs by the defender to alter the ex-post perceived benefits of an attacker. Punishment strategies increase adversary costs after a violation and denial strategies increases adversary costs in advance of a violation.

Deterrence by denial is a successful strategy in many instances; SAMs effectively deter enemy aircraft. The relative costs of upgrading certain denial tools is comparatively less than the costs of surmounting them. In the case of SAMs, the United States spent billions of dollars to defeat the S-300 missile system (~$100 million/system).[16] Following the development and use of stealth, S-300 designer Almaz upgraded its

15 Wilner, Alex S. 2015. "Deterrence Theory: Exploring Core Concepts". In *Deterring Rational Fanatics*. Philadelphia: University of Pennsylvania Press: 16-36.

16 Grazier, Dan. 2015. "The Price of the New B-21 Stealth Bomber? Sorry, That's a Secret". *The National Interest*. June 15, 2015. http://nationalinterest.org/blog/the-buzz/the-price-the-new-b-21-stealth-bomber-sorry-thats-secret-16604; 2015. "Program Dossier S-300 Surface-to-Air Missile System". Aviationweek.com. August 6, 2015. http://aviationweek.com/site-files/aviationweek.com/files/uploads/2015/07/asd_08_06_2015_dossier.pdf.

systems to the S-400 variant with greater accuracy and anti-stealth technology.[17] The cost ratio between the denial tool and offensive weapon system is approximately 1 to 1,000. The defensive and offensive capabilities, industrial, and financial resources of these two states exceed most other nations. Even with a $18.5 trillion GDP a $1 to $1,000 cost to benefit ratio is high and demonstrates how denial can be a remarkably effective strategy.

Deterrence by denial is not always successful as illustrated by the Israel – Hamas conflict. In response to Hamas' use of Katyusha rockets, Israel developed the Iron Dome System. Iron Dome batteries cost $100 million and each rocket costs $50,000.[18] To intercept an incoming Katyusha rocket, the Israelis launch 2 interceptor rockets.[19] By contrast, Hamas spends between $500 and $1,000 per rocket launch.[20] If the cost of the battery is ignored, the cost of deterrence by denial is still between 100 to 1 and 200 to 1.

Denial strategies are not passive. They require continuous modification relative to adversary capability development. Static denial strategies in cyberspace or in conventional conflict are likely to have limited credibility over time. Similarly, punishment strategies also require constant updating in relation to adversary capabilities and geopolitical considerations. In cyberspace, this involves adapting denial strategies to technological advances such as artificial intelligence, polymorphic malware and the Internet of Things, to name just a few.

Punishment strategies also require ex-ante costs. Below the nuclear threshold, threats of force are common, yet the credibility of these threats is difficult to establish. Alexander George and Richard Smoke identify three attributes important for signaling in conventional deterrence: "(1) the full formulation of one's intent to protect a nation; (2) the acquisition and deployment of capacities to back up that intent; (3) the communication of intent to a potential aggressor".[21] These three aspects are also at times limited in their ability to convey commitment to fulfill the intent.[22]

Charles Glaser, writing on cyber deterrence, established four components of basic deterrence:

17 Rogoway, Tyler. 2015. "Here's Russia's S-400 Missile System in Action, and How the US Would Deal with It". Foxtrotalpha.Jalopnik.com. December 6, 2015. https://foxtrotalpha.jalopnik.com/heres-russias-s-400-missile-system-in-action-and-heres-1746490022.

18 Morris, Benny. 2014. "Should Israel and the US Rethink Iron Dome's Usefulness?" *LA Times*, August 21, 2016. http://www.latimes.com/opinion/op-ed/la-oe-morris-iron-dome-disastrous-for-israel-20140822-story.html.

19 Ibid.

20 Ibid.

21 George, Alexander L, and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press: 64.

22 Ibid: 558.

> "1) the benefits of taking the action—the larger the benefits, the harder the adversary is to deter; 2) the probability of achieving the benefits—the higher the probability, the harder the adversary is to deter; 3) the costs the defender will impose if the adversary takes the action—the higher the costs, the more likely the adversary is to be deterred; and 4) the adversary's assessment of the probability that the defender will inflict these costs—the higher this probability, the more likely the adversary is to be deterred".[23]

George and Smoke and Glaser acknowledge the challenge of establishing not just threats of punishment, but the credibility associated with carrying out that threat.

Creating material capability (i.e. weapon systems capable of carrying out a given threat) and clear signaling might occur and yet the utilization of this capability in response to an adversary's action will lack credibility (fulfillment of commitment) unless it contains what James Fearon refers to as hand-tying within a sunk costs framework.[24] Credibility and hand-tying are most closely associated with extended deterrence, yet when expanding deterrence to cyberspace it also finds relevance. The establishment of credibility through hand-tying establishes a forcing mechanism for decisions, indicating costs have already been incurred or are likely to occur. This subsequently alters the cost-benefit calculus of retaliation. The stationing of US forces in West Berlin serves as an example of hand-tying through prospective costs.[25] An attack on West Berlin would have resulted in sunk costs and provided a strong inducement or "tripwire" to actuate US retaliatory threats. Nearly all forms of kinetic attacks against the direct interests of a nation implicitly include hand-tying. It is unclear how to effectively signal prospective costs within cyberspace to an adversary.

Charles Glaser identifies several problems associated with deterrence by punishment specific to cyberspace that extend beyond basic credibility issues. First, he notes that deterrence often relies on the attribution of an adversary's actions.[26] In cyberspace, this can be difficult and time-consuming.[27] Although the attribution problem is decreasing as more data becomes available, it does not eliminate uncertainty.[28] Second, hands-tying and other forms of credibility enhancing measures are likely lacking in cyberspace. Moreover, the ability to respond within domain simply might not be possible within certain conditions.[29] Third, Glaser identifies potential spillovers

23 Glaser, Charles. 2011. "Deterrence of Cyber-attacks and US National Security". GW-CSPRI-2011-5. Washington, DC: Cyber Security Policy and Research Institute: 2.
24 Fearon, James D. 1997. "Signaling Foreign Policy Interests". *Journal of Conflict Resolution* 41(1): 69–90.
25 Kydd, Andrew H, and Roseanne W McManus. 2017. "Threats and Assurances in Crisis Bargaining". *Journal of Conflict Resolution* 61(2).
26 Glaser. 2011: 3.
27 Ibid.
28 Rid, Thomas and Ben Buchanan. 2015. "Attributing Cyber Attacks". *Journal of Strategic Studies* 38(1-2): 4–37.
29 Ibid.

in which limited within domain options result in cross-domain, kinetic responses.[30] To date there is limited evidence of cross-domain responses and therefore lacks in credibility. Moreover, cross-domain retaliation alters the escalation framework from digital to kinetic or other and poses a challenge for states wishing to establish credibility while controlling potential escalatory behaviors.

Deterrence is more than simply threatening punishment. Deterrence requires substantial target relevant costs and the development of mechanisms to establish that further costs are credibly wagered to provide clarity for an adversary. The goal of this clarity is to establish within an adversary's calculus that their expected gains are less than any potential losses incurred. Reassessments of rational modeling and the increasing importance of cognitive modeling increase the value of tailored deterrence strategies predicated on the uniqueness of conditions and actors. Paul notes that deterrence is complex and is most logically broken down into five ideal types:

> "(1) deterrence among great powers; (2) deterrence among new nuclear states; (3) deterrence and extended deterrence involving great powers and regional powers armed with chemical, biological and nuclear weapons; (4) deterrence between nuclear states and non-state actors (5) deterrence by collective actors".[31]

It follows that tailored deterrence for cyber actors is also one potential avenue of exploration.

The potential for tailored deterrence strategies could be highlighted in numerous significant cyber incident cases. The 1998 cyber attack code-named SOLAR SUNRISE discovered by US Air Force Computer Emergency Response Team (AFCERT) stands as a prime example. The three-week hack affected more than 500 systems across the US Air Force, Navy, NASA, Lawrence Livermore Labs, MIT, Harvard, and UC Berkeley. The attack coincided with increased tensions between the United States and Iraq and resulted in high-level governmental meetings to identify a proper response action.[32] At the time, the attack was believed to be state-sponsored cyber attack focused on degrading US military capabilities. Subsequently, it was discovered that the attack was conducted by two California teenagers with guidance from Israeli hacker Ehud Tenebaum. The incident is relevant to tailored deterrence because it highlights challenges faced in developing a deterrence strategy. The adversaries were domestic, yet foreign inspired and attacked the operational infrastructure of the Department of Defense. No form of deterrence by punishment delineated above could have appropriately accounted this challenge. The only realistic

30 Ibid.

31 Wirtz, James J, Patrick M Morgan, and T V Paul. 2009. *Complex Deterrence: Strategy in the Global Age*. Chicago: University of Chicago Press: 9.

32 Healey, Jason. 2013. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*. Vienna, VA: Cyber Conflict Studies Association.

deterrence frameworks for SOLAR SUNRISE would have been deterrence by denial or punishment in cooperation with allies.

Richard Kugler writes that a strategy or general framework for deterrence in cyberspace must necessarily be tailored to differing threats, situations, and objectives.[33] The threats, situations, and objectives in cyberspace differ from the concerns addressed by first wave theorists. While the potential for physical damage through cyberspace has been demonstrated in tests such as the Aurora generator experiment that resulted in the destruction of a multi-ton diesel generator, or the Stuxnet attack that destroyed segments of a centrifuge cascade in Iran's Natanz nuclear facility, many attacks do not have kinetic parallels.[34] Building on Kugler, Jeffrey Cooper identifies three important factors that frame concepts on deterrence in cyberspace. First, there is a wide range of actors each with different capabilities and attributes as well as cost benefits structures; second, cyberspace is a unique operational domain that carries with vastly different concepts of risk and reward; third, to develop deterrence, models must be applicable to the virtual and physical aspects of the domain.[35]

This section has provided a summary of a large and robust literature on deterrence. The concepts that need to be carried forward include, the type of deterrence, the credibility of that deterrence and the attributes of the environment in which deterrence occurs, and who and what actors and weapons are to be deterred. The next section builds on the literature above, with a specific emphasis on the technical, tactical, operational and strategic attributes of cyberspace.

3. ONE SIZE DOESN'T FIT ALL

To deter adversaries in cyberspace it is helpful to first define what cyberspace is and what types of actions and actors a state would like to deter. The US Department of Defense defines cyberspace in the following way:

> "Cyberspace consists of many different and often overlapping networks, as well as the nodes (any device or logical location with an Internet protocol address or other analogous identifier) on those networks, and the system data (such as routing tables) that support them. Cyberspace can be described in terms of three

33 Kugler, Richard L. 2009. "Deterrence of Cyber-attacks". In *Cyberpower and National Security*. Edited by Larry K Wentz, Franklin D Kramer, and Stuart H Starr. Washington DC: National Defense University Press: 309–42.

34 US Department of Homeland Security. 2014. "FOIA Documents: Control Systems Security Aurora Update Brief". Washington, DC. http://s3.documentcloud.org/documents/1212530/14f00304-documents.pdf; Zetter, Kim. 2014. *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. New York: Crown Publishers.

35 Cooper, Jeffrey R. 2012. "A New Framework for Cyber Deterrence". In *Cyberspace and National Security Threats, Opportunities, and Power in a Virtual World*. Edited by Reveron, Derek S. 2012. Washington: Georgetown University Press: 105–20.

> layers: physical network, logical network, and cyber-persona. The **physical network** layer of cyberspace is comprised of the geographic component and the physical network components. It is the medium where the data travel. The **logical network** layer consists of those elements of the network that are related to one another in a way that is abstracted from the physical network, i.e., the form or relationships are not tied to an individual, specific path, or node. A simple example is any Web site that is hosted on servers in multiple physical locations where all content can be accessed through a single uniform resource locator. The **cyber-persona** layer represents yet a higher level of abstraction of the logical network in cyberspace; it uses the rules that apply in the logical network layer to develop a digital representation of an individual or entity identity in cyberspace. The cyber-persona layer consists of the people actually on the network".[36]

The inclusion of the full definition illustrates the complexity within which defense strategists and operators in the various services engage. Because the domain spans the physical, logical, and persona layers, deterrence strategies can reasonably occur within and across all three. This fundamentally differs from the conceptualization of deterrence in physical domains of land, sea, air, and space. Physical domain deterrence might include physical and cognitive aspects analogous to the cyber persona and physical network layers, however, the logical layer is wholly absent. The cyber persona layer also diverges significantly from personas within the physical domain as individuals and states have the capacity to alter their attributes within the persona, logical, and network layers.

To construct a meaningful model of deterrence in cyberspace we must first ask what it is we wish to deter. Herein lies the largest distinction between deterrence in the physical world and in cyberspace. Whereas in the physical world deterrence is directed most commonly against physical attacks against specific assets or categories of assets that when attacked provide strong, largely non-repudiable forms of attribution, in cyberspace deterrence is directed against manipulations of the elements within the environment and the environment itself. Manipulation of elements of cyberspace and the environment itself can be examined in multiple ways. Simplifying cyberspace operations into three broad categories, there are cyber attacks, cyber espionage, and cyber theft. Despite simplification, it is important to note these categories are not entirely discrete in process or function. Cyber attacks are those acts in cyberspace that degrade, deny or destroy. Acts of cyber espionage steal information for state or corporate intelligence gain. Cyber theft is the stealing of information for financial gain with no direct state utility. Attacks, espionage, and theft occur across all levels

36 US Department of Defense. 2013. "Joint Publication 3-12: Cyberspace Operations". Washington, DC. http://www.dtic.mil/doctrine/new_pubs/jp3_12R.pdf.

of actors from script kiddies to the military units of states – a problem which will be examined more below. States are most commonly concerned with cyber attacks and espionage at the national level, and theft at lower-jurisdictions.

Because attacks, espionage, and theft are perpetrated by a variety of actors against almost any target in cyberspace, sending an overt signal from one state to another, while still applicable, might not deter attacks at other levels that are of equal or greater significance. Moreover, research by Shawn Lonergan and Erica Borghard indicate a high prevalence of proxy[37] usage by states to maintain plausible deniability.[38] Using proxies to engage in cyber acts against targets deflects deterrence by threats of punishment unless sufficient evidence is present to indicate involvement by the instigating state rather than the third-party proxy. The use of proxies to engage in attacks, espionage and theft against target states outside of cyberspace has been the practice of states since Katulaya and Sun Tzu.[39] However, unlike the difficulties of non-repudiability within conventional conflicts, cyber attacks are frequently repudiable. Attackers might use Virtual Private Networks (VPNs), proxies or other means by which to engage in an attack.

Additional problems in cyberspace not frequently encountered in conventional physical domains are second and third order effects. As noted by Herbert Lin, the results of a cyber attack itself might not be identifiable, rather it is second or third order effects that generate an intended outcome.[40] Classical deterrence and tailored deterrence strategies used against terrorist organizations are unable to account for disconnected action and reaction pairs commonly found in cyberspace. The time to punish a violation can be weeks, months or years based on discovery and attribution challenges, a problem not present in classical deterrence.

Cyber attacks are incidents occurring in or through cyberspace that degrade, deny or destroy. Attacks in cyberspace can and are perpetrated by all levels of actors. The differentiation between actors is most closely correlated with targets and outcomes of attacks.[41] For example, criminal actors may use phishing attacks to ingress into a hospital's computer systems to install Cryptolocker or a similar ransomware malware on the hospital's systems. Cryptolocker is an attack that degrades civilian critical infrastructure, denies user access and has the potential to destroy critical

37 Here proxy usage refers to the authority to represent someone else not the technical usage of the term in information communications.

38 Borghard, Erica D, and Shawn W Lonergan. 2016. "Can States Calculate the Risks of Using Cyber Proxies?" *Orbis* 60(3): 395–416.

39 Kautalya and L. N. Rangarajan. 1992. *The Arthashastra*. New Delhi: Penguin Books India; Griffith, Samuel B, and Sun Tzu. 1971. *The Art of War*. New York: Oxford University Press.

40 Lin, Herbert. "Operational Considerations in Cyber-attack and Cyber Exploitation". In *Cyberspace and National Security Threats, Opportunities, and Power in a Virtual World*. Edited by Reveron, Derek S. 2012. Washington: Georgetown University Press.

41 Brantly, Aaron F. 2015. "Aesop's Wolves: The Deceptive Appearance of Espionage and Attacks in Cyberspace". *Intelligence and National Security* 31(5): 674-685.

information.[42] Very few states have national deterrence strategies aimed at sub-state actors, criminal organizations or individuals. State deterrence strategies aimed at non-terrorist sub-state actors are confined to criminological models of deterrence. Yet, if a soldier or spy from an adversary state walked into the server room at the same hospital and threatened to detonate a bomb and destroy all the files unless he was paid a ransom, the act would align more closely with a conventional deterrence framework of state-to-state deterrence by threats of punishment or tailored deterrence against terrorist actors.

Most scholars and practitioners are likely to contend that it is not the responsibility of the state to deter non-state actors (excepting terrorists), particularly criminals from cyber attacks against non-federal infrastructure outside of a criminological framework.[43] Yet, the same tool used by a criminal is available to the state and presents the same challenges associated with attribution irrespective of the perpetrator. What actions could a state undertake to deter an adversary state actor from engaging in this behavior and would these actions have a measurable effect on non-state actors as well?

Examples of cyber attacks abound and include the destruction, denial or degradation of military or civilian communications platforms. Attacks such as the Mirai (malware) botnet attack in 2016 are capable of being directed at both critical and non-critical infrastructure by both state and non-state actors. A botnet using Mirai was able to generate in excess of 1Tbps of traffic and degrade dozens of websites in the United States on 20 September 2016.[44] This same form of attack could be directed towards IP addresses of the FAA and emergency service providers or any number of Internet-enabled systems found on Shodan.io or similar services.[45]

Although DDoS attacks are generally considered to be among the least complicated forms of cyber attacks they still challenge state and sub-state entities both public and private. DDoS attacks have been used against US government infrastructure, against Estonia in 2007 and the Republic of Georgia in 2008.[46] To date, DDoS attacks against the US government or critical infrastructure have received little attention in discussions on deterrence in cyberspace. On 21 January 2016 a grand jury in the Southern District of New York indicted 7 Iranian Hackers in absentia for their involvement in DDoS

42 Winton, Richard. 2016. "Hollywood Hospital Pays $17,000 in Bitcoin to Hackers; FBI Investigating". *Los Angeles Times*. February 18, 2016. http://www.latimes.com/business/technology/la-me-ln-hollywood-hospital-bitcoin-20160217-story.html.

43 Akers, Ronald L. 2017. "Rational Choice, Deterrence, and Social Learning Theory in Criminology: The Path Not Taken" *Journal of Criminal Law and Criminology* 81(3): 1–25.

44 Bonderud, Douglas. 2016. "Leaked Mirai Malware Boosts IoT Insecurity Threat Level". securityintelligence.com. October 4, 2016. https://securityintelligence.com/news/leaked-mirai-malware-boosts-iot-insecurity-threat-level/.

45 Bodenheim, Roland, Jonathan Butts, Stephen Dunlap, and Barry Mullins. 2014. "Evaluation of the Ability of the Shodan Search Engine to Identify Internet-Facing Industrial Control Devices". *International Journal of Critical Infrastructure Protection* 7(2): 114–23.

46 Klimburg, Alexander. 2011. "Mobilizing Cyber Power". *Survival* 53(1): 41–60; Hollis, David. 2011. "Cyberwar Case Study: Georgia 2008". *Small Wars Journal*, http://smallwarsjournal.com/jrnl/art/cyberwar-case-study-georgia-2008.

attacks against US financial sector interests and a variety of other US companies occurring from 2011-2013.[47] These indictments are: (a) not deterrent threats or denials, but criminological deterrents; (b) temporally distant from the time of attack as to be ineffective at signaling deterrence; and (c) impose little to no costs on Iran or the individual perpetrators or organizers of the attack.

Beyond DDoS attacks, Russian attacks against Ukrainian electric infrastructure and US political organizations also resulted in no or weak responses that offer no indication that deterrence is making headway in cyberspace.[48] In response to massive influence operations perpetrated by the Russian Federation against the United States and its two major political parties during the 2016 Presidential election the United States expelled 35 suspected Russian intelligence operatives and placed sanctions on Russia's two leading intelligence services, the FSB and the GRU.[49] The US response imposed insignificant costs in comparison to the utility achieved by the Russian Federation.

The latter case of Russian influence and hacking during the 2016 election cycle provides a case study for why deterrence by threat in cyberspace is so difficult to achieve. The first indications of Russian interference in the 2016 election were identified by the FBI in September 2015 more than a year before the election.[50] The FBI phoned the DNC to try and alert them to a potential attack, but the call was not considered credible and was subsequently ignored by DNC staffers.[51] The progression of hacking attempts against the DNC continued and President Obama was notified in the summer of 2016. Moreover, the "attack" against the DNC was not an attack, but espionage or theft and therefore falls outside conventionally defined deterrence frameworks. Yet the impact of the espionage and the later release of private DNC emails was substantial as indicated in a declassified report by the Office of the Director of National Intelligence (ODNI).[52] The report assessed that information warfare conducted following the espionage campaign substantially degraded the DNC and engendered a loss of confidence in the US electoral system.[53] Cyber deterrence has fundamental problems including the realization that the most valuable assets in cyberspace might not be destroyed or degraded, but rather stolen and used.

47 US Federal Bureau of Investigation. 2016. "Iranian DDoS Attacks: Conspiracy to Commit Computer Intrusion". https://www.fbi.gov/wanted/cyber/iranian-ddos-attacks.

48 US Department of Homeland Security. 2016. "Cyber-Attack Against Ukrainian Critical Infrastructure | ICS-CERT". Washington, DC; Rid, Thomas. 2016. "How Russia Pulled Off the Biggest Election Hack in US History". *Esquire*. October 20, 2016. http://www.esquire.com/news-politics/a49791/russian-dnc-emails-hacked/.

49 Sanger, David E. 2016. "Obama Strikes Back at Russia for Election Hacking". *The New York Times*. New York. December 29, 2016. https://www.nytimes.com/2016/12/29/us/politics/russia-election-hacking-sanctions.html.

50 Lipton, Eric, David E Sanger, and Scott Shane. 2016. "The Perfect Weapon: How Russian Cyberpower Invaded the US". *The New York Times*. December 13, 2016. https://www.nytimes.com/2016/12/13/us/politics/russia-hack-election-dnc.html.

51 Ibid.

52 US Office of the Director of National Intelligence. 2017. "Assessing Russian Activities and Intentions in Recent US Elections" Washington, D.C. January 6, 2017. https://www.dni.gov/files/documents/ICA_2017_01.pdf.

53 Ibid.

Even in instances where specific code is used to achieve damage such has Iranian efforts to hack a spillway dam[54] or malware implants in critical infrastructure such as a German steel mill,[55] there are no formal mechanisms by which to signal a threat within cyberspace or beyond other than by referencing responses to kinetic effects. Current deterrence by threat signaling for attacks occurring in or through cyberspace is ambiguous. Efforts by the NATO CCD COE through the production of the Tallinn Manuals have begun to outline the frameworks in which deterrence could legally take place, yet the application of threats is still uncertain.[56]

Deterrence by threat within cyberspace is realistically only applicable to cyber operations that result in direct physical effects that are non-repudiable and attributed quickly. Using formal modeling in the *Decision to Attack: Military and Intelligence Cyber Decision-making* I found that most cyber attacks, with the notable exception of DDoS, operate under varying conditions of anonymity.[57] The anonymity associated with attacks is usually necessary for attacks to be successful in bypassing deterrence by denial frameworks found in the perimeter defenses of networks such as intrusion detection and prevention systems found in the logical or physical network layers of cyberspace. Threats of punishment could impact the persona layer of cyberspace as well, but as will be examined below there are some fundamental challenges unique to cyberspace posed by anonymity.

4. TECHNICAL CHALLENGES: THREATS OF PUNISHMENT WITHIN DOMAIN

Punishing an adversary in cyberspace is not cheap or fast outside of pre-established botnets or damage done to physical infrastructure. Punishment in or across any of the layers cyberspace requires what the US Department of the Army refers to as intelligence preparation of the battlefield (IPB):

> "IPB is a systemic, continuous process of analyzing the threat and environment in a specific geographic area. It is designed to support staff estimates and military decision making".[58]

54 Cylance. 2014. "Operation Cleaver". https://www.cylance.com/content/dam/cylance/pages/operation-cleaver/Cylance_Operation_Cleaver_Report.pdf.

55 Lee, Robert M, Michael J Assante, and Tim Conway. 2014. "German Steel Mill Cyber-attack". SANS Industrial Control Systems. December 30, 2014. https://ics.sans.org/media/ICS-CPPE-case-Study-2-German-Steelworks_Facility.pdf.

56 Schmitt, Michael N. (Ed.). 2013. *Tallinn Manual on the International Law Applicable to Cyber Warfare:* Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence. New York: Cambridge University Press.

57 Brantly, Aaron Franklin. 2016. *The Decision to Attack: Military and Intelligence Cyber Decision-Making*. Athens: University of Georgia Press.

58 US Department of the Army. 1994. FM 34-130 Intelligence Preparation of the Battlefield. Washington, DC.

In response to a nuclear attack on a city in the US, the proportional response would be a counter attack on an adversary city. The city itself is geographically fixed and immovable both logically and physically. Threatening in-kind retaliation is both plausible and technically feasible with ballistic missiles or air assets. The same logic does not hold in cyberspace.

Why are in kind retaliations or other forms of punishment not viable solutions for most retaliations in cyberspace? First, a state must fulfill the burden of proof in identifying the perpetrator of an action. All the above IPB and potential for retaliation still depends upon attribution of who, what, and potentially why an attack occurred.[59] Retaliation absent strong evidence is likely to lead to misidentification and unnecessary escalation.

Second, a state must retaliate within a proximate temporal range. If state X does not have detailed intelligence on the asset it wishes to retaliate against, developing intelligence along with a cyber weapon to target it increases the time horizon of response such that it is days, weeks, months or even years out from the original attack for which it is retaliating. Due to this temporal disconnect, the threat to punish in response to a given action falls into a category of what economists refer to as hyperbolic discounting. The risk of punishment for an attack is possible but so temporally, distant as to be discounted to the point of irrelevance.

Third, deterrence by punishment requires proportionality. It is necessary to have comparable assets to punish to prevent escalation or violations of international law.[60] Comparable assets are not a given within cyberspace and are often difficult to identify.[61] To punish an asset within a domain requires pre-established access or knowledge of that asset beyond its location. Whereas a city is immovable and likely to be as susceptible today as it will be tomorrow to a missile or bomb, a computer system that is penetrated today for prepositioned access, might be patched, upgraded or taken offline tomorrow.

Fourth, a state must possess a specific cyber weapon system tailored to its target. If state X alerts state Y that it is going to punish an asset or state X uses a repeated cyber weapon to attack state Y's system, it is likely to be ineffectual the longer it is used due to updated perimeter defenses, such as intrusion detection and prevention systems (IDPS), antivirus programs or a variety of other security measures. If state X wants to punish state Y it must have knowledge of the attributes of the asset it wishes to retaliate against and what the status of that asset is. State X must also develop new exploits to achieve effects or be confident that State Y has not accounted for previous exploits that have been used.

59 Brantly. 2016.

60 Schmitt, Michael N. (Ed.) 2017. *Tallinn Manual on the International Law Applicable to Cyber Operations*: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence. New York: Cambridge University Press: Kindle Location: 4530.

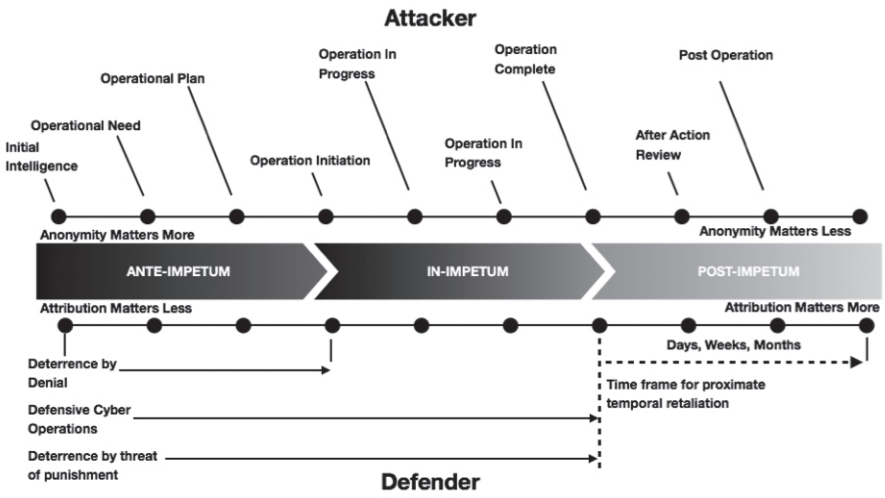
61 Libicki, Martin C. 2016. *Cyberspace in Peace and War*. Naval Institute Press: 262.

The challenges of signaling deterrence by punishment are numerous within cyberspace whether the conflict is contained within domain or crosses over domains. Advances in attribution within a timely manner and the availability and reasonable assumption that proportional assets of an adversary can be held at risk need to be improved to credibly threaten punishment. This is a challenge not isolated to within domain retaliation. While proportional target selection might be slightly easier in cross-domain retaliation, the first three issues raised above are still relevant.

Deterrence by punishment in cyberspace is possible, but it is not a reliable or credible option under most conditions absent sufficient and sustained intelligence. This assessment is not unique and is borne out in the analysis of Valeriano and Maness, who find that deterrence via punishment is generally ineffective and likely more dangerous than other means of preventing attacks.[62] Moreover, sustained invasive intelligence into adversary networks creates its own unique problems, including a security dilemma.[63] The more states engage in highly invasive intelligence via cyberspace, the more their actions are likely to be misinterpreted. Differentiating between various forms of cyber actions are difficult and can lead to miscalculation.[64]

Figure 1 illustrates the relationship between attacking and defending forces and area where both forms of deterrence function.

FIGURE 1. TIMELINE OF CYBER ATTACKS AND DEFENSE



62 Valeriano, Brandon, and Ryan C Maness. 2015. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. New York: Oxford University Press: 57-60.

63 Buchanan, Ben. 2017. *The Cybersecurity Dilemma Hacking, Trust and Fear Between Nations*. Oxford: Oxford University Press.

64 Brantly. 2016.

As seen in Figure 1, deterrence by threat of punishment and denial operate within the same temporal ranges, yet while attribution matters a great deal for threats of punishment they are generally unimportant for denial. In their initial stages both denial and punishment focus on ante-impetum means of dissuasion, yet deterrence by punishment necessarily needs post-impetum attribution for it to be used. Based on the technical realities of cyberspace and of international relations deterrence by threat of punishment is more complicated and difficult to effectively establish.

5. TECHNICAL CHALLENGES AND OPPORTUNITIES: DETERRENCE BY DENIAL

Both deterrence by denial and punishment require ante-impetum costs by the defender. The allocation of resources between denial and deterrence and the efficiency with which they deter adversaries differ. The establishment of credible deterrence by denial often starts with the allocation of financial capital to purchase technical resources and provide human capital sufficient to continually update, enhance, audit and manage complex network infrastructure.[65] Network-based and host-based defenses such as intrusion detection and prevention systems, anti-virus products and similar systems are some of the variety of overlapping expenditures that can be undertaken to increasingly make the intrusion of adversaries into a given network more difficult.[66]

In cyberspace, such expenditures are regularized and often included as overhead costs, however they are deterrent in nature.[67] Although they are not glamorous, they substantially decrease the probability of penetration. The same types of deterrence strategies are used by stores in placing electronic tracking tags on their products and detectors at doors, by banks in the construction of vaults, silent alarms and dyed packets of money, by critical infrastructure in extending the perimeter of security outward to prevent vehicle-borne improvised explosive devices, increased numbers of security guards, cameras and the use of razor wire or other physical structures. These devices signal to adversaries both criminal and terrorist alike that the costs of successfully perpetrating an attack are high and that the likelihood of success is low, although both terrorist and criminal deterrence models include deterrence by punishment through criminal proceedings and potential lethal actions against terrorist they rely far more heavily on preventive measures that deny would be adversaries.

Sceptics might contend denial mechanisms are unlikely to deter a state, yet this is in and of itself not accurate. The vast majority of probes by states do not translate into successfully attacks. The US Department of Defense suffers from millions of probes

65 Riggs, Cliff. (2004). *Network Perimeter Security*. New York: Auerbach Publications.

66 Buecher, Axel, Per Andreas, and Scott Paisley. 2009. "Understanding IT Perimeter Security". IBM. http://www.redbooks.ibm.com/redpapers/pdfs/redp4397.pdf.

67 Filkins, Barbara. 2016. "IT Security Spending Trends". SANS. https://www.sans.org/reading-room/whitepapers/analyst/security-spending-trends-36697.

a day. Yet nearly 99.99% of them are unsuccessful.[68] Moreover, in the face of a global onslaught of cyber attacks and espionage the United States re-architected much of its military network infrastructure. This restructuring allows the initial point of contact with adversaries to be chosen. In military parlance, it allowed the defenders to choose the terrain of the battle. While it did not obviate the need for denial mechanisms within the network infrastructure, it did signal increased cost imposition on adversaries and it did allow for more efficient resource allocation.

Unlike in any other battlespace, whether conventional kinetic terrorism, conventional kinetic or mass destruction military force, the opportunities for deterrence by denial are substantial in cyberspace and unique. While denial opportunities in land, sea, air, and even space are predicated on the control of a given geospatial area, the party establishing deterrence by denial has limited abilities to manipulate the nature of the domain itself. The same is not true within cyberspace. Every aspect of a defender's cyberspace from the structure of the network, to the hardware, firmware, and software within a network, to the access of individuals within and external to that network is manipulable. At every stage of an attack an adversary is always attempting to operate on or against the defender's cyberspace over which it has no control and has limited visibility.

For denial, the historical literature of deterrence theory remains relevant, in particular the second and third stages of deterrence which focused on rational game theoretic and cognitive modeling. While in conventional deterrence the emphasis was on punishment, here these same modeling techniques find applicability in deterrence by denial. Although the games might be the same, the payoffs in cyberspace manipulable and favor the defender. In few other applications of deterrence are the payoff matrices of deterrence so favorable to the defender. Despite the favorability of conditions, the ability to manipulate the potential payoff for attackers remains difficult. Although possible for defenders to reduce the probability of attack success, the potential payoff for a successful attack can remain large.

Despite conditions favoring defenders, the potential payoffs are often not affected by deterrence by denial. Minimizing the potential payoffs from attacks on data repositories requires disaggregation of data. These types of denial mechanisms come with efficiency or financial costs. Although denial offers more potential than punishment, it is not a silver bullet to the cyber deterrence problem. Denial decreases the probability of success for attackers and is likely to reduce classes of actors focused on certain targets. Despite efforts to signal through the purchase and implementation of various defensive measures, the re-architecting of network infrastructure, the cyber deterrence problem remains.

68 Howard, Travis, and Jose de Arimateia de Cruz. 2017. "The Cyber Vulnerabilities of the US Navy". *The Maritime Executive*. January 31, 2017. https://maritime-executive.com/article/the-cyber-vulnerability-of-the-us-navy.

6. BEYOND THE DETERRENCE PROBLEM

If punishment and denial are unable to fully remediate the cyber deterrence problem, are there any meaningful solutions? The core debate remains, with no simple and readily apparent solutions. The search for a single solution is likely to remain fruitless for the foreseeable future. Deterrence has never been the single tool within the toolbox of the state to dissuade or shape adversary behavior. Rather, it has always been combined with efforts that extend beyond traditional concepts of deterrence to include geopolitical and technical practices including norm development, entanglement, cumulative deterrence, research and development, policies and laws, liability structures for software and hardware, training for users and human capital development within information technology and cybersecurity.[69]

Efficient and effective cyber deterrence should extend international politics and include fields such as criminology, immunology and public health.[70] The capacity of states to punish criminals is high and the credibility of punishment actions in developed nations is strong. Despite a capacity to punish criminal behaviors, they still occur. Extending beyond punishment, states also focus on denying criminals opportunities to commit crimes. Yet crime still occurs. The root causes of crime are not simple nor isolatable to a single phenomenon. Likewise, states engage one another in cyberspace for a variety of reasons. Some reasons fit within conventional deterrence frameworks of denial and punishment and do not suffer from challenges with attribution. For instance, larger and more harmful attacks increase the probability of attribution. However, many states remain perturbed by the death by a thousand cuts phenomena which falls below thresholds and required to provide timely attribution.

Shifting the focus away from within domain deterrence focused solely on punishment and denial and changing the emphasis to a basket of strategies focused on reducing incentives, availability and anonymity fosters an environment less conducive both to hostile actions and potential malicious actors. The solution to the deterrence problem is not abandoning it, but expanding the range of alternative strategies not presently considered. By acknowledging the failures and inadequacies of deterrence strategies and the potential places where novel strategies found in other fields are applicable the intractable problem of cyber deterrence becomes more manageable.

69 Nye. 2017: 45-69; Tor, Uri. 2017. "'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence". *Journal of Strategic Studies* 40(1-2): 92–117.

70 Jaishankar, K. 2011. *Cyber Criminology: Exploring Internet Crimes and Criminal Behavior*. Boca Raton, FL: CRC Press; Brantly, Aaron "Epidemiological Approaches to National Cybersecurity". In *US National Cybersecurity: International Politics, Concepts and Organization*. Edited by Damien Van Puyvelde and Aaron Franklin Brantly. 2017. New York: Routledge.

REFERENCES

Akers, Ronald L. 2017. "Rational Choice, Deterrence, and Social Learning Theory in Criminology: The Path Not Taken". *Journal of Criminal Law and Criminology* 81(3).

Allison, Graham T. 1971. *Essence of Decision; Explaining the Cuban Missile Crisis*. Boston: Little, Brown.

Aviation Week. 2015. "Program Dossier S-300 Surface-to-Air Missile System". Aviationweek.com. August 6, 2015. http://aviationweek.com/site-files/aviationweek.com/files/uploads/2015/07/asd_08_06_2015_dossier.pdf.

Berejikian, Jeffrey D. 2004. *International Relations Under Risk: Framing State Choice*. Albany: State University of New York Press.

Bodenheim, Roland, Jonathan Butts, Stephen Dunlap, and Barry Mullins. 2014. "Evaluation of the Ability of the Shodan Search Engine to Identify Internet-Facing Industrial Control Devices". *International Journal of Critical Infrastructure Protection* 7(2).

Bonderud, Douglas. 2016. "Leaked Mirai Malware Boosts IoT Insecurity Threat Level". securityintelligence.com. October 4, 2016. https://securityintelligence.com/news/leaked-mirai-malware-boosts-iot-insecurity-threat-level/.

Borghard, Erica D, and Shawn W Lonergan. 2016. "Can States Calculate the Risks of Using Cyber Proxies?" *Orbis* 60(3).

Brantly, Aaron "Epidemiological Approaches to National Cybersecurity". In *US National Cybersecurity: International Politics, Concepts and Organization*. Edited by Damien Van Puyvelde and Aaron Franklin Brantly. 2017. New York: Routledge.

Brantly, Aaron F. 2015. "Aesop's Wolves: The Deceptive Appearance of Espionage and Attacks in Cyberspace". *Intelligence and National Security* 31(5).

Brantly, Aaron Franklin. 2016. *The Decision to Attack: Military and Intelligence Cyber Decision-Making*. Athens: University of Georgia Press.

Brodie, Bernard, Frederick Sherwood Dunn, Arnold Wolfers, Percy Ellwood Corbett, and William T. R. Fox. 1946. *The Absolute Weapon: Atomic Power and World Order*. New York: Harcourt, Brace and Co.

Buchanan, Ben. 2017. *The Cybersecurity Dilemma Hacking, Trust and Fear Between Nations*. Oxford: Oxford University Press.

Buecher, Axel, Per Andreas, and Scott Paisley. 2009. "Understanding IT Perimeter Security". IBM. http://www.redbooks.ibm.com/redpapers/pdfs/redp4397.pdf.

Cooper, Jeffrey R. 2012. "A New Framework for Cyber Deterrence". In *Cyberspace and National Security Threats, Opportunities, and Power in a Virtual World*. Edited by Reveron, Derek S. 2012. Washington: Georgetown University Press.

Cylance. 2014. "Operation Cleaver". https://www.cylance.com/content/dam/cylance/pages/operation-cleaver/Cylance_Operation_Cleaver_Report.pdf.

Fearon, James D. 1997. "Signaling Foreign Policy Interests". *Journal of Conflict Resolution* 41(1).

Filkins, Barbara. 2016. "IT Security Spending Trends". SANS. https://www.sans.org/reading-room/whitepapers/analyst/security-spending-trends-36697.

Freedman, Lawrence. 2004. *Deterrence*. Cambridge: Polity Press.

George, Alexander L, and Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.

Glaser, Charles. 2011. "Deterrence of Cyber-attacks and US National Security". GW-CSPRI-2011-5. Washington, DC: Cyber Security Policy and Research Institute.

Grazier, Dan. 2015. "The Price of the New B-21 Stealth Bomber? Sorry, That's a Secret". *The National Interest*. June 15, 2015. http://nationalinterest.org/blog/the-buzz/the-price-the-new-b-21-stealth-bomber-sorry-thats-secret-16604.

Griffith, Samuel B, and Sun Tzu. 1971. *The Art of War*. New York: Oxford University Press.

Healey, Jason. 2013. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*. Vienna, VA: Cyber Conflict Studies Association.

Hollis, David. 2011. "Cyberwar Case Study: Georgia 2008". *Small Wars Journal*. http://smallwarsjournal.com/jrnl/art/cyberwar-case-study-georgia-2008.

Hopf, Ted. 1994. *Peripheral Visions: Deterrence Theory and American Foreign Policy in the Third World*, 1965-1990. Ann Arbor: University of Michigan Press.

Howard, Travis, and Jose de Arimateia de Cruz. 2017. "The Cyber Vulnerabilities of the US Navy". *The Maritime Executive*. January 31, 2017. https://maritime-executive.com/article/the-cyber-vulnerability-of-the-us-navy.

Jaishankar, K. 2011. *Cyber Criminology: Exploring Internet Crimes and Criminal Behavior*. Boca Raton, FL: CRC Press.

Janis, Irving L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Boston: Houghton Mifflin.

Jasper, Scott. 2017. *Strategic Cyber Deterrence the Active Cyber Defense Option*. Lanham, MD: Rowman & Littlefield.

Jervis, Robert, Richard Ned Lebow, and Janice Gross Stein. 1985. *Psychology and Deterrence*. Baltimore, MD: Johns Hopkins University Press.

Jervis, Robert. 1979. "Review: Deterrence Theory Revisited". *World Politics* 31(2).

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk". *Econometrica* 4(2).

Kautalya and L. N. Rangarajan. 1992. *The Arthashastra*. New Delhi: Penguin Books India.

Klimburg, Alexander. 2011. "Mobilizing Cyber Power". *Survival* 53(1).

Knopf, Jeffrey W. 2010. "The Fourth Wave in Deterrence Research". *Contemporary Security Policy* 31(1).

Kugler, Richard L. 2009. "Deterrence of Cyber-attacks". In *Cyberpower and National Security*. Edited by Larry K Wentz, Franklin D Kramer, and Stuart H Starr. Washington DC: National Defense University Press.

Kydd, Andrew H, and Roseanne W McManus. 2017. "Threats and Assurances in Crisis Bargaining". *Journal of Conflict Resolution* 61(2).

Lee, Robert M, Michael J Assante, and Tim Conway. 2014. "German Steel Mill Cyber-attack". SANS Industrial Control Systems. December 30, 2014. https://ics.sans.org/media/ICS-CPPE-case-Study-2-German-Steelworks_Facility.pdf.

Libicki, Martin C. 2016. *Cyberspace in Peace and War*. Naval Institute Press.

Lin, Herbert. 2012. “Operational Considerations in Cyber-attack and Cyber Exploitation” in *Cyberspace and National Security Threats, Opportunities, and Power in a Virtual World*. Edited by Reveron, Derek S. Washington: Georgetown University Press.

Lipton, Eric, David E Sanger, and Scott Shane. 2016. “The Perfect Weapon: How Russian Cyberpower Invaded the US”. *The New York Times*. December 13, 2016. https://www.nytimes.com/2016/12/13/us/politics/russia-hack-election-dnc.html.

Mandel, Robert. 2017. *Optimizing Cyberdeterrence: A Comprehensive Strategy for Preventing Foreign Cyberattacks*. Georgetown University Press.

Mearsheimer, John J. 1990. *Conventional Deterrence*. Ithaca: Cornell University Press.

Morris, Benny. 2014. “Should Israel and the US Rethink Iron Dome’s Usefulness?” *LA Times*. August 21, 2016. http://www.latimes.com/opinion/op-ed/la-oe-morris-iron-dome-disastrous-for-israel-20140822-story.html.

Rid, Thomas and Ben Buchanan. 2015. “Attributing Cyber Attacks”. *Journal of Strategic Studies* 38(1-2).

Rid, Thomas. 2016. “How Russia Pulled Off the Biggest Election Hack in US History”. *Esquire*. October 20, 2016. http://www.esquire.com/news-politics/a49791/russian-dnc-emails-hacked/.

Riggs, Cliff. 2004. *Network Perimeter Security*. New York: Auerbach Publications.

Rogoway, Tyler. 2015. “Here’s Russia’s S-400 Missile System in Action, and How the US Would Deal with It”. Foxtrotalpha.Jalopnik.com. December 6, 2015. https://foxtrotalpha.jalopnik.com/heres-russias-s-400-missile-system-in-action-and-heres-1746490022.

Sanger, David E. 2016. “Obama Strikes Back at Russia for Election Hacking”. *The New York Times*. New York. December 29, 2016. https://www.nytimes.com/2016/12/29/us/politics/russia-election-hacking-sanctions.html.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven: Yale University Press.

Schmitt, Michael N. (Ed.). 2013. *Tallinn Manual on the International Law Applicable to Cyber Warfare*: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence. New York: Cambridge University Press.

Schmitt, Michael N. (Ed.). 2017. *Tallinn Manual on the International Law Applicable to Cyber Operations*: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence. New York: Cambridge University Press.

Thucydides, and Rex Warner. 1968. “The Sixth Book, Chapter XVIII”. In *History of the Peloponnesian War*. Baltimore, MD: Penguin Books.

Tor, Uri. 2017. “‘Cumulative Deterrence’ as a New Paradigm for Cyber Deterrence”. *Journal of Strategic Studies* 40(1-2).

US Department of Defense. 2013. “Joint Publication 3-12: Cyberspace Operations”. Washington, DC. http://www.dtic.mil/doctrine/new_pubs/jp3_12R.pdf.

US Department of Homeland Security. 2014. “FOIA Documents: Control Systems Security Aurora Update Brief”. Washington, DC. http://s3.documentcloud.org/documents/1212530/14f00304-documents.pdf.

US Department of Homeland Security. 2016. “Cyber-Attack Against Ukrainian Critical Infrastructure ICS-CERT”. Washington, DC.

US Department of the Army. 1994. *FM 34-130 Intelligence Preparation of the Battlefield*. Washington, DC.

US Federal Bureau of Investigation. 2016. “Iranian DDoS Attacks: Conspiracy to Commit Computer Intrusion”. https://www.fbi.gov/wanted/cyber/iranian-ddos-attacks.

US Office of the Director of National Intelligence. 2017. “Assessing Russian Activities and Intentions in Recent US Elections” Washington, D.C. January 6, 2017. https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Valeriano, Brandon, and Ryan C Maness. 2015. *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*. New York: Oxford University Press.

Wilner, Alex S. 2015. “Deterrence Theory: Exploring Core Concepts”. In *Deterring Rational Fanatics*. Philadelphia: University of Pennsylvania Press.

Winton, Richard. 2016. “Hollywood Hospital Pays $17,000 in Bitcoin to Hackers; FBI Investigating”. *Los Angeles Times*. February 18, 2016. http://www.latimes.com/business/technology/la-me-ln-hollywood-hospital-bitcoin-20160217-story.html.

Wirtz, James J, Patrick M Morgan, and T V Paul. 2009. *Complex Deterrence: Strategy in the Global Age*. Chicago: University of Chicago Press.

Zetter, Kim. 2014. *Countdown to Zero Day: Stuxnet and the Launch of the World’s First Digital Weapon*. New York: Crown Publishers.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Offensive Cyber Capabilities: To What Ends?

Max Smeets
Center for International Security and Cooperation
Stanford University
Stanford, United States
mwsmeets@stanford.edu

Herbert S. Lin
Center for International Security and Cooperation
The Hoover Institution
Stanford University
Stanford, United States
herbert.s.lin@stanford.edu

Abstract: There is a growing interest in the use of offensive cyber capabilities (OCC) among states. Despite the growing interest in these capabilities, little is still known about the nature of OCC as a tool of the state. This research therefore aims to understand if (and how) offensive cyber capabilities have the potential to change the role of military power. Drawing on a wide range of cases, we argue that these capabilities can alter the manner in which states use their military power strategically in at least four ways. OCC are not particularly effective in *deterring* adversary military action, except when threatened to be used by states with a credible reputation. However, they do have value in *compellence*. Unlike conventional capabilities, the effects of offensive cyber operations do not necessarily have to be exposed publicly, which means the compelled party can back down post-action without losing face thus deescalating conflict. The potential to control the reversibility of effect of an OCC by the attacker may also encourage compliance. OCC also contribute to the use of force for *defensive* purposes, as it could provide both a preemptive as well as preventive strike option. Finally, its symbolic value as a 'prestige weapon' to enhance 'swaggering' remains unclear, due to its largely non-material ontology and transitory nature.

Keywords: *offensive cyber capabilities, compellence, defense, deterrence, military power, swaggering*

1. INTRODUCTION

There is a growing interest in the use of offensive cyber capabilities (OCC) among states. A diverse group of states across the world including Belgium, Columbia, Germany, Finland, India, the United Arab Emirates (UAE) and Vietnam have all said they are exploring options for cyber warfare.[1] In turn, there are signs that the states such as the United States, China, Russia, Israel, the United Kingdom, Iran and North Korea continue to further develop their offensive cyber capabilities.[2] Concurrently, many states have adopted cyberspace as a new operational domain of warfare, alongside land, air, space and sea.[3] Also NATO, following the Warsaw Summit, has acknowledged cyberspace as a military domain.[4]

Despite the growing interest in these capabilities, little is known about how states use (or expect to use) OCC to further their national goals. In a recently published report, former US Secretary of Defense, Ashton Carter, expressed his disappointment in the 'cyber component' of US efforts to destroy ISIS.[5] The report highlights an important

1 This is not a comprehensive list of newcomers. On Germany see: Nina Werkhäuser, "German army launches new cyber command", *DW*, (April 1, 2017). Retrieved from: http://www.dw.com/en/german-army-launches-new-cyber-command/a-38246517; on Finland see: Secretariat of the Security Committee, "Finland's Cyber Security Strategy", (2013). Retrieved from: https://www.defmin.fi/files/2378/Finland_s_Cyber_Security_Strategy.pdf; on Vietnam see: Jim Dao, Giang The Huong Tran and Tu Ngoc Trinh, "New Law on Cyber Security in Vietnam", *Tilleke & Gibbins* (2016, June 3). Retrieved from: http://www.tilleke.com/resources/new-law-cyber-security-vietnam; on India see: Vivek Raghuvanshi, "New Indian Cyber Command Urged Following Recent Attacks", *Defense News*, (2016, June 6). Retrieved from: https://www.defensenews.com/2016/06/06/new-indian-cyber-command-urged-following-recent-attacks/; on United Arab Emirates see: Bindiya Thomas, "UAE Military To Set Up Cyber Command", (2014, September 30), *DefenseWorld*. Retrieved from: http://www.defenseworld.net/news/11185/ UAE_Military_To_Set_Up_Cyber_Command#.WW4nJYjyiUk; on Turkey see: Israel Defense, "Turkey Launched Cyber Warfare Command", (2014, April 13). Retrieved from: http://www.israeldefense.co.il/en/content/turkey-launched-cyber-warfare-command; on Columbia see: Christoffer Frendesen "Colombia sends officials to Estonia for cyber defense training", *Columbia Reports*, (2014, September 2). Retrieved from: http://colombiareports.com/colombias-govt-sends-security-forces-estonia-cyber-defense-training/.

2 On Russia see: Eugene Gerden, "Russia to spend $250m strengthening cyber-offensive capabilities", *SC Magazine UK*, (2016, February 4). Retrieved from: http://www.scmagazineuk.com/russia-to-spend-250m-strengthening-cyber-offensive- capabilities/article/470733; on the United States see Sean Lyngaas, "Pentagon Chief: 2017 budget includes $7Bn for cyber", *FCW* (February 2, 2016). Retrieved from: https://fcw.com/articles/2016/02/02/dod-budget-cyber.aspx; on Iran see: Bozorgmehr Sharafedin, "Iran to expand military spending, develop missiles", *Reuters*, (January 9, 2017). Retrieved from: https://www.reuters.com/article/us-iran-military-plan/iran-to-expand-military-spending-develop-missiles-idUSKBN14T15L; on North Korea see: David E. Sanger, David D. Kirkpatrick and Nicole Perlroth, "The World Once Laughed at North Korean Cyberpower. No More", *The New York Times* (October 15, 2017). Retrieved from: https://www.nytimes.com/2017/10/15/world/asia/north-korea-hacking-cyber-sony.html.

3 For a critical analysis on this branding see: Chris McGuffin and Paul Mitchell, "On domains: Cyber and the practice of warfare", *International Journal*, 69:3 (2014):394-412 .

4 NATO CCD COE, "NATO Recognises Cyberspace as a 'Domain of Operations' at Warsaw Summit", (2016, July 21). Retrieved from: https://ccdcoe.org/nato-recognises-cyberspace-domain-operations-warsaw-summit.html.

5 At the inaugural US Cyber Command Symposium, a more positive view of the US cyber operations against ISIS was provided. As one senior policymaker stated: "We are hitting every target, every time". Ashton Carter, "A Lasting Defeat: The Campaign to Destroy ISIS", *Report*, Belfer Center for Science and International Affairs, Harvard Kennedy School, (October, 2017). Retrieved from: https://www.belfercenter.org/LastingDefeat; Max Smeets, "US Cyber Command: An Assiduous Actor, Not a Warmongering Bully", *The Cipher Brief*, (March 4, 2018). Retrieved from: https://www.thecipherbrief.com/us-cyber-command-assiduous-actor-not-warmongering-bully.

set of issues. It rings alarm bells about the current organizational efforts of US Cyber Command.[6] It confirms findings of several scholars that the development of effective cyber capability is by no means an easy feat.[7] It also reveals the importance of contextualizing the US Cyber Command within a larger organizational structure, each component of which has its own institutional interests. Finally, Carter's statement suggests that these capabilities, even though they are very malleable and refer to a broad category of tools, may not be equally valuable in all situations against all types of actors.

The former Secretary of Defense is of course not the first senior policy maker to note disquiet about cyber weapons. In 2012, when Keith Alexander was still heading the NSA and US Cyber Command, he stated that there is "much uncharted territory in the world of cyber-policy, law and doctrine".[8] More recently, referring to Herman Kahn's classic 1959 text on nuclear strategic concepts, Michael Hayden states that "[n]o one has yet begun to write the On Thermonuclear War for cyber conflict".[9]

The purpose of this paper is therefore to explore the following question: *How and to what extent, if any, do offensive cyber capabilities have the potential to affect the roles of military power?* We do not intend to provide a highly detailed policy prescription, nor a detailed description of the requirements for the military to conduct a specific operation. Instead, this paper deals with the basic principles and aims to parsimoniously capture which goals can be realized through the use of OCC. After all, as military theorist Charles Ardant du Picq noted in the mid-19th century, "[t]he instruments of battle are valuable only if one knows how to use them".[10] As a starting point of our analysis, we use the framework developed by Robert J. Art almost four decades ago on the ends of military power. Art distinguished between four strategic roles that force can serve: i) defense, ii) deterrence, iii) compellence and iv) 'swaggering'.[11]

Our central claim is that OCC can alter the manner in which states use their military power. Offensive cyber capabilities are not particularly effective in *deterring* adversary military action, except when threatened to be used by states with a credible

6 "I was largely disappointed in Cyber Command's effectiveness against ISIS. It never really produced any effective cyber weapons or techniques. When CYBERCOM did produce something useful, the intelligence community tended to delay or try to prevent its use, claiming cyber operations would hinder intelligence collection. This would be understandable if we had been getting a steady stream of actionable intel, but we weren't. The State Department, for its part, was unable to cut through the thicket of diplomatic issues involved in working through the host of foreign services that constitute the Internet. In short, none of our agencies showed very well in the cyber fight".

7 Jon Lindsay, "Stuxnet and the Limits of Cyber Warfare", *Security Studies*, 22: 3 (2013)365-404.

8 Keith Alexander, US Senate, Committee on Armed Services, (2014, April). Retrieved from: http://www.eweek.com/security/nsa-director-says-cyber-command-not-trying-to-militarize-cyberspace.

9 Michael Hayden, *Playing the Edge: American Intelligence in the Age of Terror*, (New York: Penguin Press: 2014).

10 Charles Ardant du Picq, *Battle Studies: Ancient and Modern Battle*, trans. John Greely and Robert C. Cotton (New York: Macmillan, 1920).

11 The categories selected by Art are not analytically exhaustive. The categories are described in more detail below. Robert J. Art, "To What Ends Military Power?", *International Security*, 4:4 (1980)3-35.

reputation. However, offensive cyber capabilities do have value in *compellence*. Unlike conventional capabilities, the effects of OCC do not necessarily have to be exposed publicly, which means the compelled party can back down post-action without losing face thus deescalating conflict. The potential opportunity for the attacker to control the reversibility of effect of an OCC may also encourage compliance. At the same time, the use of OCC has escalatory potential. Cyber capabilities also contribute to the use of force for *defensive* purposes, as it could provide both a preemptive as well as preventive strike option. Finally, its symbolic value as a 'prestige weapon' to enhance 'swaggering' remains unclear, due to its largely non-material ontology and transitory nature.

The remainder of this paper consists of three parts. A study on the unique value of cyber capabilities has to start with an analysis of its distinct features. The next section therefore briefly discusses the 'rise' of OCC and assesses its characteristics. Section III, in turn, lays out the four possible functions of cyber capabilities as a tool for the state. The final section concludes and considers the implications of these findings.

2. THE RISE OF OFFENSIVE CYBER CAPABILITIES

The term 'offensive cyber capability' can have a host of different meanings.[12] We define OCC as "a capability designed to access a computer system or network to damage or harm living or material entities".[13] Adopting this definition, it also means that we exclude espionage, information warfare and information operations from our analysis. OCC encompasses a wide range of capabilities. Indeed, the cyber means used against the Ukrainian regional electricity distribution company in December 2015 are very different to those used in the DDoS attacks that swamped websites of various Estonian organizations in April 2007.[14] Rather than compile an exhaustive list of purposes and examples, we have selected three categories based on the damage

12 This is partially because the prefix 'cyber' acts like a sponge absorbing meaning. See: James Shires and Max Smeets, "The Word Cyber Now Means Everything—and Nothing At All", *Slate*, (December 1, 2017). Retrieved from: http://www.slate.com/blogs/future_tense/2017/12/01/the_word_cyber_has_lost_all_meaning.html.

13 Max Smeets, "A Matter of Time: On the Transitory Nature of Cyberweapons", *Journal of Strategic Studies*, (2017)1-28; For alternative definitions see: Thomas Rid and Peter McBurney, "Cyberweapons", *The RUSI Journal*, 157:1 (2012):6-13, p. 7; Trey Herr, "PrEP: A Framework for Malware & Cyber Weapons", *The Journal of Information Warfare*, 13:1(2014) ; Dale Peterson. "Offensive Cyber Weapons: Construction, Development and Employment", *Journal of Strategic Studies*, 36:1(2013).

14 A detailed analysis of each case goes beyond the scope of this paper. For an excellent overview on Ukraine see: Kim Zetter, "Everything We Know About Ukraine's Power Plant Hack", *Wired*, (20 January 2016). Retrieved from: https://www.wired.com/2016/01/everything-we-know-about-ukraines-power-plant-hack/; Kaspersky Lab's Global Research & Analysis Team, "BlackEnergy APT Attacks in Ukraine employ spearphishing with Word documents", Securelist, (28 January 2016). Retrieved from: https://securelist.com/blackenergy-apt-attacks-in-ukraine-employ-spearphishing-with-word-documents/73440/; Kim Zetter, "Inside the Cunning, Unprecedented Hack of Ukraine's Power Grid", *Wired*, (3 March 2016). Retrieved from: https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/;E-ISAC, SANS ICS. "Analysis of the Cyber Attack on the Ukrainian Power Grid" March 18, 2016, 4. http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf.

caused by an OCC: denial of service, file damage and physical damage.[15] Table 1 provides an overview of some of the most important cases reported by a reputable cyber security firm.

TABLE 1. IMPORTANT INSTANCES OF OCC

Denial of Service		File Damage		Physical Damage	
Name	**Year***	**Name**	**Year***	**Name**	**Year***
Estonian DDoS attacks	2007	Witty Worm	2004	Stuxnet	2010
Hacking Scientology	2008	Dozer	2009	Ukraine attacks	2015
Georgian attacks	2009	Koredos	2010		
Black DDoS	2010	Shamoon	2012		
OPI Israel	2012	Groovemonitor	2012		
		Jokra / Dark Seoul	2013		
		Destover / Sony	2014		
		Shamoon 2.0	2016		
		NotPetya	2017		

* We listed year of disclosure rather than year of compromise. **The table does not include cases of which there is no public cyber security report available, like Sands Casino in 2014.

The deployment and use of OCCs is generally extended over multiple stages. It is common to distinguish between the following four stages for advanced operations: i) reconnaissance; ii) intrusion; iii) privilege escalation; and iv) payload delivery.[16] These stages can be explained through a simple analogy of a burglar trying to get into a house. The burglar first scans the neighborhood and sees which security measures (camera system, dog, locks) the homeowner has taken (reconnaissance). The burglar then tries to get in, normally taking the path of least resistance (intrusion). When entering a specific room, they try to gain access to other rooms and hope to find the cabinet with all the keys to the cars, vault etc. (privilege escalation). Finally, the burglar decides what to do with the obtained level of access. They may not only steal the belongings of the homeowner, but also move or destroy some of the furniture in the house. Considering these stages reveals that there are close similarities between OCC and cyber espionage capabilities or, in intelligence jargon, Computer Network

15 These categories were adopted from: Steven M. Bellovin, Susan Landau and Herbert S. Lin, "Limiting the undesired impact of cyber weapons: technical requirements and policy implications", *Journal of Cybersecurity*, 3:1 (2017)59–68.

16 For example, see: FireEye, "Advanced Targeted Attacks: How to Protect Against the Next Generation of Cyber Attacks", *WhitePaper*, (2012). Retrieved from: http://www.softbox.co.uk/pub/ reeye- advanced-targeted-attacks.pdf; S. Mathew, R. Giomundo, S. Upadyaya, M. Sudit and A. Stotz, "Understanding Multistage Attacks by Attack-Track based Visualization of Heterogeneous Event Streams," *VizSEC '06, Proceedings of the 3rd International Workshop on Visualization for Computer Security* (2016)1-6.

Exploitation (CNE) and Computer Network Attack (CNA). Indeed, it is often said that there is no other weapon so strongly anchored in intelligence as cyber weapons.[17]

3. THE USES OF CYBER FORCE

Having developed a better understanding of the nature of OCC, we can now turn to potential function of these capabilities. Numerous works in security studies have been devoted to the use of force. We used the classic study of Robert J. Art – *To What Ends Military Power?* – as a starting point for our analysis. Art distinguishes between four categories that force can serve: defense, deterrence, compellence and 'swaggering'.[18]

A. Defense

The defensive use of military force serves to do two things: avert an attack or minimize damage of an attack. As Art states:

> "[f]or defensive purposes, a state will direct its forces against those of a potential or actual attacker, but not against his unarmed population. For defensive purposes, a state can deploy its forces in place prior to an attack, use them after an attack has occurred to repel it, or strike first if it believes that an attack upon it is imminent or inevitable".[19]

We commonly distinguish between a preemptive and preventive strike. A preemptive strike is when a state believes an attack upon it is imminent by an adversary. A preventive strike is when an attack is perceived to be inevitable but not imminent or known to be planned.[20]

Two prominent cases of preventive strikes in the late Cold War include Operation Scorch Sword, an airstrike by the Iranian air force in September 1980 that damaged an almost-complete nuclear reactor near Baghdad, Iraq and Operation Opera, the more successful bombing by the Israeli air force of the same nuclear reactor, almost a

17 This in turn leads to an important set of questions surrounding the organizational integration of intelligence and military capabilities. See: Max Smeets, "Organisational Integration of Offensive Cyber Capabilities: A Primer on the Benefits and Risks", 9th International Conference on Cyber Conflict, (Tallinn: NATO CCD COE Publications: 2017); Hayden, *Playing the Edge*.

18 In practice, these categories are expected to overlap and may not always be easily disentangled. Also, unlike Art, we do not explicitly distinguish between the physical and peaceful use of military power. Art, "To What Ends Military Power?".

19 Ibid. Though note that even for offensive purposes, states are prohibited from attacking unarmed populations.

20 For an excellent overview on the need to legitimize preventive and pre-emptive use of force see: Tom Sauer, "The Preventive and Pre-Emptive Use of Force: To be Legitimized or to be De-Legitimized?" The Hoover Institution. Retrieved from: http://www.ethical-perspectives.be/viewpic.php?TABLE=EP&ID=493.

year later. Stuxnet can be similarly described as a preventive strike.[21] As Kim Zetter notes, in the lead up to the cyber attack, technicians at Natanz had begun to install new centrifuges again at a rapid rate and with their performance improving.[22] Stuxnet was presented as an 'extra option' to President George W. Bush, as Sanger notes, to effectively deal with a seemingly escalating situation, especially in the eyes of the Israeli government.[23] Stuxnet was a masterpiece of work, "[b]ut Stuxnet might only have been the beginning", as Ben Buchanan notes.[24] Indeed, there was also an option developed for a large scale pre-emptive strike. In case the situation in Iran worsened, the United States had a contingency planned, reportedly code-named NITRO ZEUS. As *The New York Times* reported:

> "Nitro Zeus was part of an effort to assure President Obama that he had alternatives, short of a full-scale war, if Iran lashed out at the United States or its allies in the region. […] [T]he plan […] was devised to disable Iran's air defenses, communications systems and crucial parts of its power grid and was shelved, at least for the foreseeable future, after the nuclear deal struck between Iran and six other nations last summer [2016] was fulfilled".[25]

Although NITRO ZEUS is the only pre-emptive cyber strike option known to date, it is likely that military forces have considered the use of OCC in this manner for other situations as well, albeit on a more modest scale. Indeed, the use of a cyber capability to, for instance, neutralize the launch of an operational ballistic missile is conceivable.

B. Deterrence

The deterrent use of military force aims to dissuade an adversary from doing something by threatening him with unacceptable punishment if he does it. Deterrence hinges upon the credible threat of retaliation to dissuade an enemy from attacking. As Bernard Brodie wrote in 1958, a credible deterrent, "must be always at the ready, yet

21 Ralph Langner indicates that Stuxnet is actually not one weapon, but two. The earliest version, also referred to as Stuxnet 0.5, was in development prior to November 2005. This early version is considered to be the most sophisticated of the two, focusing on the closing the isolation valves of the Natanz uranium enrichment facility. The latter, better-known version followed a different modus operandi as it aimed to change the speeds of the rotors in the centrifuges. Ralph Langner, "Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve", (2013, November). Retrieved from: https://www.langner.com/wp-content/uploads/2017/03/to-kill-a-centrifuge.pdf; Operation Orchard led by the Israeli air force could be seen as an example of a combined preventive strike with kinetic and cyber means.

22 Kim Zetter, *Countdown to Zero day: Stuxnet and the Launch of the World's First Digital Weapon*, (New York: Crown Publishing: 2014).

23 David Sanger, *Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power*, (New York: Broadway Paperbacks: 2012).

24 Ben Buchanan, *The Cybersecurity Dilemma: Network Intrusions, Trust and Fear in the International System*, (Oxford: Oxford University Press: 2017).

25 David E. Sanger and Mark Mazetti, "US Had Cyberattack Plan if Iran Nuclear Dispute Led to Conflict", *The New York Times*, (2016, February 16). Retrieved from: https://www.nytimes.com/2016/02/17/world/middleeast/us-had-cyberattack-planned-if-iran-nuclear-negotiations-failed.html; James Ball, "US Hacked Into Iran's Critical Civilian Infrastructure For Massive Cyberattack, New Film Claims", *BuzzFeed*, (2016, February 16). Retrieved from: https://www.buzzfeed.com/jamesball/us-hacked-into-irans-critical-civilian-infrastructure-for-ma?utm_term=.ile5noYzJy#.kyVJaBdP87.

never used".[26] Defense does not necessarily buy deterrence, nor deterrence defense.[27] Where defense dissuades the adversary by means of presenting an unvanquishable military force, deterrence dissuades by presenting the certainty of a retaliatory devastation.[28]

Few cyber conflict topics have received more attention than cyber deterrence. For the most part, the existing literature uses the term to refer to deterrence of cyberattacks by an adversary, and can be grouped into three buckets. The first group of scholars argue that cyber deterrence does not have distinctive problems and works (or occasionally fails) like conventional deterrence. Dorothy Denning believes that cyberspace strongly resembles traditional domains.[29] According to her, cyber deterrence can therefore be achieved through existing regimes.[30] The second group of scholars believes that cyber deterrence has its unique set of issues, but as long as we further specify the issue area, the problems can largely be solved. Joseph Nye Jr.'s discussion of deterrence is a prominent example.[31] He notes that conventional cyber deterrence is difficult, but we could instead focus on deterrence by economic entanglement and norms to overcome barriers.[32] Lucas Kello argues that cyber deterrence does not work as a strategy, but we could aim for punctuated deterrence instead; we should not deter individual actions but a series of actions.[33] The last group of scholars argues that cyber deterrence does not work and will *never* work. Richard Harknett argues that cyber deterrence is impossible due to the structure of cyberspace.[34] In his view, we need to move away from the deterrence paradigm and consider different forms of strategy, such as persistence.[35] This paper does not address cyber deterrence as defined above; instead, it focuses on the use of a cyber capability to deter a certain type of (military) means of an adversary.

26 Bernard Brodie, "The Anatomy of Deterrence", *RAND Corporation*, (1958, July 23). Retrieved from: https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM2218.pdf.

27 Art, "To What Ends Military Power", p. 7.

28 Ibid. Some scholars instead distinguish between deterrence by detail and deterrence by punishment.

29 The scholars note that "Studies of 'cyber deterrence' raise as many problems as would be raised by a comparable study of 'land deterrence.' Dorothy E. Denning, "Rethinking the Cyber Domain and Deterrence", *JFQ*, 77 (2015)8-15. Retrieved from: http://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-77/jfq-77_8-15_Denning.pdf, p. 15.

30 Ibid.

31 Joseph S. Nye, "Deterrence and Dissuasion in Cyberspace", *International Security*, 43:3 (Winter, 2016/2017)44-71.

32 Ibid.

33 Lucas Kello, *Virtual Weapon and International Order*, (Yale: Yale University Press: 2017); also see: Uri Tor, "'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence", *Journal of Strategic Studies*, 40:1-2(2017)92-117.

34 Richard J. Harknett and Joseph S. Nye, "Is Deterrence Possible in Cyberspace?" International Security, 42:2 (2017)196-199; Also see: Brad D. William, Meet the scholar challenging the cyber deterrence paradigm, (July 19, 2017) *The Fifth Domain*. Retrieved from: https://www.fifthdomain.com/home/2017/07/19/meet-the-scholar-challenging-the-cyber-deterrence-paradigm/; Richard J. Harknett and Michael P. Fischerkeller, "Deterrence is Not a Credible Strategy for Cyberspace", Orbis 61:3 (2017)381-393.

35 Ibid.

OCC tend to be transitory in nature, meaning they only have the "temporary ability to access a computer system or network to cause harm or damage to living and material entities".[36] The transitory nature of a capability is determined by both technical (e.g. type of vulnerability, access and payload used) and non-technical (e.g. the number and type of actors the capability is used against) factors.[37] This feature, combined with their clandestine nature, makes it difficult to prove you have a specific type of capability pre-deployment. Hence, state actors can talk about offensive cyber capabilities whether or not they actually have them; such talk is intended to convey to other actors the impression that the talking nation does have the talked-about capabilities. But since the fact of possession cannot be verified by those other actors nor demonstrated by the talking state, such talk is cheap talk.[38]

Cheap talk, however, is not by definition meaningless and may under certain circumstances still have an impact. One of the key factors which is said to affect the effectiveness of cheap talk is reputation.[39] More specifically, post-hoc revelations about an actor's capability – either intentionally or non-intentionally – can add to the reputation and credibility of the actor's cheap talk on the intention and ability to conduct an offensive cyber operation. This has led to a number of paradoxical dynamics for cyber conflict.

The release of the classified National Security Agency (NSA) documents by Edward Snowden has been described as the most embarrassing episode in the history of the secretive US intelligence agency. It revealed how the NSA maintained a mass-surveillance program over its own citizens, accessed data from companies, intercepted data from global communications networks and stored information of millions of people. Yet, it also exposed the impressive arsenal of the agency. Not least from the Snowden disclosures, *The Washington Post* reported that the US government mounted at least 231 offensive cyber operations in 2011.[40] As Gompert and Libicki note, in

36 Max Smeets, "A Matter of Time".

37 OCC exploiting software vulnerabilities are both quantitatively and qualitatively different from conventional weapons in their transitory nature. They are quantitatively different as the introduction of countermeasures - that is, the remediation (patching) of vulnerabilities - occurs on a very rapid and continuing basis. They are also qualitatively different; patching does not only prevent successful exploitation against one system but against any administrator uploading the patch. Even though there are different ways in which patches can be distributed after a software vulnerability is exploited, a defense for one creates a defense for all. Ibid.

38 Joseph Farrell and Matthew Rabin, "Cheap Talk", *Journal of Economic Perspectives*, 10:3 (1996):103-118; Clayton L. Thyne, "Cheap Signals with Costly Consequences: The Effect of Interstate Relations on Civil War", *Journal of Conflict Resolution*, 50:6 (2006)937-961; Joseph Farrell and Robert Gibbons, "Cheap Talk with Two Audiences", *The American Economic Review*, 79:5 (1989)1214-1223.

39 Thomas Schelling, *Arms and Influence* (Yale: Yale University Press: 1966), p.124; Alexandra Guisinger and Alastair Smith, "Honest threats: The interaction of reputation and political institutions in international crises", *Journal of Conflict Resolution*, 46: (2002)175-200; Anne Sartori, "The Might of the Pen: A Reputational Theory of Communication in International Disputes", *International Organization*, 56 (2002)121-50.

40 Barton Gellman and Ellen Nakashima, "US spy agencies mounted 231 offensive cyber-operations in 2011, documents show", *The Washington Post*, (2013, August 30). Retrieved from: https://www.washingtonpost.com/world/national-security/us-spy-agencies-mounted-231-offensive-cyber-operations-in-2011-documents-show/2013/08/30/d090a6ae-119e-11e3-b4cb-fd7ce041d814_story.html.

this way, the leaks have ironically "helped it to broadcast how deeply the NSA can supposedly burrow into the systems of others".[41]

Overall, it is more difficult to use OCC as means to deter compared to most other forms of military force. However, it does not mean that it is impossible at all. Especially if an actor is able to show repeatedly what is capable and willing of doing through cyber means it can benefit from this reputation in the future.[42]

C. Compellence

The term compellence in International Relations originates from Thomas Schelling, conceptualizing it as the second form of coercion alongside deterrence.[43] The compellent use of military force serves one of two purposes: i) to stop an activity undertaken by an adversary, or ii) to get an adversary to do something he has not yet undertaken.

The difference between deterrence and compellence hinges upon initiative and timing. The deterrent use of force is based on a promised reaction following an action of the adversary, the timing of which is in principle automatic. The compellent use of force, in turn, is based on a more active strategy of the threatener. For compellence, timing is a critical factor: "too strict a deadline makes compliance impossible, while one too lenient makes compliance unnecessary".[44] Deterrence is usually said to be easier to achieve than compellence; as the deterred party need not to do anything visible, it does not suffer from any reputational damage and can simply argue or imply that it never intended to conduct the activity.

Cyber capabilities have a distinct advantage in this respect. Its effects do not necessarily have to be exposed publicly, which means the compelled party can back down post-action without losing face. More specifically, the compelled actor can deny that the effect was caused by OCC. For example, a three-day disruption of computer systems at an airport leading to massive financial losses and delays could be attributed to a 'general system failure' (a company mistake) whilst in reality it was due to a cyber attack.

This opens up new opportunities for the use of force, although it is dependent on a number of conditions. Not least, the cyber attack needs to cause significant levels of harm or damage to be perceived as a substantial enough cost to change action and delineate the action from the 'constant state' of cyber activity. Whereas plausible deniability is often an advantage to the attacker, in this case the actor should find a

41 David C. Gompert and Martin Libicki, "Waging Cyber War the American Way" Survival, 57:4 (2015)7-28; also see: Martin Libicki, *Cyberspace in Peace and War*, (Annapolis, Naval Institute Press: 2016), p. 198.

42 It remains unclear however whether the Snowden revelations helped deterrence or not.

43 Schelling, *Arms and Influence*, p. 69–91.

44 Gregory F. Treverton, "Framing Compellent Strategies", *RAND Corporation* (2000). Retrieved from: http://slantchev.ucsd.edu/courses/pdf/treverton-compellence.pdf.

way – either through the design of the weapon or other means – to show that it is conducting this cyber attack in response to the adversary's activity.[45] Finally, in case a compelled actor does not want to reveal it has been attacked, a cyber security firm could instead write a public report exposing the activity.[46] As much of the attribution capability lies with private companies, oftentimes having a strong incentive to publish, this could be a serious concern for states.[47]

OCCs have another distinct advantage when it comes to the compellent use of military force. Unlike kinetic weapons, the attacker can sometimes control the reversibility of the effects of cyber capabilities. Control is based on two dimensions: i) "the adversary's inability to stop or revert the effects of the cyber attack"; and ii) "[the] attacker's ability to stop or revert the effects of the attack at any given time desired".[48] The most detailed account on how reversibility may be achieved is provided by Neil Rowe describing four techniques: i) reversible cryptography, where data is encrypted to prevent use, but can be decrypted after adversary complies; ii) system obfuscation, in which a computer is obfuscated in a reversible manner; iii) data retainment and restoration, where important data is withheld but can be restored; and iv) compromise deception in which adversaries mistakenly think that their system is compromised, but after compliance find out they have been deceived.[49]

The potential reversibility of effect of an OCC may encourage compliance. The adversary may know that, if it backs down, the 'old' situation can be restored. A simple characterization of a conventional situation may be: 'I will keep bombing your critical infrastructure until you stop attacking me'. In this situation, the utility the attacker gains by ceasing the attack is that no further costs (i.e. damage to its critical infrastructure) will be incurred. But the attacker still has to take in its earlier infrastructure losses that were caused during the initial stages of the conflict. In the

45 See discussion on 'loud cyber weapons', which has primarily been about how to "possibly deter future intrusions". Yet, as this discussion suggests, it should also be considered for the compellent use of force. Chris Bing, "US Cyber Command director: We want 'loud,' offensive cyber tools", *FedScoop*, (2016, August 3). Retrieved from: https://www.fedscoop.com/us-cyber-command-offensive-cybersecurity-nsa-august-2016; Herb Lin, "Developing "Loud" Cyber Weapons", *Lawfare*, (2016, September 1). Retrieved from: https://www.lawfareblog.com/developing-loud-cyber-weapons; Herb Lin, "Still More on Loud Cyber Weapons", *Lawfare*, (2016, October 19). Retrieved from: https://www.lawfareblog.com/still-more-loud-cyber-weapons.

46 In the case of Stuxnet, for example, the Iranian government has for a long time denied its systems were compromised. Instead, it was researchers from VirusBlokAda, Symantec and the Langner group which initially reported on the sophisticated attacked.

47 Also see: Max Smeets, "The Strategic Promise of Offensive Cyber Operations", *Strategic Studies Quarterly*, Forthcoming.

48 Ibid.

49 Neil Rowe, "Towards Reversible Cyberattacks", *Proceedings of the 9th European Conference on Information Warfare and Security*, ed. J. Demergis (Reading: Academic Publishing Ltd: 2010), 261-267. Note, however, that reversibility is often a question of time scale. The kinetic destruction of a bridge can be "reversed" by rebuilding the bridge, albeit over a time scale of weeks or months rather than minutes. And in any case, a human death that results from a "reversible" cyberattack on a critical system will not be resurrected when the effects of that cyberattack are reversed. That is, while the direct effects of a cyber capability may be reversible, the consequential effects are almost never reversible. The key issue of reversibility lies in the fact that the reversibility can be implemented by the attacker rather than the defender.

case of a cyber attack, the scenario may be characterized as follows: 'I will corrupt data on 'X' amount of your critical computer systems for every day you keep attacking me'. In this situation, the incentive structure for the attacker has changed; if the actor backs down it will no longer incur costs in the future and retrieves earlier corrupted data.

D. Swaggering

Whereas defense, deterrence and compellence are widely used concepts, 'swaggering' is not part of the common political science vocabulary.[50] As Art indicates:

> "[s]waggering is in part a residual category, the deployment of military power for purposes other than defense, deterrence, or compellence. Force is not aimed directly at dissuading another state from attacking, at repelling attacks, nor at compelling it to do something specific. The objectives for swaggering are more diffuse, ill-defined and problematic than that. Swaggering almost always involves only the peaceful use of force and is expressed usually in one of two ways: displaying one's military might at military exercises and national demonstrations and buying or building the era's most prestigious weapons. The swagger use of force is the most egoistic: it aims to enhance the national pride of a people or to satisfy the personal ambitions of its ruler [...] Swaggering is pursued because of the fundamental yearning of states and statesmen for respect and prestige".[51]

OCC seem to be less valuable for swaggering purposes.[52] Cyber capabilities have a largely non-material ontology, making it difficult to publicly showcase or 'parade' these capabilities. Second, the transitory nature of cyber capabilities is also a problem for swaggering. Cyber capabilities' transitory nature is primarily due to the malleability of cyberspace affecting the life-cycle of a vulnerability and effectiveness of an OCC. The life cycle of vulnerabilities is subject to three delays: i) the awareness delay; ii) the patching delay; and iii) the adaptation delay.[53] The moment actors reveal their capability, it inevitably increases the likelihood of a vendor learning about the vulnerability and assigning a high level of priority to developing a patch (i.e. reducing the awareness and patching delay).[54] Overall, as a document from the East West Institute concludes:

50 The concept has been used once before in relation to cyber attacks by Neuman and Poznansky. They however misapplied the concept as swaggering is not a form of coercion. Craig Neuman and Michael Poznansky, "Swaggering in Cyberspace: Busting the conventional wisdom and cyber coercion", *War on the Rocks*, (2016, June 28). Retrieved from: https://warontherocks.com/2016/06/swaggering-in-cyberspace-busting-the-conventional-wisdom-on-cyber-coercion/.

51 Art, "To What Ends Military Power", p. 10-11.

52 However, this does not mean that a cyber *command or program* cannot be established for prestige purposes.

53 Smeets, "A Matter of Time".

54 Ibid.

> "[m]ilitary forces will have distinct interests in keeping cyber weapons secret. [...] Those nations that are developing the most advanced weapons have a strong interest in being able to protect the intelligence surrounding such capabilities".[55]

4. CONCLUSION

Considering the growing interest in the use of offensive cyber capabilities as a tool for the state, this study assessed to what degree these capabilities have the potential to change the role of military power. We have shown that OCCs have the potential to significantly affect how states use their military power in several ways. First, OCCs have downgraded the role of *deterrence*, except for those states with a credible reputation for being able and willing to conduct offensive cyber operations. However, we indicated that *compellence* is no longer ruled out as a function of military power considering several features of cyber capabilities. Unlike conventional capabilities, the effects of offensive cyber capabilities do not necessarily have to be exposed publicly, which means the compelled party can back down post-action without losing face. The potential to control the reversibility of effect of a cyber capability by the attacker may also encourage compliance. As OCCs can be used as both a preemptive and a preventive strike option, it reemphasizes the potential to use of force for *defensive* purposes. Finally, due to its largely non-material ontology and transitory nature, its symbolic value as a prestige weapon to enhance swaggering remains unclear.

Major powers reap benefits from their nuclear arsenal without using them physically and risk high costs when they are used. This in turn incentivizes the avoidance of warlike behavior and exploitation of peaceful use. Yet, this logic breaks down for cyber capabilities: the benefits from non-use are lower given the limits of deterrence and swaggering; the costs of non-use are higher due to the transitory nature of these capabilities; and the risks of using cyber capabilities are lower. Overall, it means less powerful incentives exist for restraint.

As we have only provided a primer on the topic, there are several avenues for future research. This paper was consciously limited to only assess the role of OCCs with regard to state power. Given that OCCs are normally part of a broader arsenal of capabilities, it is important to discuss the military use of OCC in relation to military capabilities. Further research may therefore conduct a comparative analysis of other assets (nuclear weapons, drones, covert actions) to gain a more holistic understanding of the military contribution of each capability. Also, it has been noted that the growth of the private sector market for OCC leads to new opportunities for states to acquire,

55 EastWest Institute, "Working Towards Rules for Governing Cyber Conflict Rendering the Geneva and Hague Conventions in Cyberspace", (2011). Retrieved from: https://www.eastwest.ngo/sites/default/files/ideas-files/US-Russia%20(1).pdf.

deploy and use these capabilities. It remains unclear, however, to what degree this trend also changes the way in which OCCs can be strategically used by states as a function of military power.

REFERENCES

Alexander, Keith, US Senate, Committee on Armed Services, (2014, April). Retrieved from: http://www.eweek.com/security/nsa-director-says-cyber-command-not-trying-to-militarize-cyberspace.

Anonymous, "Magnitude 4.3 – NORTH KOREA", USGS, (2006, October 9). Retrieved from: https://web.archive.org/web/20140427050803/http://earthquake.usgs.gov/earthquakes/eqinthenews/2006/ustqab/.

Ardant du Picq, Charles, *Battle Studies: Ancient and Modern Battle*, trans. John Greely and Robert C. Cotton (New York: Macmillan, 1920).

Art, Robert J., "To What Ends Military Power?", *International Security*, 4:4 (1980)3-35.

Ball, James, "US Hacked into Iran's Critical Civilian Infrastructure for Massive Cyberattack, New Film Claims", *BuzzFeed*, (2016, February 16). Retrieved from: https://www.buzzfeed.com/jamesball/us-hacked-into-irans-critical-civilian-infrastructure-for-ma?utm_term=.ile5noYzJy#.kyVJaBdP87.

Bellovin, Steven M., Susan Landau and Herbert S. Lin, "Limiting the undesired impact of cyber weapons: technical requirements and policy implications", *Journal of Cybersecurity*, 3:1 (2017)59–68.

Bing, Chris, "US Cyber Command director: We want 'loud,' offensive cyber tools", *FedScoop*, (2016, August 3). Retrieved from: https://www.fedscoop.com/us-cyber-command-offensive-cybersecurity-nsa-august-2016.

Brodie, Bernard, "The Anatomy of Deterrence", RAND Corporation, (1958, July 23). Retrieved from: https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM2218.pdf.

Buchanan, Ben, *The Cybersecurity Dilemma: Network Intrusions, Trust and Fear in the International System*, (Oxford: Oxford University Press: 2017).

Carter, Ashton, "A Lasting Defeat: The Campaign to Destroy ISIS", Report, Belfer Center for Science and International Affairs, Harvard Kennedy School, (October 2017). Retrieved from: https://www.belfercenter.org/LastingDefeat.

Collier, Jamie, "State Proxies & Plausible Deniability: Challenging Conventional Wisdom", *Cybersecurity Intelligence*, (2015, September 24). Retrieved from: https://www.cybersecurityintelligence.com/blog/state-proxies-and-plausible-deniability-challenging-conventional-wisdom-644.html.

Dao, Jim, Giang The Huong Tran and Tu Ngoc Trinh, "New Law on Cyber Security in Vietnam", *Tilleke & Gibbins* (2016, June 3). Retrieved from: http://www.tilleke.com/resources/new-law-cyber-security-vietnam.

Denning, Dorothy E., "Rethinking the Cyber Domain and Deterrence", *JFQ*, 77 (2015)8-15. Retrieved from: http://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-77/jfq-77_8-15_Denning.pdf.

E-ISAC, SANS ICS, "Analysis of the Cyber Attack on the Ukrainian Power Grid" (2016, March 18). Retrieved from: http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf.

EastWest Institute, "Working Towards Rules for Governing Cyber Conflict Rendering the Geneva and Hague Conventions in Cyberspace", (2011). Retrieved from: https://www.eastwest.ngo/sites/default/files/ideas-files/US-Russia%20(1).pdf.

Farrell, Joseph, and Matthew Rabin, "Cheap Talk", *Journal of Economic Perspectives*, 10:3 (1996):103-118.

Farrell, Joseph, and Robert Gibbons, "Cheap Talk with Two Audiences", *The American Economic Review*, 79:5 (1989)1214-1223.

FireEye, "Advanced Targeted Attacks: How to Protect Against the Next Generation of Cyber Attacks", *WhitePaper*, (2012). Retrieved from: http://www.softbox.co.uk/pub/ reeye- advanced-targeted-attacks.pdf.

Frendesen, Christoffer, "Colombia sends officials to Estonia for cyber defense training", *Columbia Reports*, (2014, September 2). Retrieved from: http://colombiareports.com/colombias-govt-sends-security-forces-estonia-cyber-defense-training/.

Gellman, Barton, and Ellen Nakashima, "US spy agencies mounted 231 offensive cyber-operations in 2011, documents show", *The Washington Post*, (2013, August 30). Retrieved from: https://www.washingtonpost.com/world/national-security/us-spy-agencies-mounted-231-offensive-cyber-operations-in-2011-documents-show/2013/08/30/d090a6ae-119e-11e3-b4cb-fd7ce041d814_story.html.

Gerden, Eugene, "Russia to spend $250m strengthening cyber-offensive capabilities", *SC Magazine UK*, (2016, February 4). Retrieved from: http://www.scmagazineuk.com/russia-to-spend-250m-strengthening-cyber-offensive- capabilities/article/470733.

Gompert, David C., and Martin Libicki, "Waging Cyber War the American Way", *Survival*, 57:4 (2015)7-28.

Guisinger, Alexandra, and Alastair Smith, "Honest threats: The interaction of reputation and political institutions in international crises", *Journal of Conflict Resolution*, 46: (2002)175-200.

Harknett, Richard J., and Joseph S. Nye, "Is Deterrence Possible in Cyberspace?" *International Security*, 42:2 (2017)196-199.

Harknett, Richard J., and Michael P. Fischerkeller, "Deterrence is Not a Credible Strategy for Cyberspace", *Orbis* 61:3 (2017)381-393.

Hayden, Michael, *Playing the Edge: American Intelligence in the Age of Terror*, (New York: Penguin Press: 2014).

Herr, Trey, "PrEP: A Framework for Malware & Cyber Weapons", *The Journal of Information Warfare*, 13:1(2014).

Israel Defense, "Turkey Launched Cyber Warfare Command", (2014, April 13). Retrieved from: http://www.israeldefense.co.il/en/content/turkey-launched-cyber-warfare-command.

Kaspersky Lab's Global Research & Analysis Team, "BlackEnergy APT Attacks in Ukraine employ spearphishing with Word documents", Securelist, (2016, January 28). Retrieved from: https://securelist.com/blackenergy-apt-attacks-in-ukraine-employ-spearphishing-with-word-documents/73440/.

Kello, Lucas, *Virtual Weapon and International Order*, (Yale: Yale University Press: 2017).

Langner, Ralph, "Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve", (2013, November). Retrieved from: https://www.langner.com/wp-content/uploads/2017/03/to-kill-a-centrifuge.pdf.

Libicki, Martin, *Cyberspace in Peace and War*, (Annapolis, Naval Institute Press: 2016).

Lin, Herbert, "Developing 'Loud' Cyber Weapons", *Lawfare*, (2016, September 1). Retrieved from: https://www.lawfareblog.com/developing-loud-cyber-weapons.

Lin, Herbert, "Still More on Loud Cyber Weapons", *Lawfare*, (2016, October 19). Retrieved from: https://www.lawfareblog.com/still-more-loud-cyber-weapons.

Lindsay, Jon, "Stuxnet and the Limits of Cyber Warfare", *Security Studies*, 22:3 (2013)365-404.

Lyngaas, Sean, "Pentagon Chief: 2017 budget includes $7B for cyber", *FCW* (February 2, 2016). Retrieved from: https://fcw.com/articles/2016/02/02/dod-budget-cyber.aspx.

Mathew, S., R. Giomundo, S. Upadyaya, M. Sudit and A. Stotz, "Understanding Multistage Attacks by Attack-Track based Visualization of Heterogeneous Event Streams", *VizSEC '06, Proceedings of the 3rd International Workshop on Visualization for Computer Security* (2016)1-6.

McGuffin, Chris, and Paul Mitchell, "On domains: Cyber and the practice of warfare", *International Journal*, 69:3 (2014):394-412.

Michael, Melissa, "NotPetya and Wannacry: Have we seen the last?" *F-Secure* (2017, July 7). Retrieved from: https://business.f-secure.com/notpetya-and-wannacry-have-we-seen-the-last.

NATO CCD COE, "NATO Recognises Cyberspace as a 'Domain of Operations' at Warsaw Summit", (2016, July 21). Retrieved from: https://ccdcoe.org/nato-recognises-cyberspace-domain-operations-warsaw-summit.html.

Neuman, Craig, and Michael Poznansky, "Swaggering in Cyberspace: Busting the conventional wisdom and cyber coercion", *War on the Rocks*, (2016, June 28). Retrieved from: https://warontherocks.com/2016/06/swaggering-in-cyberspace-busting-the-conventional-wisdom-on-cyber-coercion/.

Nye, Joseph S., "Deterrence and Dissuasion in Cyberspace", *International Security*, 43:3 (Winter, 2016/2017)44-71.

Peterson, Dale, "Offensive Cyber Weapons: Construction, Development and Employment", *Journal of Strategic Studies*, (2013)36:1.

Raghuvanshi, Vivek, "New Indian Cyber Command Urged Following Recent Attacks", *Defense News*, (2016, June 6). Retrieved from: https://www.defensenews.com/2016/06/06/new-indian-cyber-command-urged-following-recent-attacks/.

Rid, Thomas, and Ben Buchanan, "Attributing Cyber Attacks", *Journal of Strategic Studies*, 38:1-2 (2015)4-37.

Rid, Thomas, and Peter McBurney, "Cyberweapons", *The RUSI Journal*, 157:1 (2012):6-13.

Rowe, Neil, "Towards Reversible Cyberattacks", *Proceedings of the 9th European Conference on Information Warfare and Security*, ed. J. Demergis (Reading: Academic Publishing Ltd: 2010), 261-267.

Sanger, David E., and Mark Mazetti, "US Had Cyberattack Plan if Iran Nuclear Dispute Led to Conflict", *The New York Times*, (2016, February 16). Retrieved from: https://www.nytimes.com/2016/02/17/world/middleeast/us-had-cyberattack-planned-if-iran-nuclear-negotiations-failed.html.

Sanger, David E., David D. Kirkpatrick and Nicole Perlroth, "The World Once Laughed at North Korean Cyberpower. No More", *The New York Times* (October 15, 2017). Retrieved from: https://www.nytimes.com/2017/10/15/world/asia/north-korea-hacking-cyber-sony.html.

Sanger, David, Confront and Conceal: *Obama's Secret Wars and Surprising Use of American Power*, (New York: Broadway Paperbacks: 2012).

Sartori, Anne, "The Might of the Pen: A Reputational Theory of Communication in International Disputes", *International Organization*, 56 (2002)121-50.

Sauer, Tom, "The Preventive and Pre-Emptive Use of Force: To be Legitimized or to be De-Legitimized?", *The Hoover Institution*. Retrieved from: http://www.ethical-perspectives.be/viewpic.php?TABLE=EP&ID=493.

Schelling, Thomas, *Arms and Influence*, (Yale: Yale University Press: 1966).

Secretariat of the Security Committee, “Finland’s Cyber Security Strategy”, (2013). Retrieved from: https://www.defmin.fi/files/2378/Finland_s_Cyber_Security_Strategy.pdf.

Sharafedin, Bozorgmehr, “Iran to expand military spending, develop missiles”, *Reuters*, (2017, January 9). Retrieved from: https://www.reuters.com/article/us-iran-military-plan/iran-to-expand-military-spending-develop-missiles-idUSKBN14T15L.

Shires, James, and Max Smeets, “The Word Cyber Now Means Everything—and Nothing at All”, *Slate*, (2017, December 1). Retrieved from: http://www.slate.com/blogs/future_tense/2017/12/01/the_word_cyber_has_lost_all_meaning.html.

Smeets, Max, “A matter of time: On the transitory nature of cyberweapons”, *Journal of Strategic Studies*, (2017)1-28.

Smeets, Max, “Organisational Integration of Offensive Cyber Capabilities: A Primer on the Benefits and Risks”, *9th International Conference on Cyber Conflict*, (Tallinn: NATO CCD COE Publications: 2017).

Smeets, Max, “US Cyber Command: An Assiduous Actor, Not a Warmongering Bully”, *The Cipher Brief*, (March 4, 2018). Retrieved from: https://www.thecipherbrief.com/us-cyber-command-assiduous-actor-not-warmongering-bully.

Thomas, Bindiya, “UAE Military to Set Up Cyber Command”, (2014, September 30), *DefenseWorld*. Retrieved from: http://www.defenseworld.net/news/11185/. UAE_Military_To_Set_Up_Cyber_Command#.WW4nJYjyiUk.

Thyne, Clayon L., “Cheap Signals with Costly Consequences: The Effect of Interstate Relations on Civil War”, *Journal of Conflict Resolution*, 50:6 (2006)937-961.

Tor, Uri, “‘Cumulative Deterrence’ as a New Paradigm for Cyber Deterrence”, *Journal of Strategic Studies*, 40:1-2(2017)92-117.

Treverton, Gregory F., “Framing Compellent Strategies”, *RAND Corporation* (2000). Retrieved from: http://slantchev.ucsd.edu/courses/pdf/treverton-compellence.pdf.

Weaver, Nicholas, and Dan Ellis, “Reflections on Witty: Analyzing the Attacker”, *Security*, 29:3 (2004) 34-37.

Werkhäuser, Nina, “German army launches new cyber command”, *DW*, (April 1, 2017). Retrieved from:http://www.dw.com/en/german-army-launches-new-cyber-command/a-38246517.

William, Brad D., “Meet the scholar challenging the cyber deterrence paradigm”, (July 19, 2017) *The Fifth Domain*. Retrieved from: https://www.fifthdomain.com/home/2017/07/19/meet-the-scholar-challenging-the-cyber-deterrence-paradigm/.

Zetter, Kim, “Everything We Know About Ukraine’s Power Plant Hack”, *Wired*, (20 January 2016). Retrieved from: https://www.wired.com/2016/01/everything-we-know-about-ukraines-power-plant-hack/.

Zetter, Kim, “Inside the Cunning, Unprecedented Hack of Ukraine’s Power Grid”, *Wired*, (3 March 2016). Retrieved from: https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/.

Zetter, Kim, *Countdown to Zero day: Stuxnet and the Launch of the World’s First Digital Weapon*, (New York: Crown Publishing: 2014).

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Understanding and Countering Cyber Coercion

Quentin E. Hodgson
RAND Corporation
Santa Monica, California, United States
qhodgson@rand.org

Abstract: The past decade has seen the rise of cyberspace as a topic of popular, political and scholarly discourse, from the highest reaches of government to the movie screen. States are grappling with how to address the rising tide of cyber threats to their economies, to their citizens' personal information and increasingly to political and social cohesion. States are using cyber capabilities as a tool of statecraft to achieve political objectives. This paper seeks to develop an understanding of how states use cyber capabilities to coerce others for political objectives. Cyber coercion is defined as the use of cyber capabilities to compel an opponent to undertake an action it would not normally wish to perform and avoid an undesirable outcome. The paper seeks to address: how a state can employ cyber capabilities to compel another state (or non-state actor) to accommodate its ambitions; how cyber coercion might take place; and ways that the United States and its partners can recognize, respond to and counter attempts at cyber coercion. The paper examines the use of cyber operations by North Korea and Russia in recent years as part of their broader strategies to exert influence over their neighbours, showing how the context in which such operations occur is critical.

Keywords: *cyber operations, coercion, deterrence*

1. INTRODUCTION

The past decade has seen the rise of cyberspace as a topic of popular, political and scholarly discourse, from the highest reaches of government to the movie screen. Military organizations from the United States to the People's Republic of China

have grappled with issues from how to address threats emanating from cyberspace to how to integrate cyberspace capabilities into military doctrine (US Department of Defense 2015; Stefan-Gady 2015). In this time, the general public has been exposed to a growing body of reporting on cyberspace issues from the hacking of government agencies, hospitals, public transportation systems and beyond. The US Defense Science Board has issued several reports calling into question both the resilience of military systems to cyber threats (Defense Science Board 2013) and outlining how to deter cyber attacks (Defense Science Board 2017). Government authorities worry that adversaries and others may use cyber means to attack critical infrastructure (Rogers 2017), and more recently the prospect of adversaries undermining democratic processes through disinformation campaigns and even outright corruption of electoral processes has come to the fore.

At the same time, the use of cyber capabilities in a variety of contexts to further nation state interests has grown, from the US' purported targeted operations against Iranian nuclear facilities and North Korea's ballistic missile programme to Iranian-attributed campaigns against Western banks and its regional neighbours. This gives rise to questions about how states are using cyber capabilities as yet another tool of statecraft, including to intimidate, coerce and compel others to do their bidding. This paper seeks to address the question of how states have used cyber capabilities to coerce other states or non-state actors either to pursue courses of action they might not otherwise pursue or to refrain from such actions.[1] More importantly, it compares two actors – Russia and North Korea – which have used cyber operations against their neighbours and others to understand the dynamics of cyber coercion and attempt to isolate factors that indicate when cyber coercion may occur.[2] The paper begins with a discussion of cyber coercion and how it fits into broader deterrence and coercion strategies, followed by an examination of examples from Russia and North Korea. The paper will then suggest some ways that countries can seek to prevent or lessen the impact of cyber coercion.

2. WHAT IS CYBER COERCION?

Any discussion of coercion naturally begins with Thomas Schelling's classic writing on the topic, particularly his seminal work *Arms and Influence*. Schelling described two forms of coercion: active coercion, or compellence, and passive coercion, or deterrence (Schelling 1966). The former involves the active use of force in some form to compel action by another, while the latter involves the threatened use of force to motivate an action or restraint from an action. In reality, the distinction is more of a

1 The focus of this paper will be on state-to-state interactions, but the author acknowledges Travis Sharp's valuable contribution to the literature that a state may seek to coerce a non-state actor and *vice versa*. See Sharp 2017.

2 These case studies are intended to inform a broader research project to develop a framework for cyber coercion that ties into response and defensive actions to thwart attempts to coerce through cyber means.

continuum, as some states may combine compellence actions with the threat of more devastating consequences to accomplish their ends. The literature has often focused on the use of force by states, not necessarily because these concepts do not apply to other actors, but rather because the motivation for examining these concepts in the 20th century was to understand the nature of state-to-state relations. As one author recently noted, scholars have often used analogies to more localized conflicts, such as Schelling's reference to teenager hot-rodding and Robert Jervis's reference to village stag-hunting (Sharp 2017).

In recent years, popular, political and academic discourse has tried to find appropriate analogies or comparable historical instances from other domains to explain cyberspace operations, to clarify the concepts of deterrence, or to distinguish cyberspace from everything else (Nye 2011). This paper begins with the premise that, although cyberspace is indeed a man-made domain, its characteristics are more a matter of distinction rather than fundamental difference from other domains when it comes to international relations. States will seek to use cyber capabilities as one tool of statecraft, just as they seek to use other forms of military force, economic power or social and humanitarian influence to further their interests. The same applies to the use of cyber capabilities as a means to exert influence or pressure on others to shape behaviour, deter adverse actions and even compel another actor (either another state, a multinational organization, or even a single individual). As one scholar has noted, coercion is "the use of threatened force, including the limited use of actual force to back up the threat, to induce an adversary to behave differently than it otherwise would" (Johnson, Mueller and Taft 2003). This definition does not require a certain level of force, so cyber weapons do not have to have the same potential impact as nuclear or even conventional weapons to be credibly used to exert influence, nor does the threatened use of cyber capabilities need to be explicit to have a coercive effect.

Coercion in international relations is not the same as it is with, for example, an abduction, although some of the literature uses formulations that more closely resemble abduction than the dynamics of inter-state relations. This is important for two reasons: 1) context is critical to understanding whether coercion is occurring; and 2) the potential for miscommunication between the coercer and the coerced can be significant, even if there is a long-standing relationship between states, as we shall see in the two case studies in this paper. In an abduction, there is usually an explicit demand for action, whether it is demanding a monetary ransom or some other form of compensation such as the release of political prisoners. The scholarly literature describes a logic for the dynamic between coercer and coerced: "if you do not do X, I will do Y" (Borghard and Lonergan 2017). Another form this takes is when a coercive action or threat "demands clarity in the expected result… [and] be accompanied by some signal of urgency" (Whyte 2016). But in reality, the demands are not always

so clear. The threat actor may not make a clear threat or identify itself explicitly. To express this difference, we can articulate the theoretical ideal and observed practice as follows:

Theory: coercion = *f* (clear threat + actor claims responsibility + explicit desired behaviour)

Observed practice: = *f* (vague threats + implied actor + implicit desired behaviour)

The observed practice is not always a combination of all three; it could involve a clear attribution and explicit desired behaviour, but the threat could be vague. This reality complicates the ability to understand when a state is seeking to coerce another and take steps to counteract or blunt the threat. This paper will return to the differences between theory and observed practice shortly to address whether cyber coercion is successful.

The coercer and coerced may not perceive the messages in the same way (Jervis 1976). Some scholars have noted that cyber coercion is less likely to achieve objectives because the coercive message will signal the threat and allow the coerced to respond or to defend itself, reducing the effectiveness of the coercive measure (Gartzke 2013), but these conclusions are based on a couple of assumptions that do not hold up under scrutiny. Their first assumption is that the coercive measure will be explicit and specific enough to provide the coerced the opportunity to pre-empt the action or prepare its defences. But this is rarely the case, and growing vulnerability to cyber attacks, particularly in more technologically advanced societies, means that the prospective attack surface is so large that adequate preparation is unlikely. The US government, for example, has focused on the protection of critical infrastructure from cyber attack for more than 20 years, starting when President Bill Clinton's Commission on Critical Infrastructure Protection issued its report in 1997 (President's Commission on Critical Infrastructure Protection 1997). The insecurity of critical infrastructure has grown, not diminished, since then.

Their second assumption is that the coercer will signal the means they will use to threaten an opponent. Coercion, however, does not have to state the exact means that will be employed to be credible. The coerced merely has to believe that the coercer has the capability to inflict harm; they do not need to specify "and I will do so with my cyber armies". States are aware of their opponents' capabilities, or they become aware of them over time, and can intuit the potential outcomes. For example, it is highly likely that most states and relevant non-state actors have very little real insight into US cyber capabilities, and in fact may have an inflated picture based on Hollywood movies and the stature of the US civilian technology sector. Couple that with the public

belief that the United States probably employed these capabilities to attack both the Iranian nuclear programme and the North Korean ballistic missile programme, and we can see that the actual capabilities that the United States possesses are less important than the perception of them[3] (Sanger and Broad 2017).

This paper is not advancing an argument about the likely success of cyber coercion; several scholars have addressed its apparently low rate of success (Jensen, Valeriano and Maness n.d.; Borghard and Lonergan 2017). A successful attempt at cyber coercion should result from a combination of a successful cyber operation, in which the targeted system or network was disrupted, with a change in behaviour by the coerced. Even in cases where the operation itself achieves its aims, it appears that behavioural changes are few, whether because the actor carrying out the operation overestimated the impact or underestimated the capacity of the adversary to withstand pain. Despite this poor track record, however, states persist in developing cyber capabilities and appear to believe, rightly or wrongly, that the promise of cyber coercion exists. Therefore, we can expect states to continue to pursue coercive actions through cyberspace.

3. NORTH KOREA

Of any state, North Korea is arguably the most likely to employ cyber capabilities as part of a coercive strategy. Despite broad consensus about the country's technological backwardness,[4] the North Korean regime has shown a remarkable astuteness and dedication in investing in militarily relevant technologies, most prominently in its nuclear and ballistic missile programme, but also in recent years in its cyber capabilities (Ball 2017). North Korea has a long history of coercive action, from shooting down a US spy-plane in the 1960s to shelling off-shore islands and sinking a South Korean naval vessel in 2010 (Terry 2013). For North Korea, these actions have largely paid off, resulting in concessions and economic aid from South Korea and the United States as often as more economic sanctions. Sharp (2017) has argued that the North Korean attack on Sony Pictures Entertainment in 2014 was a form of cyber coercion aimed at destabilizing Sony's leadership, imposing costs and seeking to retaliate for perceived insults to the regime with the impending release of a comedy film, *The Interview*, the plot of which is focused on an attempt to assassinate the Dear Leader (Sharp 2017).

The case of North Korea's reaction to the film has been the subject of several analyses, but it is worth briefly reviewing the timeline of events and the broader context in which this case occurred. The proximate cause of the events was the impending release of the film and the North Koreans' strong objections to it. As early as June

3 One could argue that this is one area where cyber weapons and nuclear weapons are more alike. The United States has not used a nuclear weapon in conflict since 1945 and has not conducted a nuclear test since 1992, but few states if any are likely to doubt the US nuclear arsenal's size or capabilities.

4 Including reportedly only 28 registered websites. See http://www.bbc.com/news/world-asia-37426725.

2014, the North Korean government condemned *The Interview* in a Foreign Ministry statement and subsequently sent a letter to the UN Secretary General accusing the United States of terrorism and an act of war (Brzeski 2014). After postponing release of the film until December, Sony received emailed demands for money from a group calling itself God'sApstls, followed by a malware attack that resulted in corruption of the master boot records of numerous computers, rendering them inoperable. A group called Guardians of Peace claimed responsibility for the attack and began releasing embarrassing emails and yet-to-be released films in the Sony library (Roman 2014). This was followed by threats of violence against movie theatres and "doxing" of Sony executives through release of internal documents that showed them in a bad light. The North Korean government denied responsibility for the attacks or the threats but referred to the acts as "righteous deed[s]" and speculated that "supporters and sympathizers" of the North Korean regime were involved (Reuters 2014). Sony pulled the movie from theatres, but later reversed its decision after coming under criticism, including from the President of the United States, for appearing to capitulate to threats.

It is important to take a moment to reflect on this point, since if we take the critiques of cyber coercion to heart, the fact that the North denied its involvement would appear to undermine the argument that it was intended as a coercive measure. But the timing of this case is important, as is the context. North Korea clearly indicated its displeasure with the film for several months prior to the events. In the summer, the North Korean Foreign Ministry said "[if] the US administration connives at and patronizes the screening of the film, it will invite a strong and merciless countermeasure" (Brzeski 2014). Totalitarian regimes often fail to understand how western countries operate and conduct their own mirror-imaging. North Korea could very well have believed that *The Interview* was part of an official US government propaganda campaign against the regime. North Korea has a long history of strong rhetoric, but it has also shown itself willing to use force of various kinds with little compunction, whether through directly attacking military targets like soldiers along the Demilitarized Zone and South Korean naval vessels, or civilian targets in the South. From North Korea's perspective, it possibly felt it had conveyed its message clearly and publicly through official channels. The fact that it chose to then follow up on its (failed) coercive rhetoric with cyber attacks through proxies does not detract from the original intent of the threats. The first phase of coercion, which did not explicitly state the form in which subsequent pain would be inflicted, simply failed to achieve the desired outcome of stopping the film, so the North had to escalate from threats to action. At that point, the North Koreans were transitioning from the threat of consequences to seeking to impose those consequences, and who delivered the effects is less important. At the same time, US officials noted that they were not clear on how the threat against the movie theatres was intended to be carried out, which nevertheless did not deter them from treating it as a serious threat (Sharp 2017).

Whether the North Koreans truly believed that the use of proxy fronts (likely for the Reconnaissance General Bureau and the Korean People's Army) would obfuscate the origins of the threats is an interesting question, but currently unanswerable. If the North Koreans had sought to hide their direct involvement, then it is questionable whether it would contribute to the credibility of future coercive threats. That said, North Korean has routinely denied physical attacks, such as the sinking of the South Korean naval vessel *Cheonan* in 2010, when no other credible perpetrators have presented themselves (Terry 2013). It is conceivable that North Korean denies its involvement as a *pro forma* matter as opposed to truly seeking to avoid blame. It also plays to their domestic audience, for whom the regime has to portray itself constantly as the victim rather than the aggressor. Sharp concludes that, while not necessarily achieving all of its aims, the Sony case shows a successful use of cyber capabilities, coupling cyber exploitation (stealing data) with offensive cyber capability to disable computers, coerce Sony's leadership and even lead to the downfall of several senior leaders there (Sharp 2017). Whether the coercive actions were intended to shape other actors is unclear, but North Korea has not limited itself to using cyber to attack private companies. In recent years, it has also employed cyber operations as part of its coercive campaign against the Republic of Korea. Suspected North Korean cyber operations against the South have included targeting the financial, media and energy sectors, as well as government agencies. In some cases, including the attack on a virtual currency exchange in Seoul in May 2017, financial interests may have been the stronger motivation (Perper 2017). The 2013 attacks against South Korean television stations, a bank and bank machines, however, may have been part of an escalatory exchange following a two-day Internet outage in the North (Branigan 2013). These cases are less clearly overt acts of attempted coercion, but they show a willingness to engage in a cyber tit-for-tat and to inflict damage on the South.

North Korea's cyber capabilities are not exclusively retaliatory, nor does the regime likely see them as a replacement for other forms of coercion (Jun, LaFoy and Sohn December 2015). The nuclear and missile programmes are probably still seen as guarantors of regime survival, but cyber capabilities provide a flexible new tool to achieve a variety of ends: theft to improve the regime's finances, espionage and the ability to threaten and inflict pain and damage on its adversaries. The recent cyber events also establish a track record of use that could play a role in future coercive scenarios. Returning to the theoretical construct for coercion (coercion = f (clear threat + attribution + explicit desired behaviour)) we can code the cases as follows:

Case	Threat	Threat Actor Responsible	Desired Behaviour
Sony	Ambiguous	Disputed attribution, but likely North Korea	Clear
South Korea television and banks	Ambiguous	Disputed attribution, but likely North Korea	Unclear

4. RUSSIA AND UKRAINE

Russian cyber activity has gained in prominence, beginning with the denial of service attacks against large segments of the Estonian economy and government in 2007 and as part of the conflict with Georgia in 2008, which some sources have attributed to the Russian government or to patriotic hackers acting on the government's behalf (Davis 2008; Hollis 2010). More recently, the focus has turned to Russian disinformation campaigns and alleged interference in elections in the United States, Germany and France, among others (FireEye January 2017). Russian actors, some more closely affiliated with the government and others playing a more ambiguous role, have established online personas on multiple Internet platforms, including Twitter and Facebook, to disseminate falsified news stories and develop narratives sympathetic to Russia's views (Coats 2017). In the midst of such campaigns, it appears that Russia has also started to use cyber capabilities as a coercive tool. Here we will focus on Russian activity in Ukraine, but this is not intended to downplay or diminish Russian activity in other countries. It is also important to acknowledge that Russian disinformation campaigns, although not the focus of this analysis, could very well be coercive measures intended to destabilize its neighbours and seek to either promote more pro-Russian parties and social movements or motivate current elites to accommodate Russian demands.

The dynamics of Russian-Ukrainian relations are complex and long-standing, which underscores the importance of understanding the context in which the events of recent years have occurred. The Russians have historically seen Ukraine as a part of the border region of Russian territory, rather than as a separate geographic and political entity (in Russian, Ukraine roughly means "on the border"). The Russian military campaign in 2014 to seize Crimea was seen domestically more as a means to correct a quirk of history than an invasion, as Crimea was a gift to Ukraine during Nikita Khrushchev's tenure as leader of the Soviet Union (McCauley 1993). The Crimea also serves as the home port for the Russian Navy's Black Sea Fleet, which makes it strategically important for Russia. Russia's apparent actions to destabilize Ukraine through various means, including cyber operations, supporting proxy fighters and sending military forces into Eastern Ukraine, stem from a desire to maintain Ukraine in Russia's orbit and prevent further integration with the West (Treisman 2016). Ukraine's negotiations in 2013 to conclude a political and trade deal with the European Union also threatened to put Ukraine more squarely in the West's camp.

After then-President Viktor Yanukovich reversed course, protests erupted in Kiev. Police moved in to confront the protesters and violence ensued, resulting in dozens of deaths (Applebaum 2017). In the aftermath of these protests, pro-Russian groups in Eastern Ukraine began to seize control of government institutions, prompting the

government to respond militarily. Following the change of President in May 2014, fighting continued and, despite a negotiated ceasefire in February 2015, the conflict continued throughout the year.

In the midst of the horrific fighting and civilian suffering, particularly in Eastern Ukraine, the country suffered the first significant cyber attack on its electric grid in December 2015. The attack affected approximately 250,000 customers for some hours, but appeared to cause no lasting damage despite targeting the Supervisory Control and Data Acquisition (SCADA) controllers in addition to business system workstations and servers (SANS Institute 2016). The malware employed was a set of tools including the BlackEnergy Trojan and the KillDisk eraser and targeted at least three geographically dispersed regional power sub-stations (Greenberg 2017). The impact on the energy sector received the most attention, particularly coming during the winter, but the cyber attacks against Ukraine had also impacted other sectors including media, finance and transportation in the preceding months. Security researchers have attributed the BlackEnergy tool and the actions in Ukraine to the Sandworm intrusion set, which many believe is a Russian hacker group (Hultquist 2016). The Ukrainian government has been more explicit in tying this activity to Russian security services. Attacks on various sectors continued in 2016, including another attack on the energy sector almost exactly a year after the December 2015 attacks that hit the Kiev transmission station; this time the outage lasted barely an hour.

The Russian government has not claimed responsibility for these cyber attacks and routinely denies involvement in cyber operations against other countries, reminding audiences of evidence that the United States in particular has engaged in the widespread use of cyber operations (Russian Ministry of Foreign Affairs 2016). The Russian government did not appear to make explicit demands of the Ukrainian government or public, either in advance of the attacks or afterwards. In the context of the broader conflict, however, the Russian strategy appears to include: establishing facts on the ground through the manoeuvre of military forces and the use of proxies; spreading disinformation to portray the West and pro-western Ukrainians as enemies of the Ukrainian people; and using cyber operations to reinforce that messaging. Cyber operations in this context appear to be intended to broadly destabilize the political and social cohesion in Ukraine.[5] The ultimate outcome, therefore, is predicated on the Ukrainian government acquiescing to Russian influence on the country and halting its integration with the West. In that sense, the coercion appears focused less on seeking to promote specific actions and more towards shaping Ukrainian behaviour for the long term.

5 There is also speculation that the Russians are using the conflict with Ukraine to 'test' its cyber capabilities in a real-world laboratory as a prelude to potential use against other countries such as the United States. Although this may be a collateral benefit, there is little public evidence to support this as the primary reason.

Russian cyber operations against Ukraine show the importance of understanding the context in which conflict occurs. Analysis that examines cyber operations in isolation will fail to identify the implicit outcomes that the instigator seeks, which often go unstated because the parties already know what they are. It is also evident that the Russians are not looking for the Ukrainians to undertake a single, specific action to forestall future cyber coercion, but that it is conducting a broader campaign to prevent Ukraine's integration with the West. The theoretical framework would therefore appear in this case to be as follows:

Case	Threat	Threat Actor Responsible	Desired Behaviour
Russia-Ukraine	Ambiguous	Disputed attribution, but likely Russia	Somewhat clear

5. WHAT CAN WE DO ABOUT IT?

The North Korean and Russian cases demonstrate that states may indeed be using cyber capabilities to attempt to coerce others, but that the ambiguous nature of these campaigns, with their unclear threats, ambiguous attribution and lack of clarity of desired behaviour, makes it less likely that the coercion will succeed, although that has not appeared to diminish their occurrence. That said, these coercive campaigns are not without cost to the victims, which would indicate that some work is needed to counter or mitigate them. Traditional deterrence theory postulates two primary means for response: a threat of punishment for an action that is credible and (one presumes) unacceptable to the opponent, and denial of gains from an action. Professor Joseph Nye (2016/7) has added to these two by postulating that entanglement and normative taboos can play a role. Addressing the threat of cyber coercion will have to account for these mechanisms, but there are practical difficulties in implementing them that need to be addressed. Before addressing these means, however, we should examine how to recognize that cyber coercion is occurring.

The two case studies presented in this paper highlight two key points when seeking to assess whether cyber coercion is occurring. The first is to recognize that the instigator will not always present explicit demands; there may not be the equivalent of a ransom demand. In many cases of state-on-state conflict, the relationship is long-standing and complex, and therefore the nature of the demands may be more implied than explicitly stated. The Sony Pictures case shows a counter-example, where it appears that the demand was clearly stated: do not release the film. But in that case the second point comes to the fore, that the demand will not state explicitly in all cases the form in which threatened consequences will come. In fact, the Sony case included threats of physical harm to movie theatres that never materialised and may have been intended

to instil fear with no prospect of the threat ever being carried out; of course, US law enforcement authorities could not take that chance and treated the threat seriously. North Korea may have used the subsequent cyber operations as a means to destabilize Sony Pictures' leadership, as one scholar claimed, but it is just as likely that it presented a tangible way for North Korea to inflict pain when other options were not open to it or would have proved too costly.

These considerations give rise to a set of questions to consider in similar circumstances:

- *Does a state's adversary have demonstrated or emerging cyber capabilities that it should track seriously?* This has implications not only in terms of intelligence collection and analysis, but also in challenging basic assumptions. Both Iran's and North Korea's cyber capabilities took Western governments off-guard because they had simply assumed that these countries did not have the technological capabilities.
- *What is the broader context in which conflict is developing?* Thinking about a country's cyber capabilities in isolation risks missing emerging signals that a coercive campaign is beginning or potentially entering a new phase where cyber operations could occur.
- *Does the coercer have long-standing demands?* Identifying these contributes to understanding what potential outcomes the coercer may seek and could assist in anticipating potential cyber coercive actions.

The threat to impose costs on others for using cyber capabilities has not prevented state use of cyber, though it is impossible to prove an assertion that perhaps current US and Western policies have prevented more egregious actions. It is far more likely that countries such as Russia or North Korea see little reason to fear retaliation at apparently low thresholds of cyber use because the consequences have been spread out over time and have not resulted in loss of life or significant damage to property that would normally invite such a response. The case studies in this paper indicate, however, that there is significant ambiguity around coercive actions using cyber capabilities, which complicates a state's response. States that may be subject to cyber coercion will have to carefully examine the circumstances in which they perceive threats and determine whether a lower threshold for response or even pre-emption may be required. This carries escalation risks, of course, and could even lead to action against an entirely innocent state (at least in the particular situation evaluated).

Given the broad attack surface and the thousands, if not millions, of targets that present themselves in cyberspace, denial of an adversaries' objectives seems an impossible task. The US government identifies 16 critical infrastructure sectors (with "elections" being an ambiguous addition in 2016) that encompass some 1,000,000 owners and operators. Even if a small portion of these are truly critical, such as the list of entities

deemed at greatest risk and potentially causing greatest harm (the so-called "section 9" list, referring to the Obama administration's cyber security executive order which was adopted by the Trump administration in its first cyber security executive order), adequately defending them against a vast array of threats is no easy task. That being said, there is evidence that states seeking to coerce others underestimate their capacity to endure pain, and therefore improving resiliency (as opposed to simple defence) is likely a vital component of a counter-coercion strategy (Jensen, Valeriano and Maness n.d.).

In Nye's formulation, entanglement "refers to the existence of various interdependences that make a successful attack simultaneously impose serious costs on the attacker as well as the victim" (Nye 2016, p. 58). It is possible that this consideration has influenced states such as Russia and China to pursue a less integrated Internet; indeed, Russia has announced plans to create its own form of domain name system to undo its entanglement with the United States (Tucker 2017). Given that, this approach may be useful as a supporting line of effort but is unlikely to prove decisive.

Finally, norms of state behaviour were a central thrust of the Obama administration's work in the UN Group of Governmental Experts (UN GGE) and in its bilateral discussions with the Russian and Chinese governments (Finnemore 2017). For a period of time, this path seemed to have achieved some success, with consensus reports emerging over several years. However, the 2016-2017 UN GGE group failed to achieve consensus and concluded its work without a report on which the 25 participating countries could agree (Korzak 2017). Of course, the establishment of norms as statements of principle are only the first step. Much like customary law, norms gain stature as nations demonstrate through their actions that they are adhering to these norms. The failure of the UN GGE does not in itself signal the death of cyber norms; it simply highlights the challenge of gaining consensus on these issues in a diverse group of countries that do not all necessarily trust each other to negotiate in good faith.

Each of these four proposed approaches has a role to play, but clearly there is no miracle cure that addresses the potential for states to use cyber capabilities to threaten and coerce those whom they seek to bend to their will. The first step is to develop the ability to recognize when cyber coercion could come to pass and seek to head it off, including with explicit warnings and leveraging the four methods Professor Nye identified, with particular focus on improving resiliency in the face of cyber threats.

6. CONCLUSION

This paper has argued that cyber capabilities can indeed be used as coercive tools of statecraft, but recognizing when they may be used and how a state can reduce their impact is no easy task. The context in which cyber coercion may occur is important, as are the capabilities that a state may develop. The increasing commodification of cyber attack tools, the growing legitimate, grey and black markets for these tools and the increasing attack surface all make cyber coercion an increasingly attractive tool for states.

The case studies presented here demonstrate that cyber coercion often occurs in contexts of significant ambiguity. The threat actor may not make an explicit threat, may choose to work through proxies or deny involvement outright, or the desired behaviour may not be clearly stated. In the case of North Korea's attack on Sony, there were vague threats at the beginning from the North Korean government, followed by more specific threats from an apparent proxy. The desired outcome was clear from the beginning, although the coercive campaign ultimately failed to prevent the release of the film. In the denial of service attacks on South Korean television and banking, there was no specific threat, nor a clear claim of responsibility in the immediate aftermath. Indeed, the North Korean government never made a specific demand of the South, but a broader examination of North Korean behaviour over decades indicates that the threats and desired response are long-standing and understood. Similarly, in the Russia-Ukraine context, Russian cyber actors are not explicitly tied to the Russian government, although many observers believe they are at least loosely linked. The desired outcome – Ukraine's drawing back from Western integration and remaining in Russia's orbit – is long-standing. In each of these examples, cyber capabilities appear to have played a role in a broader strategy. Examining them as stand-alone cases misses the broader context in which cyber capabilities are used. More work is needed to develop this context for states of concern to detect, respond and mitigate the effects of cyber coercion.

ACKNOWLEDGMENTS

I would like to acknowledge the work of former MITRE colleagues Peter Sheingold, Cynthia Wright and Mark Peters on grey zone operations that led me to start to explore the concept of cyber coercion. I also want to acknowledge John Parachini, Laura Baldwin, Cynthia Dion-Schwarz and Sina Beaghley for encouraging me to pursue this project and providing support. My thanks to three anonymous reviewers whose comments helped improve the paper immensely. This paper is dedicated in memory

of Shawn Brimley, one of the finest national security professionals it has been my privilege to work with and know.

REFERENCES

Applebaum, Anne. 2017. "Why does Putin want to control Ukraine? Ask Stalin." October 20. Accessed January 6, 2018. https://www.washingtonpost.com/outlook/why-does-putin-want-control-ukraine-ask-stalin/2017/10/20/800a7afe-b427-11e7-a908-a3470754bbb9_story.html?utm_term=.9fb81.

Ball, Tom. 2017. "Crowdstrike CTO: Theft and destruction are 'just a few keystrokes' apart." *Computer Business Review*. September 29. Accessed December 29, 2017. https://www.cbronline.com/news/cybersecurity/crowdstrike-cto-theft-destruction-just-keystrokes-apart/.

Branigan, Tania. 2013. "South Korea on Alert for Cyber Attacks after Major Network Goes Down." November 20. Accessed January 6, 2018. https://www.theguardian.com/world/2013/mar/20/south-korea-under-cyber-attack.

Broad, William J., and David E. Sanger. 2017. "Trump Inherits a Secret Cyberwar Against North Korean Missiles." *The New York Times*, March 5: A1.

Brzeski, Patrick. 2014. "North Korea Files Complaint With United Nations Over 'The Interview'." *Hollywood Reporter*. July 11. Accessed December 29, 2017. https://www.hollywoodreporter.com/news/north-korea-files-complaint-united-717943.

Coats, Dan. 2017. "Worldwide Threat Assessment of the Intelligence Community." Washington, DC: Director of National Intelligence, May 11.

Davis, Joshua. 2008. "Hackers Take Down the Most Wired Country in Europe." August 21. https://www.wired.com/2007/08/ff-estonia/.

Defense Science Board. 2017. *Cyber Deterrence*. Task Force, Washington, DC: US Department of Defense.

Defense Science Board. 2013. *Resilient Military Systems and the Advanced Cyber Threat*. Task Force, Washington, DC: US Department of Defense.

Elkind, Peter. 2015. "Inside the Hack of the Century." June 25. Accessed January 4, 2018. http://fortune.com/sony-hack-part-1/.

Finnemore, Martha. 2017. *Cybersecurity and the Concept of Cyber Norms*. November 30. Accessed December 2, 2017. http://carnegieendowment.org/2017/11/30/cybersecurity-and-concept-of-norms-pub-74870.

FireEye. January 2017. APT 28: *At the Center of the Storm*. Special Report, FireEye iSight Intelligence.

Gartzke, Erik. 2013. "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth." *International Security* 38(2): 41-73.

Greenberg, Andy. 2017. "How an Entire Nation Became Russia's Test Lab for Cyberwar." June 20. Accessed July 6, 2017. https://www.wired.com/story/russian-hackers-attack-ukraine/.

Hollis, Davis. 2010. "Cyberwar Case Study: Georgia 2008." *Small Wars Journal*.

Hultquist, John. 2016. *Sandworm Team and the Ukrainian Power Authority Attacks*. January 7. Accessed January 5, 2018. https://www.fireeye.com/blog/threat-research/2016/01/ukraine-and-sandworm-team.html.

Jensen, Benjamin M., Brandon Valeriano and Ryan C. Maness. n.d. "Cyber Compellence: Applying Coercion in the Information Age." http://www.brandonvaleriano.com/uploads/8/1/7/3/81735138/cyber_victory.pdf.

Johnson, David E., Karl P. Mueller and William H. Taft. 2003. "Conventional Coercion Across the Spectrum of Operations: The Utility of U.S. Military Forces in the Emerging Security Environment." RAND Corporation, Santa Monica, CA.

Jun, Jenny, Scott LaFoy and Ethan Sohn. December 2015. *North Korea's Cyber Operations: Strategy and Responses*. A Report of the CSIS Korea Chair, Washington, DC: Center for Strategic & International Studies.

Korzak, Elaine. 2017. "UN GGE on Cybersecurity: The End of an Era?" July 31. Accessed September 15, 2017. https://thediplomat.com/2017/07/un-gge-on-cybersecurity-have-china-and-russia-just-made-cyberspace-less-safe/.

McCauley, Martin. 1993. *The Soviet Union 1917-1991*. London: Longman.

Nye, Joseph S. Jr. 2016/2017. "Deterrence and Dissuasion in Cyberspace." *International Security* 41(3): 44-71.

Nye, Joseph S. Jr. 2011. "Nuclear Lessons for Cyber Security?" *Strategic Studies Quarterly* 5(4): 18-38.

Perper, Rosie. 2017. "North Korea may be behind a massive cyber attack on a South Korean bitcoin exchange that caused it to collapse." December 21. Accessed January 6, 2018. http://www.businessinsider.com/north-korea-south-korea-bitcoin-heist-2017-12.

President's Commission on Critical Infrastructure Protection. 1997. "Critical Foundations: Protecting America's Infrastructures." Washington, DC.

Rogers, Michael S. 2017. "Statement Before the Senate Committee on Armed Services." May 9. https://www.armed-services.senate.gov/imo/media/doc/Rogers_05-09-17.pdf.

Roman, Jeffrey. 2014. "Sony Pictures Cyber-Attack Timeline." December 23. Accessed December 30, 2017. https://www.bankinfosecurity.com/sony-pictures-cyber-attack-timeline-a-7710.

Russian Ministry of Foreign Affairs. 2016. *Comment by Foreign Ministry Spokesperson Maria Zakharova on new threats of sanctions from the United States*. December 28. Accessed January 6, 2018. http://www.mid.ru/en/foreign_policy/news/-/asset_publisher/cKNonkJE02Bw/content/id/2581641.

SANS Institute. 2016. *Confirmation of a Coordinated Attack on the Ukrainian Power Grid*. January 9. Accessed January 5, 2018. https://ics.sans.org/blog/2016/01/09/confirmation-of-a-coordinated-attack-on-the-ukrainian-power-grid.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Sharp, Travis. 2017. "Theorizing cyber coercion: The 2014 North Korean operation against Sony." *Journal of Strategic Studies* 40(7): 898-926.

Stefan-Gady, Franz. 2015. "China to Embrace New Active Defense Strategy." May 26. Accessed December 29, 2017. https://thediplomat.com/2015/05/china-to-embrace-new-active-defense-strategy/.

Terry, Sue Mi. 2013. "North Korea's Strategic Goals and Policy towards the United States and South Korea." *International Journal of Korean Studies* 17(2): 63-92.

Treisman, Daniel. 2016. "Why Putin Took Crimea: The Gambler in the Kremlin." April 18. Accessed January 6, 2018. https://www.foreignaffairs.com/articles/ukraine/2016-04-18/why-putin-took-crimea.

Tucker, Patrick. 2017. "Russia Will Build Its Own Internet Directory, Citing US Information Warfare." November 28. Accessed November 29, 2017. http://www.defenseone.com/technology/2017/11/russia-will-build-its-own-internet-directory-citing-us-information-warfare/142822/.

US Department of Defense. 2015. *DoD Cyber Strategy*. Washington, DC: US Department of Defense.

Whyte, Christopher. 2016. “Ending cyber coercion: Computer network attack, exploitation and the case of North Korea.” *Comparative Strategy* 35(2): 93-102.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Targeting Technology: Mapping Military Offensive Network Operations

Daniel Moore
Department of War Studies
King's College London
London, United Kingdom
daniel.d.moore@kcl.ac.uk

Abstract: State-sponsored network intrusions are publicly and frequently exposed but assessing how militaries conduct offensive network operations remains difficult. Data can be transmitted near-instantaneously, yet cyber-attacks can take months or even years to mature, complicating attempts to integrate them into joint operations. What challenges, risks, opportunities and advantages are associated with attacking networks? This paper argues that military offensive network operations can be usefully cast into a two-part taxonomy: event-based attacks and presence-based attacks. These are then applied to practical use-cases drawn from existing strategies, case studies and current military platforms. Event-based operations include all instances in which the target is directly and in real time attacked by compromise of its software and may appear roughly analogous to physical weapons. Presence-based operations include all network intrusions in which the attackers traverse compromised networks until targets are located, assessed, and weaponized for later activation, more analogous to a clandestine sabotage operation. Distinguishing between these two types is crucial; they offer different solutions, encompass varying risks, and may require different resources to accomplish. Event-based attacks can offer a tactical advantage against a single adversary platform or network. A successful presence-based operation may result in a strategic advantage against a stronger force. Each of the two operation types is broken into phases as defined by the US Department of Defense Common Cyber Threat Framework. The model envisions four steps in the network operation life-cycle: preparation, engagement, presence and effect. By anchoring the assessment using the framework, the unique characteristics of both operation types become easier to analyze.

Keywords: *cyber warfare, network operations, cyber attacks, offensive cyber*

1. INTRODUCTION

Military use of offensive network operations (MONOs) epitomizes the desire for cleaner, quicker, and less violent conflict. If strategic adversary coercion can be achieved by targeting the digital infrastructure used for both national security needs and daily life, enemy resolve should theoretically decrease to the point of surrender. This is an understandably appealing concept, but not entirely accurate. Network operations can assist both tactical and strategic combat efforts if all their particular advantages and disadvantages are accounted for. While nations occasionally release slivers of information on how they employ offensive network capabilities, doctrine and strategy remain understandably murky on how operational success is achieved in and through networks.

At the core of this work is the argument that *MONOs can broadly be grouped into two classes;* ***presence-based*** *and* ***event-based***. Presence-based operations are offensive network activities which include a lengthy intrusion component meant to establish a persistent presence within adversary assets, and then traverse networks and locate objectives. Event-based operations primarily include direct attacks intended to cause immediate effect against a targeted platform. Many of the currently known state-sponsored network attacks would fall into the former category, while many network attacks against military hardware and software in the battlefield would fit the latter. All can be carried out for military objectives.

A typology for network warfare matters. When all offensive operations are assessed together, the results often seem muddled and difficult to translate to military doctrine.[1] Examined separately, presence- and event-based operations are shown to have distinctive characteristics embodying unique advantages and disadvantages. They require different manpower, resources and operational approaches, and can be applied against different targets for different effects. Some may be more easily relegated to battlefield use, while others are best kept for strategic maneuvers. Activating a presence-based operation may entail losing a crucial source of intelligence, while event-based attacks are inherently suitable for recurring military use. By identifying the parameters under which an operation or capability can be relegated to each of the categories, it empowers decision-makers to "release" some capabilities to battlefield commanders, while retaining sensitive measures within the higher echelons.

Event-based operations are roughly analogous to firing a weapon. When such an attack is launched, virtual ordnance traverses one or more networks, where it connects with the adversary's defenses. Impact on the target – if successful – is immediate or near-immediate. They are meant to be reusable, and the attack may be launched by a local fire team, a warfighting platform or from remote territory. These types of attacks – like

1 For an example of the deliberations around these challenges, see Atkin, McLaughlin, and Moore (2016).

their kinetic counterparts – often have localized effects meant to augment or support kinetic strikes (US Army 2014, 31). They may disrupt an aircraft's onboard systems, degrade radar functionality or impair a regional network by way of a destructive worm that wipes endpoints and servers. As a corollary, such tactical network warfare works well in a combined arms package, jointly deployed alongside kinetic capabilities.

Presence-based operations are roughly analogous to clandestine sabotage operations. A precursor successful intelligence operation results in sustained remote access to an adversary's networks. From that point, attacker assets are maneuvered to enumerate servers and endpoints, gathering information and identifying weak points that may subsequently be attacked for effect. Specialized implants are fielded where needed, with the intent to activate when the order to do so arrives. This can manifest as a multi-year intrusion campaign into an adversary's command and control network, logistics framework or critical infrastructure. The potential risks to friendly weapons and capabilities of discovery are far greater due to the extended presence "behind enemy lines", as is the chance of failure. But the potential benefit is commensurately immense, possibly resulting in an advantage of strategic proportions. These operations may serve as the surprise prelude to an offensive campaign or as a means of exerting pressure on adversary governments.

This article offers an in-depth analysis of MONOs for both event-based and presence-based attacks. The model chosen as the theoretical scaffolding is the US Department of Defense's *Common Cyber Threat Framework* (US DNI 2013), which capably aggregates different industry and public-sector models to provide a useful approach towards assessing wider network campaigns rather than focusing on individual intrusions. The four primary phases presented in the Common Cyber Threat Framework – *preparation, engagement, presence, and effect* – are assessed for both presence and event-based operations.

While official information on MONOs is scarce, this does not imply a dearth of sources. The increasing tenacity of the information security industry in unravelling nation-state cyber capabilities provides a useful window into well-resourced network operations. Industry network defenders working to deconstruct organized adversaries have generated useful analytical models such as Lockheed Martin's Cyber Kill Chain (Hutchins, Cloppert, and Amin 2011) and the Diamond Model (Caltagirone, Pendergast, and Betz 2013). Official publications do indeed exist, and include tactical accounts of how units operate on the field (Kimmons 2017), joint publications on doctrine (US Joint Chief of Staff 2013), strategic guidelines (Chairman of the Joint Chiefs of Staff 2006), oversight reports (US DoD 2017) and even operational integration roadmaps (US DoD 2003). Although employed cautiously, even leaks

of highly-classified materials from network operations units such as the NSA[2] can contribute information on context and capabilities.

Military network operations do not exist in a vacuum. In contrast to some existing models, they do not begin with target reconnaissance and do not end after activating offensive payloads (Hutchins, Cloppert, and Amin 2011, 4–5). There are several strategic and tactical phases predicating the operation itself, and several that follow it. Similarly, there are processes that run concurrently to the network intrusion, interacting with work carried out by network operators to facilitate their success and feeding off it. These additional components are not peripheral; they are instrumental to an operation's success and are an integral part of understanding offensive military capabilities in cyberspace.

Some limitations accompany the scope of this work. Firstly, while the sources and case studies below are not limited to the US, they do favor them significantly due to their relative availability. Secondly, the proposed distinction is meant as a useful generalization for the allocation of resources and division of forces rather than a catch-all classification. Some niche cases may not fall neatly within one category or the other, and some attacks may present elements of both, such as a presence-based intrusion which is then used to launch subsequent event-based attacks.

2. PREPARATION

Preparation encompasses all efforts preceding contact with the enemy. The Cyber Threat Framework defines preparation as all collective efforts to identify targets, develop capabilities, assess victim vulnerability and define the scope of the operation (US DNI 2013, 2). Each of these processes reflects months and perhaps years of investment in resources, both material and operational. Thus, while it is the least discussed, the preparation phase of any offensive network operation may often be its longest.

Before operators first interact with adversary networks, planners must first initiate a *targeting* cycle. This may seem deceptively trivial; an actor seeking to target an adversary will simply pursue its networks. In reality, locating, identifying and enumerating relevant networks for attack can be difficult (Monte 2015, 20). Modern militaries employ dozens of disparate networks even within a single organizational entity (Burbank et al. 2006, 39–42). Identifying which to attack is no negligible feat. It requires in-depth intelligence and an understanding of the adversary's order of battle. In many cases, sensitive or operational networks do not interface directly

2 There were at least three separate leaks in the US alone. These include NSA leaks by former contractor Edward Snowden and by a group calling itself the Shadow Brokers in 2016, and a purported CIA leak in 2017, see Wikileaks (2017, 7).

with the Internet or perhaps even with any other networks.[3] This makes the notion of identifying them and securing access that much harder. The force commander will choose to pursue a target through networks only if it is deemed to be the most effective means of attaining the objective (Ducheine and van Haaster 2014, 313–14).

Targeting cycles are decidedly different for presence and event-based operations. Targeting for presence-based operations is most commonly conducted by the strategic intelligence entities that have network intrusion capabilities. Traditionally, it is within the remit of signals intelligence (SIGINT) organizations, which in varying jurisdictions are either civilian or military.[4] As such, it is often a derivative component of those entities' prioritized intelligence requirements (PIRs). PIRs form a fundamental national security agenda towards which agencies are expected to work, whether by collecting intelligence or preparing for eventual network attacks (US DoD 2013, 24–25). Targeting is therefore a long-term process in which intelligence on the adversary is accumulated, increasingly providing information required to properly prioritize between networks by balancing feasibility and relevance to the objectives at hand. The result is a highly curated list of specific targets.

Targeting for event-based operations would reasonably take place in proximity to the attack itself (Conti and Raymond 2017, 181–82). As a result, this cycle could commonly be conducted by the theatre force commander, or perhaps even a tactical unit lead against a limited objective. This, alongside the employment of pre-packaged network capabilities, entails that the decision-making process is both faster and conducted with far available resources. In order to identify which networks should be selected for subsequent engagement, the commander must identify the adversary's local centers of gravity which, if compromised, would reduce enemy effectiveness. To accomplish this, reconnaissance assets conducting spectrum analysis and automated network mapping procedures may identify adversary networks in the region, possibly even auto-assigning ordnance against them.

Some targets may be chosen for both event and presence-based operations, reflecting varying goals and opportunities. Over the last two decades, the United States has gradually modernized battlefield connectivity for its deployed forces. A part of this process, titled Warfighter Information Network – Tactical, or WIN-T, is a prime example of how saturated the network landscape can be. A combination of dedicated line-of-sight radios and satellite-communication terminals (Coile 2009, 5) services a host of networks including the general-purpose NIPRNet, SIPRNet[5], and local compartmentalized data and voice networks (Epperson 2014). Many of these

3 The idea of separating a network from all other networks is called "air-gapping" and is a widely accepted methodology of reducing a network's potential attack surface.

4 In the United States, the NSA is a civilian agency. In the Israeli example, it is military unit 8200.

5 Non-Secure Internet Protocol Router Network (NIPRNet) and Secure Internet Protocol Router Network (SIPRNet): US Department of Defense networks used for unclassified and classified communications between and within partner organizations.

networks enable unclassified, ancillary functions that are not mission critical. Others carry sensitive targeting information, communications or intelligence data. Some of these networks may be inaccessible as they are transmitted over a medium to which the attacker has little hope of gaining access. Others rely on commercial satellites and even the Internet as their transmission medium. Completing the targeting process by successfully classifying which networks both matter and are pragmatically reachable is therefore a challenge. In some cases, these networks may be subjected to long-term compromise in the form of a presence-based operation. In other cases, locally accessible datalinks such as a regional network cell might be the target of an event-based attack. Interestingly, the WIN-T project has now been officially terminated by the US military, citing concerns that the project's architecture is indeed too vulnerable to a determined, well-resourced adversary (Crawford, Mingus, and Martin 2017, 6–8).

One crucial pre-operation process is *capability acquisition and development*. Capabilities in network warfare include all hardware and software used to affect enemy platforms. There is some limited merit in downplaying the complexities of this process; unlike actual weapons, network intrusion tools can ostensibly be developed by anyone. Similarly, the development cycle for a potent so-called "cyber-weapon" is also typically deemed to be much shorter (Rattray 2001, 171), easier and cheaper (Nye 2010, 5). Again, there is some reason to this assertion. However, the unique circumstances of developing capabilities to attack networks are well worth examining. Each supposed advantage is mirrored by an equal or greater disadvantage.

Presence-based attack tools must be stealthy, agile, and modular. They must be stealthy as the majority of their life-cycle will be spent clandestinely embedded in adversary networks. They must be agile to enable operators to use them creatively to traverse adversary networks, collect intelligence and weaponize valuable targets. Finally, they must often be modular to allow operators to only deploy necessary capabilities at any given moment, thereby reducing the footprint of the tool, a further operational security mechanism (Monte 2015, 124). Each deployment of a highly engineered network attack tool must be carefully managed to include only the components currently needed to facilitate success. The expectation that presence-based operational tools must be stealthy introduces a significant weakness: these tools become quite brittle in use. The pervasive notion that offensive network tools are single-use stems from this very issue (Libicki 2009, 83). The defensive cycle for a network adversary is demonstrably shorter, as detected malware can result in detection signature within days of its discovery by a capable defender. It is not just the particular deployment that is threatened; detection of an offensive platform risks its compromise against all targets against which it is currently employed. That is a momentous risk of capabilities, which explains in part why intelligence agencies often guard them so carefully.

It is almost inconceivable that network attack tools could enjoy the same operational longevity as their kinetic counterparts. One of the longest known offensive network operations platforms – codenamed Regin by its private-sector discoverers – was ostensibly operating from at least 2003 (Kaspersky Lab 2014, 3) and widely attributed to the NSA (Rosenbach, Schmundt, and Stöcker 2015). At the time of its discovery in 2014, security company Kaspersky claimed that it was "…one of the most sophisticated attack platforms we have ever analyzed" (Kaspersky Lab 2014, 23). Once publicized and with its various mechanisms for communication and stealth thoroughly mapped and defended against, NSA operators would have had to immediately cease all intrusion activity until sufficient changes could be made and new evasion mechanisms deployed. Such an event is both an enormous investment in time and resources and also potentially a major operational compromise.

Conversely, event-based attack tools must be robust, aggressive, fool-proof and intuitive to operate. As they would likely be deployed by frontline units, no expertise must be needed to wield them effectively. They must be able to operate against a wide range of targets in a slew of contingencies, while generating similarly predictable effects. Battlefield operators will not have time to dynamically redeploy modules or carefully orchestrate network traversal. The weapon must therefore be capable of autonomously completing its objectives without further assistance. Resource exhaustion attacks, such as the often-seen denial of service attack or generic destructive payloads, are common examples of event-based capabilities.

Both presence and event-based capabilities require investment in *vulnerability research*. This entails all efforts to locate exploitable flaws in software and hardware used by the adversary: flaws that can be subverted to compromise the target and get it to either behave unexpectedly or preferably to run arbitrary code. Vulnerability research runs the gamut from generic-use software such as Microsoft Windows to dedicated software used by military hardware and other niche platforms. It is a crucial component in most network attack tools.

Software vulnerabilities are difficult to find both for attackers and defenders. From the offensive perspective, effectively exploiting critical software in a manner conducive to intrusions is increasingly difficult (Symantec 2017, 16). At the same time, there is no shortage of vulnerabilities, as data indicates that publicly disclosed, high severity submissions have nearly doubled in 2017 (NIST 2017). From the defender's perspective – as a RAND report indicated in 2017 – unless the tool weaponizing them is somehow discovered, vulnerabilities last an average of almost seven years without being exposed (Ablon 2017, 11). Thus, maintaining an expert workforce entrusted with continuously hunting for new useful vulnerabilities is paramount.

For event-based operations, the final component of preparation is integrating capabilities for use with forward-deployed warfighting platforms. Presence-based operations are often handled by remote operators, much like drones. However, in many cases, especially those involving segregated networks used to communicate sensitive data, proximity or line-of-sight access is required. In these cases, military forces may find themselves delivering fire directly in the field, be it by aircraft, naval vessel, ground vehicle or actual boots on the ground.

There are recent examples of event-based attacks in which network capabilities were supposedly integrated into battlefield platforms. The United States military operates infantry cyber teams to work alongside electronic warfare assets to map out enemy networks and identify targets (Kimmons 2017). The Russian military has, allegedly, disrupted Royal Air Force sorties over Syria by way of a network attack launched from a deployed electronic warfare vehicle (Giannangeli 2017). Developing a reliable, robust, battlefield-deployable offensive cyber capability is increasingly becoming viable, albeit expensive. Thus, while attacking networks may seem to be low-cost, attaining battlefield readiness and conducting event-based offensive operations may include hefty development, targeting and intelligence cycles.

3. ENGAGEMENT

The Cyber Threat Framework defines the initial engagement phase as: "Threat actor activities taken prior to gaining access but with the intent to gain unauthorized access to the intended victim's physical or virtual computer or information system(s), network(s), and/or data stores" (DNI US 2013, 4). Put simply, this phase embodies the attempts to intrude upon the enemy; it is the first active contact with its networks, intent on establishing a digital beach-head. What the framework obfuscates is the characteristics of this phase. Adopted from the operational typology used by Buchanan, the engagement phase may occur months in advance for presence-based operations or adjacent to the desired effect for event-based attacks (Buchanan 2017, 76–84). Not all cases are created equal, but all share one notable commonality; the engagement phase starts the operational clock.

A ubiquitous approach to network intrusion is compromising an internet-facing server or device. Identifying and compromising these may be easier than directly penetrating segregated networks, but not all such targets are inherently useful. Operations may also commence by interacting with an individual rather than a machine. Strategic network operations intended to gain entry to sensitive networks may first need to compromise those who routinely use them and hold trusted access to their assets. The reason for this is two-fold: first, there may not be a viable technological intrusion

vector, as many sensitive networks are cut off from external inputs; and second, the users are often the most vulnerable element in an otherwise secure network (Barrett 2003). They are prime targets for social engineering as an intrusion vector, but that does not mean it is always a trivial endeavor. Successfully getting individuals to usefully compromise their own security without arousing suspicion often requires expertise, preferably provided by dedicated personnel.

In event-based operations, the engagement phase can occur in seconds. As the targeting cycle is similarly shortened, there is no time to craft phishing emails tailored to human targets or set up elaborate honeypots. Instead, the engagement phase will focus on compromising accessible targets by exploiting remote software and hardware vulnerabilities. Particularly when using automated capabilities to target warfighters or other connected devices, it is sometimes possible to directly attack the software to gain entry. The engagement phase for event-based operations may not always result in full access to the target, but depending on what the desired effect is, that may not be necessary. For example, simply attempting to exhaust available resources or corrupt a target's means of communication may be possible without ever being able to execute code directly on the target and if the goal is to prevent the target from functioning as intended, that may be sufficient. Such scenarios are more easily placed within a military context; see for example denial of service attacks, which bear some similarities to conventional electromagnetic jamming.[6]

The potential perpetrators for event-based operations are far more varied than their presence-based counterparts. In many cases, these could be forward-deployed offensive cyber units, such as both the US and the UK are increasingly using (US Army 2014, 30–32). In other instances, field staff such as human intelligence assets or specific warfighters may be required to facilitate the actual engagement. As Edward Snowden revealed in a leaked top-secret document in 2013, the NSA's GENIE program to facilitate semi-automated network operations would at times rely on such assets. When necessary, field operators would physically infect adversary devices, plant hardware, or conduct short-range offensive SIGINT (NSA 2013). SIGINT agencies with global or regional reach could also deliver payloads from remote facilities.

4. PRESENCE

The presence phase is where most of the friction occurs between intruder and target. It is where persistent malicious software is continuously employed to understand, dissect, and establish a hold within the targeted network or networks, gradually extending the intruder's access until it locates servers or devices suitable to achieve the task at hand (US DNI 2013, 5). It is the process of extending and cementing

6 This aligns nicely with US military doctrine that situates Cyber and Electromagnetic Activities (CEMA) as a unified operational function, see US Army (2014).

the reach into the adversary's networks, two processes respectively called lateral movement and persistence.

The presence phase embodies the biggest discrepancy between the two operational categories – time spent on target. Where presence-based operations unsurprisingly spend most of their lifecycle in the presence phase, event-based operations may have an inconsequential or even non-existent presence phase. When nation-state intrusion campaigns are analyzed and reported to take months prior to detection, this primarily refers to the presence phase. The key difference in timespan reflects applicability to two wholly different operational tempos. For presence-based operations, the presence phase is essentially a cyclical process of expanding micro-intrusions in which additional nodes in the network are scanned, breached and subsequently assessed for mission relevance. This is represented well in the Kill Chain model, which threads multiple compromises on targeted networks into a single campaign with shared features (Hutchins, Cloppert, and Amin 2011, 7–8). Each intrusion must be handled with care to avoid tripping any alarms or informing network defenders of an active intrusion against them.

Presence-based offensive operations are first intelligence operations. Until such a time as a more active measure is needed, malicious software is tasked with either remaining dormant or collecting information, identical to the behavior in an intelligence mission (Lin 2010, 64). As a corollary, operators in the presence phase must rely extensively on the assistance of intelligence analysts to assist in further targeting and dissection of materials exfiltrated from the target (Malone 2010, 16). In some cases, the offensive is carried out entirely by the intelligence agency (GCHQ 2012). The presence phase is thus both assessing the independent intelligence value of the target, and simultaneously gathering information needed to help steer the operators towards the server or servers where attacking would result in achieving the desired objective.

When Russian operators initially infiltrated the Ukrainian power grid in 2015, they did not immediately wreak havoc on all they encountered. Instead, earlier intrusion efforts cleverly used the specialized protocols unique to these industrial networks to traverse the network, map its layout and glean the information required to develop robust offensive capabilities (Dragos 2017, 9). In a subsequent operation, the presence phase included pivoting from the power company's corporate network onto its industrial network, leveraging an attack against both to simultaneously cripple the grid and prevent operators from fixing it (Dragos 2017, 10). Finally, advancements eventually allowed the operators to "…de-energize a transmission substation on December 17, 2016" (Dragos 2017, 4) by way of the CRASHOVERRIDE malware tailored to affect even relatively well-defended energy grids. The Russians had achieved a malware-induced blackout, but they had done so after a considerable amount of time from the

initial engagement phase. Success would not have been possible without expertise and accrued experience.

For event-based offensive operations, the presence phase is nearly imperceptible. This is intrinsic to the attack vector; capabilities employed in an event-based attack are meant to impact the target directly and then disappear, leaving as few lingering artefacts as possible. Were tell-tale indicators to remain, such as residual code left running or files persisting in the target's file system, it would simplify subsequent efforts by the adversary to develop future countermeasures. Thus, it is significant for an event-based capability to be only minimally present on enemy assets.

A cascading effect – intentional or otherwise – may result in an event-based attack having a limited period of network presence. For example, an automated network attack tool designed to propagate through networks and rapidly destroy all infected endpoints and servers would require a limited presence to ensure subsequent infections of additional targets. A good example of such an attack is the NotPetya destructive malware, which in 2017 heavily affected Ukrainian networks before cascading beyond its scope to adversely affect various other entities globally (Perlroth, Scott, and Frenkel 2017). The attack, which resulted in extensive damage to victims worldwide, was unusually publicly attributed by numerous Western intelligence agencies to the Russian military.[7]

The potential cost incurred in discovery is arguably the most meaningful deterrent to attacking via cyberspace. In recent years, a growing trend amongst large vendors in the information security market has been to uncover massive nation-state surveillance efforts, often facilitated by highly sophisticated malicious software. The immediate result of this compromise is an attempted rollback of all deployed assets, both by the original offender attempting to effect damage control and the victims who enjoy updated configurations for their defensive products. The product of this is a partial collapse of the aggressor's intrusion infrastructure and, more importantly, the defender's near-immediate inoculation against future attempts to use the same tool in an offensive capacity. The presence phase is thus the most sensitive component in many offensive network operations. The continuous friction with different adversary networks and the need to collect intelligence means that discovery and eventual inoculation are a big risk to attackers. Presence operators must therefore continuously work to conceal their moves, clean up evidence and establish stable, covert communication channels that would reliably allow decision-makers to activate positioned offensive payloads when necessary (Peterson 2013, 123).

7 See, for example, US Press Secretary (2018).

5. EFFECT

The final effect phase is where triggers are pulled. Ordnance is activated, disabling, disrupting or manipulating targets. Effects either translate into objectives, fizzle uselessly, or have unintended and potentially disastrous collateral effects. For presence-based operations, the effect phase is the culmination of possibly months of planning, targeting, intelligence collection, infection attempts and dedicated development (Rattray and Healey 2010, 79). For event-based operations, the effect phase represents the primary thrust of the attack. When Richard Clarke declared in 2009 that "strikes in cyber war move at a rate approaching the speed of light" (Clarke 2009, 32), he was not referring to the entire span of an operation, but rather to the period of time between the activation of the ordnance and its detonation on the target, the manifestation of the effect phase. Even so, ordnance may be instantly triggered but may still take time to deliver its intended effect.

Distilling various official definitions, there are three "attack" types when targeting networks – disruptive, manipulative, and destructive.[8] Disruptive, or suppressing, attacks inflict "temporary or transient degradation by an opposing force of the performance of a weapon system below the level needed to fulfil its mission objectives" (US DoD 2017, 229). Their utility increased with the rise of electronic warfare, where electromagnetic transmissions could be jammed to produce a temporary but potent effect (Army Headquarters 2003, 7). The concept of disruptive attacks has made a natural transition to cyberspace, where temporarily degrading the capacity of military resources can adversely affect the efficacy of an adversary force (US Army 2014, 9).

Disruptive network attacks are commonplace even outside military scenarios. So-called denial-of-service attacks capable of levying massive throughput of network traffic routinely disrupt the functionality of online services, big and small. The targets range from global gaming communities such as the Sony PlayStation Network (Samit 2016) to major banks (Hamill 2014). Typically, these attacks either exploit an implementation flaw in the targeted technology or simply attempt to overwhelm its available resources. No legitimate connections can interact with the platform as intended, rendering it temporarily disabled for its original purpose. Similar approaches may be applied to military technology, platforms and protocols.

Manipulation effects attempt to alter information or functionality in the adversary networks, thereby deceiving operators or preventing intended system functionality. Such attacks attempt to alter perception, preventing an adversary from acting properly to further its own objectives. A scenario could include introducing a nearly imperceptible deviation to a weapon's targeting process, causing strikes to miss due

8 Adapted from the US Military's taxonomy of "…deceive, degrade, deny, destroy, or manipulate…", see US Army (2014, 17). Libicki similarly speaks of attacks aimed at eruption (target illumination), disruption, and corruption. See Libicki (2009, 145).

to what could appear to be a technical glitch. Kinetically, this is hard to accomplish but could be roughly analogized to physically tampering with a missile's warhead to secretly render it inert. When the missile fires, it seemingly behaves as normal until impact, when the warhead does not detonate. During the heat of conflict and until it happens repeatedly and consistently, it would be difficult to identify the fault as an attack. By the time it is discovered, it would likely already be too late. As the Stuxnet campaign demonstrated (Falliere, Murchu, and Chien 2011; Farwell and Rohozinski 2011), masking a manipulative effect to increase its longevity can cause an effect to be repeatedly successful over time. Hiding an effect does, however, require incrementally introducing it; an immediate and blunt change of circumstance markedly increases the probability of detection.

Destructive attacks are intended to inflict damage on adversary networks, either on hardware, software or both. These types of attacks are firmly rooted in conventional warfare, where destruction of enemy assets and personnel is often seen as the primary method of reducing its combat effectiveness.[9] When applied to network operations, a destructive attack could cause permanent software damage, such as in the case of malware which completely erases all critical files on target servers,[10] or even permanent hardware damage, such as the previously mentioned Stuxnet worm targeting the Iranian nuclear project (Langner 2011).

6. CHALLENGES AND OPPORTUNITIES

Delineating between event-based and presence-based operations allows a discussion on how militaries are integrating these capabilities into doctrine and strategy. They are markedly different in characteristics, duration, challenges, and opportunities and thus must not be lumped together, but fundamental similarities exist between the two categories and are certainly helpful in understanding networks as a medium for warfare; but useful observation of military capabilities will remain limited unless we recognize that not all capabilities must be treated the same.

Event-based operations represent the instances in which network attacks are somewhat analogous to the kinetic. Like firing a weapon, an event-based operation entails sending a payload from attacker to target in the hope of immediately reducing its integrity or capacity to operate. As a result, these capabilities are often more tactical in nature, easier to integrate with existing military OODA loops,[11] and are promising candidates for joint warfare. They are, however, limited in scope, may require extensive research

9 The classic approach to warfare - most commonly codified by Prussian strategist Carl von Clausewitz – favours destruction as the sole means of achieving military coercion. See Clausewitz (1873) for the original school of thought.

10 See, for example, the 2012 Shamoon attack, in which a presumably Iranian attacker wiped thousands of computers at Saudi's national gas company, Aramco (Bronk and Tikk-Ringas 2013).

11 OODA loop – A process in which combatants Observe, Orient, Decide, and Act. Military vernacular for conceptualising decision-making process in combat. See Boyd (1995).

and development, and could be limited to a specific subset of adversary equipment. A weapon suitable for disabling a US Navy destroyer may exploit hardware-specific vulnerabilities,[12] rendering it unsuitable against other targets. Consequently, battlefield operators deploying such weapons must have immaculate understanding of their adversary and a firm control of their own options.

Presence-based operations are intelligence missions with an offensive finisher; a form of digital sabotage. They may initially appear indistinguishable as operators infect networks and gather information necessary to craft an attack. In these phases, even if the target detects the malware present in its assets, it is very difficult to assess motive and intent. Only once offensive modules are deployed can confidence in hostile intent increase. This adds an unfortunate layer of political nuance, as overly successful network intrusions may be misconstrued by the target as unduly aggressive. The risk of potentially undesired escalation has been aptly covered by Buchanan when discussing the "cybersecurity dilemma" (Buchanan 2017), an application of the classic security dilemma to network intrusions between nations.

Presence-based operations can potentially be high-risk, high-reward capabilities. Successfully pre-positioning assets in military or otherwise critical networks may potentially have meaningful impact on the course of conflict if used to facilitate strategic surprise or large-scale reduction in enemy capacity to operate. At the same time, presence-based operations are notoriously brittle, and their discovery can undo years of focused labor. By nature, such operations require tight, intensive, unyielding support of friendly intelligence assets to map the threat, generate initial persistent access, and successfully maneuver through complex tangles of military networks until the right targets are found. It is therefore understandable why these campaigns are often spearheaded by intelligence agencies with core expertise on network intrusions rather than deployed military forces.

The Lockheed-Martin F-35 Lightning II fighter aircraft is a fascinating example of a platform potentially vulnerable to both presence-based and event-based attacks. After two decades of development, the aircraft had started active deployment accompanied by a host of issues with its onboard software. These included major in-flight failures of the radar system (Gallagher 2016), issues with its onboard avionics (US DoD 2016, 35), and "…276 deficiencies in combat performance [designated] as 'critical to correct'…" (US DoD 2017, 48). Additionally, both the onboard systems and the logistical software used to manage the F-35 have demonstrated numerous vulnerabilities during security testing procedures, many yet to be addressed as of 2017 (US DoD 2017, 103–4). While onboard systems are unlikely to be directly connected to the internet (Lin 2010, 66), targeting one or more of the F-35's prized array of sensory inputs and communication methods is possible for a knowledgeable adversary.

12 These vulnerabilities do indeed exist, see for example US DoD (2017, 3).

An event-based attack might try to overwhelm or otherwise compromise some of the F-35's tactical data links, used to share data with allied assets in the air and on the ground. For compatibility purposes, this communication commonly occurs via the Link-16 protocol, an encrypted legacy protocol used by NATO forces since 1975. While it has undoubtedly undergone improvements over its lifecycle, the limitations in encrypting reliable airborne tactical traffic and the vast array of opportunities for US adversaries to intercept, analyze and exploit Link-16 protocol vulnerabilities raise the option that it may be compromised during an attack. Link-16 includes targeting information, location of friendly forces and directives from command forces (Hura et al. 2000). Interestingly, even oversight reports have indicated some issues with the Link-16 data that forced pilots to revert to voice communication (US DoD 2017, 70). Others have indicated intermittent problems with the Multifunction Advanced Data Link (MADL) system used to communicate between fifth generation stealth aircraft,[13] causing pilots to 'lose tactical battlefield awareness' (US DoD 2017, 71). Successfully compromising the F-35's data links is thus not unfeasible and may severely degrade aircraft battlefield performance.

The effects phase in this particular instance could include one of several options. As an example, a manipulation attack could alter the pilot's perception of the battlefield by adding, removing, or moving specific targeting points fed to the radar subsystem by external channels. A disruptive attack could try to overwhelm sensory input or prevent the aircraft from awareness of being acquired by a ground-based air-defense battery. The effects would thus be nearly instantaneous, limited in scope to the targeted aircraft, and tactical in nature.

A presence-based attack against the F-35 could take months to prepare, culminating in an elaborate effects phase saved for evoking strategic surprise or in dire need. Rather than targeting a single aircraft or sortie, attackers would instead target the peripheral networks that interface with the F-35 during its operational life cycle. These could be on-base networks, maintenance forces or third-party software providers. By doing so, an adversary may temporarily degrade or completely disable a large number of aircraft.

One supposed innovation in the F-35's software is the Autonomic Logistic Information System, or ALIS. With one ALIS station present at each unit operating F-35s, it allows semi-automated fleet management, mission management, logistics, and maintenance (Lockheed Martin 2009). As with other parts of the Joint Strike Fighter program, ALIS has been plagued with critical faults which are instructive in two relevant aspects: how ALIS might be vulnerable to presence-based operations; and how exploiting these vulnerabilities could lead to a strategic advantage when triggered in the effects phase.

13 Currently for the US, the F-22 and the F-35.

The issues in ALIS are varied. Attempts to deploy it in test environments have forced support personnel to lower network security settings to allow users to log on (US DoD 2017, 96). Incorrectly handled maintenance data has resulted in one instance in "major damage to a weapons bay door" (US DoD 2017, 96) from an incorrectly loaded bomb that got loose and struck the aircraft. In June 2017, a software error in ALIS grounded an entire F-35 unit until the issue was addressed (Freedberg Jr. 2017). It would therefore seem that the system can both be a boon to aircraft operators and an attack vector for offensive network operators. A single warfighting platform now presents a diverse, varied attack surface that can potentially be exploited during wartime.

All military offensive network operations can be a tremendous boon to military objectives across all levels of operation. Each type has unique characteristics, requires different support staff, and may weave into doctrine at varying locations. Where event-based operations may assist in crippling a local adversary network to facilitate joint strikes, a well-placed presence-based capability may sufficiently delay adversary decision-making and resource martialing to strategically diminish the capacity for effective response. From sowing tactical chaos to deceiving a carrier strike group, the potential is vast – if each category is understood, respected, and contextually integrated.

REFERENCES

Ablon, Lillian. 2017. Zero days, thousands of nights: the life and times of zero-day vulnerabilities and their exploits. Santa Monica: Rand Corporation.

Army Headquarters. 2003. "US Army Field Manual 3-13 - Information Operations."

Atkin, Thomas, James McLaughlin, and Charles Moore. June 26, 2016. Hearing Before the House Armed Service Committee. Washington DC. http://docs.house.gov/meetings/AS/AS00/20160622/105099/HHRG-114-AS00-Wstate-AtkinT-20160622.pdf.

Barrett, Neil. 2003. "Penetration Testing and Social Engineering: Hacking the Weakest Link." *Information Security Technical Report* 8 (4): 56–64.

Boyd, John. 1995. "The Essence of Winning and Losing." June 28. http://pogoarchives.org/m/dni/john_boyd_compendium/essence_of_winning_losing.pdf.

Bronk, Christopher, and Eneken Tikk-Ringas. 2013. "The Cyber Attack on Saudi Aramco." *Survival* 55 (2): 81–96. https://doi.org/10.1080/00396338.2013.784468.

Buchanan, Ben. 2017. *The Cybersecurity Dilemma Hacking, Trust and Fear Between Nations*. Oxford: Oxford University Press.

Burbank, Jack L., Philip F. Chimento, Brian K. Haberman, and William T. Kasch. 2006. "Key Challenges of Military Tactical Networking and the Elusive Promise of MANET Technology." *IEEE Communications Magazine* 44 (11). http://ieeexplore.ieee.org/abstract/document/4014472/.

Caltagirone, Sergio, Andrew Pendergast, and Christopher Betz. 2013. "The Diamond Model of Intrusion Analysis." DTIC Document. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA586960.

Chairman of the Joint Chiefs of Staff. 2006. "National Military Strategy for Cyberspace Operations."

Clarke, Richard A. 2009. "War from Cyberspace." *The National Interest*, 31–36.

Clausewitz, Carl Von. 1873. *On War*. 3rd ed. Vol. 1. London: N. Trubner & Co.

Coile, Gregory. 2009. "WIN-T SATCOM Overview Briefing." Program Executive Office Command Control Communications-Tactical. http://www.afcea-aberdeen.org/files/presentations/afceaaberdeen_ltccoile_28may2013.pdf.

Conti, Gregory, and David Raymond. 2017. *On Cyber: Towards an Operational Art for Cyber Conflict*. Kopidion Press.

Crawford, Bruce T., James J. Mingus, and Gary P. Martin. 2017. The United States Army Network Modernization Strategy. http://docs.house.gov/meetings/AS/AS25/20170927/106451/HHRG-115-AS25-Wstate-CrawfordB-20170927.pdf.

Dragos. 2017. "CRASHOVERRIDE: Threat to the Electric Grid Operations." Dragos. https://dragos.com/blog/crashoverride/CrashOverride-01.pdf.

Ducheine, Paul, and Jelle van Haaster. 2014. "Fighting Power, Targeting and Cyber Operations." In *Cyber Conflict (CyCon 2014), 2014 6th International Conference On*, 303–327. IEEE. http://ieeexplore.ieee.org/abstract/document/6916410/.

Epperson, Lynn. 2014. "Satellite Communications Within the Army's WIN-T Architecture" Program Executive Office Command Control Communications-Tactical. http://studylib.net/doc/18136899/satellite-communications-within-the-army-s-win.

Falliere, Nicolas, Liam O Murchu, and Eric Chien. 2011. "W32.Stuxnet Dossier" Symantec. https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf.

Farwell, James P., and Rafal Rohozinski. 2011. "Stuxnet and the Future of Cyber War." *Survival* 53 (1): 23–40. https://doi.org/10.1080/00396338.2011.555586.

Freedberg Jr., Sydney J. 2017. "ALIS Glitch Grounds Marine F-35Bs." *Breaking Defense* (blog). June 22, 2017. http://breakingdefense.com/2017/06/breaking-alis-glitch-grounds-marine-f-35bs/.

Gallagher, Sean. 2016. "F-35 Radar System Has Bug That Requires Hard Reboot in Flight." *Ars Technica*. March 10, 2016. https://arstechnica.com/information-technology/2016/03/f-35-radar-system-has-bug-that-requires-hard-reboot-in-flight/.

GCHQ. 2012. "Full-Spectrum Cyber Effects". https://snowdenarchive.cjfe.org/greenstone/collect/snowden1/index/assoc/HASH8311.dir/doc.pdf.

Giannangeli, Marco. 2017. "Russians 'Hacking into' RAF Crews over Syria." *The Daily Express*. January 15, 2017. http://www.express.co.uk/news/world/754236/russia-raf-bombers-syria-hacking-missions-military-army.

Hamill, Jasper. 2014. "Bank-Busting Jihadi Botnet Comes Back To Life. But Who Is Controlling It This Time?" *Forbes*. June 30, 2014. https://www.forbes.com/sites/jasperhamill/2014/06/30/bank-busting-jihadi-botnet-comes-back-to-life-but-who-is-controlling-it-this-time/#3df4bb0f6f07.

Hura, Myron, Gary McLeod, James Schneider, Daniel Gonzales, Daniel M. Norton, Jody Jacobs, Kevin M. O'Connel, William Little, Richard Mesic, and Lewis Jamison. 2000. "Tactical Data Links." In *Interoperability: A Continuing Challenge*, 107–21. Chapter 9 - Tactical Data Links: RAND.

Hutchins, Eric M., Michael J. Cloppert, and Rohan M. Amin. 2011. “Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains.” *Leading Issues in Information Warfare & Security Research* 1: 80.

Kaspersky Lab. 2014. “The Regin Platform: Nation State Ownage of GSM Networks” https://securelist.com/files/2014/11/Kaspersky_Lab_whitepaper_Regin_platform_eng.pdf.

Kimmons, Sean. 2017. “Cyber Teams Throw Virtual Effects, Defend Networks against ISIS.” United States Army. February 15, 2017. http://www.army.mil/article/182400/cyber_teams_throw_virtual_effects_defend_networks_against_isil.

Langner, Ralph. 2011. “Stuxnet - Dissecting a Cyberwarfare Weapon.” *IEEE Security and Privacy* 9 (3): 49–51.

Libicki, Martin C. 2009. *Cyberdeterrence and Cyberwar.* Santa Monica, CA: RAND.

Lin, Herbert S. 2010. “Offensive Cyber Operations and the Use of Force.” *Journal of National Security Law and Policy* 4: 63.

Lockheed Martin. 2009. “Autonomic Logistics Information System (ALIS).” Lockheed Martin.

Malone, Jeff. 2010. “Intelligence Support Requirements for Offensive CNO.” presented at the Cyber Warfare and Nation States Conference, Canberra, Australia, August 23.

Monte, Matthew. 2015. *Network Attacks & Exploitation: A Framework*. Indianapolis, IN, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119183440.

NIST. 2017. “NVD - CVSS Severity Distribution Over Time.” NIST. 2017. https://nvd.nist.gov/vuln-metrics/visualizations/cvss-severity-distribution-over-time.

NSA. “Computer Network Operations - GENIE.” 2013. National Security Agency. https://www.eff.org/files/2015/02/03/20150117-spiegel-excerpt_from_the_secret_nsa_budget_on_computer_network_operations_-_code_word_genie.pdf.

Nye, Joseph S. 2010. “Cyber Power.” DTIC Document. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA522626.

Perlroth, Nicole, Mark Scott, and Sheera Frenkel. 2017. “Cyberattack Hits Ukraine Then Spreads Internationally.” *The New York Times*, June 27, 2017, sec. Technology. https://www.nytimes.com/2017/06/27/technology/ransomware-hackers.html.

Peterson, Dale. 2013. “Offensive Cyber Weapons: Construction, Development, and Employment.” *Journal of Strategic Studies* 36 (1): 120–24. https://doi.org/10.1080/01402390.2012.742014.

Rattray, Gregory J. 2001. *Strategic Warfare in Cyberspace*. Cambridge, Mass: MIT Press.

Rattray, Gregory J., and Jason Healey. 2010. “Categorizing and Understanding Offensive Cyber Capabilities and Their Use.” In *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy*. Washington D.C.: National Academic Press.

Rosenbach, Marcel, Hilmar Schmundt, and Christian Stöcker. 2015. “Source Code Similarities: Experts Unmask ‘Regin’ Trojan as NSA Tool.” *Spiegel Online*, January 27, 2015, sec. International. http://www.spiegel.de/international/world/regin-malware-unmasked-as-nsa-tool-after-spiegel-publishes-source-code-a-1015255.html.

Samit, Sarkar. 2016. “Massive DDoS Attack Affecting PSN, Some Xbox Live Apps.” Polygon. October 21, 2016. https://www.polygon.com/2016/10/21/13361014/psn-xbox-live-down-ddos-attack-dyn.

Symantec. 2017. “Internet Security Threat Report.” Symantec. https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf.

US Army. 2014. “Army Field Manual 3-38 - Cyber Electromagnetic Activities.”

US DNI. 2013. “A Common Cyber Threat Framework: A Foundation for Communication.” Office of the Direction of National Intelligence. https://www.dni.gov/files/ODNI/documents/features/Threat_Framework_A_Foundation_for_Communication.pdf.

US DNI. 2013. “Cyber Threat Framework Lexicon.” Office of the Director of National Intelligence.

US DoD. 5/17. “Joint Publication 1-02: DoD Dictionary.” US Department of Defense.

US DoD. 2003. “Information Operations Roadmap.” US Department of Defense.

US DoD. 2013. “Joint Publication 2-0: Joint Intelligence.” US Department of Defense.

US DoD. 2016. “Fiscal Year 2015 DoD Programs - F-35 Joint Strike Fighter (JSF).” US Department of Defense.

US DoD. 2017. “Aegis Modernization Report Program - Fiscal Year 2016.” US Department of Defense.

US DoD. 2017. “Fiscal Year 2016 DoD Programs - F-35 Joint Strike Fighter (JSF).” US Department of Defense.

US Joint Chief of Staff. 2013. “Joint Publication 3-12: Cyber Operations.” US Joint Chief of Staff.

US Press Secretary. 2018. “Statement from the Press Secretary.” The White House. February 15, 2018. https://www.whitehouse.gov/briefings-statements/statement-press-secretary-25/.

Wikileaks. 2017. “Vault 7: CIA Hacking Tools Revealed.” Wikileaks. March 7, 2017. https://wikileaks.org/ciav7p1/.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Drawing Inferences from Cyber Espionage

Martin C. Libicki
Center for Cyber Security Studies
U.S. Naval Academy
Annapolis, MD, United States
libicki@usna.edu; libmazo@gmail.com

Abstract: To survive a confrontation, it helps to understand other side's capabilities and intensions. Estimates of opposing capabilities rest on an empirical basis but understanding the other side's intentions is inferred from words and deeds.

Therein lies a dilemma common across all military domains: acts to alter the balance of a confrontation can also shape the inferences that the other side draws about one's intentions. The dilemma also operates in cyberspace, but in unique ways.

First, efforts by one side to acquire information on the other can be read by the other side as preparations for a cyber attack prefatory to a military attack.
Second, others may draw inferences from the fact of cyber espionage alone, even though the basis for believing in a cyber security dilemma is weak.
Third, there are ways of carrying out cyber espionage that can mitigate inferences that others draw about the imminence of cyber attack by, for example, limiting which components within a network are targeted for eavesdropping or by using penetration methods that do not leave arbitrary code behind.
Fourth, defenders themselves can also modulate their reactions in ways that limit drawing unnecessary inferences.
Fifth, expectations of how well modulating cyber espionage can convey peaceful intentions should be very modest.

All these are complicated by difficulties in the target's ascertaining a penetration's date, characterization, and authorship. We conclude with a call for those who would penetrate military-related systems to think about the inferences that the other side may draw if such penetrations are discovered.

Keywords: *cyber espionage, cyber attack, signaling*

1. INTRODUCTION

To survive a confrontation, it helps to be mindful of the other side's capabilities and intentions. Estimates of opposing capabilities often take painstaking work, but at least rest on an empirical basis. But understanding the other side's intentions is something that needs to be inferred from words and deeds.[1]

Therein lies a dilemma present across conflict domains. Acting can alter the terms of a confrontation to the actor's advantage, but it can also shape the inferences that the other side draws about one's intentions. Some inferences can both help *and* harm. One side may want to signal its resolve to attain and defend some objective. It does so by demonstrating capability, readiness, and a willingness to put people and assets in harm's way. It hopes that the other side backs off. But the other side may infer not only that its potential foe is prepared and willing, but also that it is facing a now higher level of aggression. Perhaps the objective has grown or the willingness to take risks to achieve it has risen. So, the other side sees a growing threat – one that forces it to do something to recover its former level of security. Therefore, it decides to bolster its own capability, readiness, and willingness to fight.[2] The advantages that one side reaps from its actions can be thereby nullified by the indirect disadvantages because the other side is drawing unhelpful inferences about its adversary's intensions.

We contend that the dilemma operates in cyberspace, but in a unique way – efforts by one side to acquire information on the other can be read by that other side as preparations for a cyber attack prefatory to a military attack.[3] It hardly helps stability when the high degree of ambiguity present in cyberspace combines with the thin experience base of cyber attacks and its non-physical (hence non-intuitive) nature. Perhaps needless to add, what happens in cyberspace matters to conventional military affairs more than it used to.

This essay walks through the problems and issues that may arise when inferences are drawn from activity in cyberspace, particularly those that take place during a crisis or confrontation. One might imagine, for reference purposes, that China and the United States are at odds over the South China Sea; neither is certain what the other side wants or how far it is willing to go, even if each has a good idea of what physical assets are to hand. So, what considerations should go into each side's rules of engagement in cyberspace?

1 The classic treatment being Robert Jervis, *Perception and Misperception in International Politics*, Princeton NJ (Princeton University Press), 1978.

2 Elsa Kania, "Cyber Deterrence in Times of Cyber Anarchy: Evaluating the Divergences in U.S. and Chinese Strategic Thinking," November 11, 2016; unpublished paper.

3 The logic that links a cyber attack to a kinetic attack is that because many of the effects of a cyber attack are temporary and reversible, carrying one out is pointless unless the intent is to exploit a temporary interruption or degradation of the other side's information services by using kinetic forces to make permanent changes in the military balance or outcomes.

In addressing this question, this paper distinguishes cyber espionage, which is unauthorized access to systems in order to acquire information, from cyber attack, which entails accessing systems in order to disrupt their operations or corrupt their information. To put this in the language of the CIA triad: cyber espionage affects only confidentiality while cyber attacks affect integrity and availability. Unfortunately, popular use generally applies "cyber attack" to a broad array of mischief in cyberspace, including the manipulation of social media. Cyber attack, in this paper, is also distinguished from "attack," which is used to mean kinetic attack using physical force.

2. INFERRING CYBER ATTACKS FROM CYBER ESPIONAGE

Cyber espionage can create knowledge *and* help set up cyber attacks; yet, if discovered, it may alter the target's assessment of the intruder's capabilities and intentions. The first is generally helpful. The second is generally harmful, in that the target may conclude that the intruder is preparing to fight and to do so soon.

Although caution is therefore advised in penetrating systems whose disturbance may enflame the other side, in a crisis a country may want to carry out *more* cyber espionage in order to determine the status, readiness, and intentions of the other side's armed forces. Indeed, as with spy satellites in the 1960s whose imagery persuaded U.S. leaders not to panic over the size of Soviet ICBM arsenals, or as former British intelligence officials would argue,[4] better intelligence tends to foster stability. It substitutes fact for doubt in situations in which leaders believe they must assume the worst, and hence gird for conflict. Some risk is inescapable. Even if traditional espionage uses tools clearly different from those used in warfighting, the heightened effort to collect intelligence prefatory to bolstering defense is nearly indistinguishable from efforts to collect intelligence prefatory to offense. Thus, any discovery of heightened intelligence efforts may lead the target to react badly.

Moreover, because a malware implant designed for cyber espionage is often identical to one designed for cyber attack, discovering and attributing[5] one in a critical system could easily be viewed as a *direct* precursor to attack. This normally would lead the target to raise its alert levels, which, in and of itself may exacerbate tensions.[6] In a crisis, not only are alert levels high to begin with, but so is suspicion of the other side's motives.

4 Based on remarks by Nigel Inkster (personal communications) and Sir David Osmand (http://carnegieendowment.org/2017/03/20/concurrent-session-i-cyber-weapons-and-strategic-stability-pub-67884).

5 Although attribution can be uncertain, the paper focusses on two countries in a confrontation at the time of discovery. Thus, the target is probably more apt to blame the intrusion on the other side (because it is easier to impute a motive) than if there were no confrontation.

6 Paul Bracken described how ominous signs could make the other side raise its alert level in his "Strategic War Termination," in Ashton B. Carter, John D. Steinbruner, and Charles A. Zraket, eds., Managing Nuclear Operations (Washington, D.C.: The Brookings Institution, 1987), pp. 197–214.

One important facet in drawing inferences from an implant is that its implantation would reflect conditions true at the time of its implantation rather than at the time of its discovery. Good forensic teams working on well-monitored networks can often figure out when an intrusion took place, and hence shed light on why.[7] If the penetration predated the crisis, it may be deemed not to be part of a dynamic of escalating alert levels. Nothing, of course, prevents one country from implanting malware against the day it might be needed for attack, but discovery alone cannot support the supposition that any such attack will take place imminently.

However, because many countries lack access to good forensics or fail to monitor their networks assiduously, the age of the intrusion may not be obvious to *them*. And until the other side figures out *when* the first penetration that resulted in a system's compromise took place, it may, in fear, conclude that the penetration was recent enough to have been motivated by the crisis itself.

The target need not be not forced into one conclusion. Perhaps what looks like cyber espionage was just fact-finding. Yet even cyber espionage unrelated to any possible cyber attack is not necessarily innocent. If the compromised system tracks military units in real time, an implant into it is still cyber espionage, but can also be used for later adversary targeting. Discovering that such a system was compromised regardless of how long ago, *should* raise concerns, just not ones that require going onto a war footing.

Now, what if the target infers that the intrusion was meant to be seen?[8] Granted it is difficult to distinguish between: (1) the desire to be seen; (2) an indifference to being seen which leads to a relaxation of operational security, thereby raising the likelihood of being seen; and (3) simple bad luck on the intruder's part. The target, in drawing inferences from what it has discovered, may also forget that the characteristics of discovered intrusions are not necessarily characteristics of undiscovered ones.[9]

7 The fact, for instance, that intrusions against the DNC started in the summer of 2015 strongly suggests that their motivation was more anti-Clinton than pro-Trump, whose nomination was hardly assured at that point.

8 The DNC had been penetrated for roughly a year before discovery (Dmitri Alperovitch, "Bears in the Midst: Intrusion into the Democratic National Committee", June 15, 2016: At the DNC, COZY BEAR intrusion had been identified going back to summer of 2015, while FANCY BEAR separately breached the network in April 2016). Yet the FBI still argued, "The most startling exchange at this week's hearing involved questions about why Russian hackers were so indiscreet when they stole e-mails from the Democratic National Committee and from the head of the Clinton campaign. That 'loudness' looks deliberate, Mr Comey replied." (source: "The FBI says it is investigating the president's campaign," March 23, 2017; http://www.economist.com/news/united-states/21719491-slice-country-hears-president-victim-government). See also Julian Borger, "Trump-Russia collusion is being investigated by FBI, Comey confirms", March 20, 2017; https://www.theguardian.com/us-news/2017/mar/20/fbi-director-comey-confirms-investigation-trump-russia: "The Russian intervention in the election was 'unusually loud', as if Moscow did not care about being caught."

9 Presumably, intrusions that are discovered are those that are easiest to discover. Their discoverability may not characterize the discoverability of the average intrusion (unless all of them are eventually discovered).

Still, the target's perception that the intruder was brandishing its capabilities by allowing its implants to be discovered – when spies normally go to great lengths to hide *theirs* – may persuade it to see coercion taking place. It could then ask: for what purpose? And why now? This could have been a periodic reminder and hence not indicative of an imminent threat. Logically, it should not indicate an imminent attack, since the attacker should be at pains to mask its intentions until they are suddenly revealed. But it could be a warning to back down, by containing the implicit message that failure to do so would be dangerous.

Another complicating factor with cyberspace operations arises from the question: how can countries underscore the credibility of deterrence instruments (such as retaliatory cyber attacks) without revealing the particulars of such capabilities and thereby inducing countermeasures?[10] Because countermeasures do not emerge immediately when systems prove broadly vulnerable, the target may infer that the other side is signaling its urgency by revealing what it can do *and* that it will not be needing such capabilities for long. If the target concludes from the intruder's presumed willingness to burn exploits that the intruder needed to make a quick impression, the target may then ask what the occasion is or will be.

The target may also conclude that the intrusion was undertaken to test the efficacy of and reaction to a cyber attack to be launched at some later date. This conclusion would be reinforced if it was a cyber attack, albeit a small one, that had taken place. Evidence for that may include the location of the intrusion, the identity of the affected systems, or the presence of attack code within the implant. Its placement or characteristics may persuade the target that the attacker has little confidence of being able to access the implant once the system goes to war.[11] But even such a discovery would not be particularly good evidence of an imminent attack, especially if the characteristics of the implant suggested the attacker's confidence that it could persist indefinitely without discovery.

Conversely, if the target concludes that a nominal cyber attack was carried out primarily as a final test prior to deployment, it may expect that use to be imminent. Its fears may rise if the implant's placement, characteristics and, especially, its implantation date suggest that the attacker was risking a high likelihood of discovery to validate or characterize a particular type of cyber attack. It is but a short step for the target to infer that discovery is evidence of discoverability, and thereby conclude that detonation is coming sooner rather than later. Further evidence of imminent use may be an implant's fragility, in that it is not robust against the run of changes that systems undergo. Other indications are recent rises in the frequency or scale of communications between the

10 See, for instance, Austin Long, Brendan Green, "Clandestine Capabilities and Deterrence in World Politics", unpublished.

11 This raises the question of how to activate the cyber attack if the implant is unreachable, but the answer may be that activation – a one-bit decision – can be triggered on the malware's assessment of network events in cases where malware cannot build attack code on the fly.

implant and its controller, or tests of the ability of the implant to support a certain payload. The latter can sometimes be inferred from reading logs.

Finally, any particular intrusion may serve several purposes. Concluding that one purpose may have been relatively benign hardly proves that more malign purposes are absent.

3. INFERENCES FROM THE FACT OF CYBER ESPIONAGE ALONE

A country's reaction to having simply been spied on may reflect its take on the security dilemma. Countries that believe that someone else's gain is automatically their loss are apt to interpret intrusions more darkly than those that believe that both sides can simultaneously be more secure. Those inclined to believe that the other is implacably hostile will read events as proof of dark design; those inclined to impute a mix of motives to the other side will hold many differing interpretations and delay imputing malevolence to system intruders pending further evidence. Some will see Munich in 1938; others, Sarajevo in 1914. The usual caveats apply: countries with different political cultures may draw inferences differently, the various bureaucracies within a single country may disagree with one another, and members of the public, elite opinion, and private organizations may each have their own opinion.

Furthermore, what seems innocent after the crisis has passed may seem otherwise during the crisis. The human tendency to impute intent to random circumstance may lead to conclusions that *because* the discovery of implants happened to produce fear, they were meant to induce fear and their discovery was part of that plan.

That noted, the technical basis for imagining a security dilemma *in cyberspace* is weak, particularly compared to contests such as nuclear missiles versus nuclear missiles or WWI-era land forces versus similar land forces. There are several reasons why. *First*, the contest in cyberspace is asymmetric: the best measures against cyber attack are cyber defenses, not an opposing cyber attack capability used for counterforce purposes.[12] Most measures that increase defense do not allow one's own attackers to enjoy greater success.[13] *Second*, because the element of surprise is intrinsic to the

12 In other words, the cost-effectiveness of carrying out cyber attacks on the attackers themselves would be low, in large part because the primary assets used in cyber attacks, computer code and intelligence, are essentially indestructible, and the hardware used is easily replaced. This consideration has nothing to do with the relative cost-effectiveness of offense versus defense, or with deterrence in cyberspace.

13 Ben Buchanan (in *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*, Oxford 2017) has argued that NSA intrusions have provided information on adversary intrusion (and hence attack) capabilities that have permitted stronger defenses. Thus, stronger defenses *by potential attackers* against penetration would have yielded weaker defenses on the part of defenders allied with the penetrators. But even if true, information is available only on some actors not all, such information is only part of what it takes for defense, and networks that benefit from NSA-acquired information are only a fraction of the total networks in the United States (albeit perhaps disproportionately important ones).

success of a cyber attack, it would take great confidence in such defenses before one side is sufficiently emboldened by the prospect of impunity to launch its own cyber attacks. *Third*, even if all system defenses were perfect, the logic that in cyberspace impunity emboldens aggression must also presume that the other side will not escalate into physical combat. This presumption is valid only if the stakes involved are too small to merit violence. *Fourth*, the strong commercial consensus on the need for better cyber security in general means that actions that improve cyber security for one (e.g., the discovery of a vulnerability that leads to a patch, an improved understanding of cost-effective practices) often improve cyber security for all.

Cyber espionage, like espionage in general, also permits information to be transmitted in particularly credible ways. If one side in a confrontation were to aver that it lacked active planning for aggression, the other side may well dismiss its avowals as motivated. But if one were to *steal* corroborating information from potential foes, one would have to be very suspicious indeed to conclude that such information was deliberately planted there, particularly if finding it was hard.

Such deception *could* happen,[14] but carrying on ostensibly confidential communications under the assumption they were wiretapped and would therefore be transmitted to the other side's leadership requires either giving up all confidential channels or knowing in advance which channels would stay confidential and which would be penetrated. The same holds with even more weight if the deception involved physical evidence, such as the disposition of military forces. Thus, however irritated one side may be at being penetrated, a salve on this irritation is the presumption that one's peaceful intentions have been more credibly communicated than mere narrative would allow.

4. HOW TO KEEP ON WITH CYBER ESPIONAGE WITHOUT SO MUCH RISK

How might cyberspace spies suppress unhelpful inference-making? One way is to loosen the correlation between being spied on and being attacked. Presumably, countries will not credibly promise never to attack in cyberspace; doing so forgoes a potentially significant military advantage and anyway would not be believed. Nevertheless, the correlation between espionage and attack *can* be weakened by copious acts of cyber espionage *not* correlated with a cyber attack. But this may backfire if the other side thinks that this is being done deliberately – that is, to inhibit the target from raising its guard after discovering intrusions that really were prefatory to cyber attack. Besides, being caught spying a lot tends to make one look unfriendly to begin with.

14 A great deal depends on how widely system owners start using deception. One case is France's then-candidate Emmanuel Macron suspecting that Russia would penetrate his campaign's networks and lacing false documents in his networks. See Adam Nossiter, David Sanger, and Nicole Perlroth, "Hackers Came but the French were Prepared," May 9, 2017; https://www.nytimes.com/2017/05/09/world/europe/hackers-came-but-the-french-were-prepared.html.

Another possible way to reduce the risk is to ensure that one's cyber espionage implants lack the characteristics that would permit leveraging them for cyber attack. The implant may be placed, say, in a router for the purpose of capturing messages from an internal office system; a cyber attack launched against a router would, at worst, be an inconvenience that lasted no longer than it takes to round up and install a replacement. So, no reasonable inference about a future cyber attack could be made. In practice, making such fine distinctions requires: (1) that the target has systems worth eavesdropping on that can be distinguished from those worth attacking; (2) that the intruder knows which are which; (3) that the target (the network owner) also knows which are which and believes the intruder may want to make that distinction; and (4) that such differences can and will be communicated correctly to the target's leadership. The first condition is clearly not up to the penetrators. The second is an assumption that requires a great deal of prefatory cyber espionage in the first place, reintroducing the very risks of discovery that the strategy was attempting to modulate. The third may require insight into the intruder, since the point is to understand whether the intruder meant simply to spy or to also set up a cyber attack. As for the fourth, one can only guess.

The target's technical experts may point out that a penetration in, say, a well-guarded albeit Internet-linked network is no indication of how well the more critical and hence often air-gapped (i.e., electronically isolated) military systems can survive attack. This is particularly true for a cyber attack whose effectiveness depends on good timing, hence on an ability to exercise real-time command and control over the implants. But might such leaders also remember the same technical experts arguing that these dearly-acquired guards would protect their conversations? And while technical experts may remind leaders of the many caveats that follow all assessments of cyber security, lay-folk often disregarded them or view them as attempts to evade responsibility for being wrong. Leaders may therefore be skeptical of arguments that a penetration here does not mean an attack there. Again, the essential role played by surprise in cyber operations erodes assurances of all sorts.

Lastly, is it in one country's interest to improve another country's *confidence* in the resilience of its armed forces in the face of cyber attack? Success at calming the other side would reduce the risks of overreaction that might follow penetrations into the networks of its military. Confidence makes it easier to dismiss the implications of having found the implants, because the target will conclude that they cannot affect a military force resilient to cyber attack. But feeding such confidence also obviates the value of brandishing one's weapons in cyberspace and vitiates the corresponding deterrence value of one's cyberspace capabilities. Furthermore, unless the argument is generic – we are resilient to such attacks, so you probably are also resilient – demonstrating the resilience of another side's military systems with any credibility

would have to show a level of insight into the details of their systems which would be anything but reassuring.

So, increases in cyber espionage unavoidably create risks if getting caught raises fears.

5. THE DEFENDER'S OPTIONS

Although the target of a discovered intrusion may well infer an imminent attack and raise its alert levels in ways that lead to mutual escalation which culminates in war, nothing *compels* defenders to act that way. Wars are costly and risky and actions such as raising alert levels are not risk-free. The questionable value of running these risks because intrusions *might* be precursors to attack and pre-emption *might* improve the odds of surviving an attack suggests a place for alternative reactions.

A great deal depends on whether such intrusions are an *indicator* of future aggression (specifically, evidence that the odds of physical aggression need to be revised upward) or just an enabler. If an indicator, then countries need to attend to what happens on the ground, so to speak. If an *enabler*, then policies to stop intrusions merit consideration, as they always should.

Warning against further intrusions may bolster deterrence; it signals discovery, displeasure, and, most importantly, that the target takes these intrusions as indicators of potential attack. Although the standard cyber deterrence challenges apply, such as what constitutes an infraction that merits a response and what the response should be, the issue of grandfathering also merits note. Contrast cyber attacks with cyber espionage; if you warn the other side to stop immediately, then later attacks can be assumed to reflect acts of volition that took place *after* the warning; attacks tend to announce themselves at the time. Intrusions, however, do not announce themselves. An intrusion discovered tomorrow may have been carried out yesterday. Thus, being able to time-stamp the last *hostile* volitional activity (not simply the first intrusion) is important in a coherent deterrence posture.

Unfortunately, correct characterization of the intruder's post-warning activity is not trivial, and the problem is worse if the intrusion leaves behind an autonomous implant, one that takes some actions on its own. The intruder can try to erase or deactivate the implant, but then imagine a target's ire in discovering the intruder's post-warning footprints. Even if discovery does not activate reprisals, it could provide a clue as to how the intruder penetrated otherwise inaccessible systems. After all, if the intruder was confident that, even in wartime, it could command and control the intrusion in real-time, then the implanted code would not need autonomous capabilities. Thus,

the existence of such capabilities suggests that the system is hard to access. *Telling* the target about the intrusion so that the target can de-activate it runs into similar problems *and* connotes an obeisance that one rival may not wish to convey to another.

So, unless the target *wants* to build a narrative that would justify fighting the intruder, it needs to exercise forbearance or even forgiveness when it catches what look like violations following a warning.

6. DELIBERATING SIGNALING

Similar issues bedevil using cyber espionage to signal broader intent, in contrast to using it to brandish capabilities. A 2016 study[15] suggests that, if given what they think is the opportunity, policy-makers will try to signal their intentions through cyberspace. In the words of then-CIA-director John Deutsch, they may believe that the "electron is the ultimate precision-guided munition",[16] allowing precision signaling. Or, they may conclude that signaling in cyberspace is far cheaper than moving, say, warships. In one war game examined by the study:

> Strict rules of engagement—to include no network exploitation of strategic command and control and limited military command and control—were placed on computer network exploitation with the assumption that these activities would be detected and would be interpreted as signals of the United States' [lack of] desire to escalate the crisis.

There are two reasons for being skeptical that such signaling would have the desired effects.

One is general to all signaling: there is no guarantee that they will correctly infer what you imply.[17] Some inferences are contrary to fact; for example, that you have forces hidden when in fact you do not. Other inferences are contrary to what you were signaling: you brandish cyber attack capabilities to show how prepared you are, but they think you emphasized non-lethal capabilities because you are afraid to use lethal capabilities. A litany of fairly prosaic reasons can be adduced to explain inaccurate inference, but the simplest is that people make mistakes: they do not see all the evidence or they do not know how to evaluate everything they see. Being busy, as decision-makers typically are, they fail to pay the requisite attention to what they

15 Jacquelyn Schneider, U.S. Naval War College, *Cyber and Crisis Escalation: Insights from Wargaming*, unpublished paper, January 2017.

16 U.S. Senate Committee on Government Affairs on the subject of "Foreign Information Warfare Programs and Capabilities." June 25, 1996.

17 See, for instance, Max Fisher, "Do U.S. Strikes Send a 'Message' to Rivals? There's No Evidence", April 21, 2017; www.nytimes.com/2017/04/21/world/do-us-strikes-send-a-message-to-rivals-theres-no-evidence.html.

do see. Being people, they have confirmation bias: they see what they want to see and when evidence comes along they emphasize their prior perceptions and discard what contradicts it. They themselves may be good evaluators but work for organizations that, collectively, exercise confirmation bias. People also tend to mirror-image: if they see you doing something that they could have done, they may well infer that you are doing it for the same reasons they would have. Leaders with a high regard for their own personal perspicacity (which is reinforced by sycophantic assistants) may rely on their intuition over the painstakingly-generated insights of their intelligence community. Finally, the signal's receivers may be aware of things that signalers are not – and they, in turn, may be aware of things that they think the receivers should have been aware of but were never exposed to. What you see as a signal of yours, they interpret as arising from internal machinations at their end.

Unfortunately for clarity, signalers may have too little idea of what things look like from the perspective of receivers (who, themselves, often take pains to keep others in the dark). Signalers have too little idea of why recipients would think the signal should be read in a certain way. In the end, the signaler may be wrong, but error is beside the point. The reactions of those receiving the signal are entirely determined by facts and circumstances as *they* see them. Neither reality nor what the signaler intended to signal count, if the point is to influence their thinking.

The other set of reasons is specific to cyberspace. Even though cyber espionage may be misinterpreted as preparations for cyber attack, the failure to discover cyber espionage may not necessarily be correctly interpreted as a lack of desire to carry out a cyber attack. Such an interpretation would require that the other side *expects* to find evidence of cyber espionage and then concludes that an absence of a discovery means the absence of activity. It also assumes that they do not find cyber espionage from third parties and erroneously conclude that it came from their potential foes, the most likely guess under the circumstances. They may easily conclude that penetrations carried out *because of the* crisis would not be discovered, because advanced persistent threats even from countries as casual about operational security as China has been can linger undiscovered for several months. Those from more careful penetrators such as Russia or the United States may linger undetected far longer. Even if the penetrators made themselves easy to find in the more benign parts of the other side's network and scarce in the more sensitive areas, the more likely conclusion may be that they took greater pains to be stealthy in the latter case.

Hostile signals – look at us in your system – *should* have a greater fidelity than non-signals. At least there is something to work with. And penetrators should want to take more pains going in than going out, lest they be blocked prematurely. But, to reverse all the cautions noted above, unless the penetration was found where it would clearly

be prefatory to a cyber attack, the other side could interpret their finding as evidence of mere cyber espionage, which may imply nothing out of the ordinary.

Perhaps the difficulty of drawing the correct inferences from discoveries of penetrations in general, or implants in particular, may be eased as cyberwar examples accrete. But would they? While cyberspace is a very dynamic place, few cyber attacks have taken place at nation-state scale, as distinct from cyber espionage and cybercrimes.[18] Thus, by the time enough incidents have accumulated to support conclusions, years may have passed and, more importantly, the world that such incidents describe may have changed so much that earlier evidence is immaterial. The problem is not that the technological basis of computation and communication is so fluid – with the possible exception of what artificial intelligence *might* bring, there is a fair degree of year-to-year stability – but that the interaction between people and markets and between attackers and defenders is constantly evolving. Consider the many ways of creating flooding attacks: volunteers on their own computers, large botnets (involuntarily recruited zombie computers), medium-sized botnets amplified by packet reflection, web servers (e.g., those that support WordPress), cloud servers, and networked devices (e.g., video cameras) – with no guarantee that novel techniques may not be added to the list. The technology behind ransomware was largely available twenty years ago, but did not take off[19] until someone showed that it could work; then many others jumped into the business aided, in part, by the emergence of digital currencies such as Bitcoin. Because measures beget countermeasures which beget counter-countermeasures, techniques may morph rapidly in the hothouse environment that is cyberspace. Meanwhile, other tricks die off. Spam is no longer the problem for consumers that it once was,[20] and changes in Microsoft Windows over the last ten years have complicated any strategy that relies on USB sticks as an infection vector. Correctly interpreting any one penetration against such a dynamic background is difficult.

Speculatively, future years may see a shift from first-order attack methods (the insertion of arbitrary executable code into target systems) to second-order (shaping inputs to yield unexpected outputs in the target system). This could arise because preserving the integrity of a system's code base is a workable problem (e.g., by burning instructions into hardware, if nothing else) while ever-increasing system complexity leads to an exponential increase in the interaction space. Furthermore, the NSA at least (according to the former head of its Tailored Access Office, Rob Joyce[21]) tends to rely on hijacking credentials as much as or more than inserting malware into

18 Notably, system intrusions for the ultimate purpose of getting money, the best example of which was the theft of $81 million from the Bank of Bangladesh, putatively by North Korea (which has also been associated with bitcoin-related theft).

19 For instance, Dan Bilefsky and Yonette Joseph, "Cyberattack in U.K. Hits 16 Health Institutions," May 12, 2017; https://www.nytimes.com/2017/05/12/world/europe/uk-national-health-service-cyberattack.html.

20 "Spam email levels at 12-year low," July 17, 2015; http://www.bbc.com/news/technology-33564016.

21 See his address to the USENEX Enigma 2016 conference: https://www.youtube.com/watch?v=bDJb8WOJYdA.

systems, and hijacked credentials are less useful for cyber attack because the damage you can do with them is limited to the damage that the credential's true owner can carry out. So, credentials may be good enough for tapping the flow of information but not for altering it. If so, the methods used for cyber espionage and cyber attack may diverge, making the world free for cyber espionage.

7. CONCLUSIONS

In a crisis, countries will be looking at indicators of all sorts, not just from within their network. But, as with all things cyberspace, intrusions into networks are likely to garner greater importance over time. As long as the methods of cyber espionage – notably implants – look like the methods of cyber attack, the discovery of one will raise fears about the imminence of the other. Unfortunately for stability, the link between the two is unpredictable. Discovery may or may not happen, but it is more likely to happen in a crisis when systems are being scrubbed more diligently. Figuring out *when* the intrusion took place (the earlier, the more benign) is a forensic art not possessed by all, and without such information the target may assume the worst. The target's reaction, in turn, may be colored by its understanding of the security dilemma in cyberspace. If so, the course of wisdom may be to counter with one's own signals, perhaps deterrent signals. Conversely, signaling through the manipulation of cyber espionage traces likely offers less fidelity than other signaling methods, which themselves have often been misread.

The lesson is to consider what message you want your cyber espionage to carry if and when it is discovered. If you do not want to inflame tensions, double down on operational security, but do not assume success. Thus, also avoid adding military targets to spy on when in crisis, or at least approach them with techniques that look very different from those used to set up cyber attacks. If you are brandishing capabilities or signaling intent, generate a narrative that anticipates discovery. But think this through *beforehand*.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

The Topography of Cyberspace and Its Consequences for Operations*

Brad Bigelow
Principal Technical Advisor
SHAPE DCOS CIS and Cyber Defence
Mons, Belgium
brad.bigelow@shape.nato.int

Abstract: For all the focus on cyberspace as a source of security threats and a domain of military operations, there has been little progress on establishing a consistent approach to describing what constitutes cyberspace. Dozens of definitions of the term "cyberspace" have been developed, but consensus on its essential attributes has yet to be achieved. Similarly, a number of different models have been offered to describe cyberspace in terms of layers, such as the physical, logical and cyber persona layers used in US Joint Publication 3-12, *Cyberspace Operations*. This paper argues that cyberspace as a label for a domain should not be confused with the individual networks – some interconnected ("open") and some relatively isolated ("closed") – involved in military operations. As illustrated by the STEADFAST COBALT exercise, military operations often involve a complex set of networks. The paper then uses the example of the Internet to illustrate the need to take a topographical approach – one that identifies the features of the objects or entities and their structural relationships – to enable effective military operations. This more detailed topographical view of the Internet is used to illustrate how cyberspace considerations relate to existing operational doctrine such as concepts from the operational environment (Joint Operational Area and Area of Interest). Some considerations fit well within this framework. Others require some adaptation, such as shifting some responsibilities to a centralized and persistent function such as the Cyberspace Operations Centre (CyOC) being established by NATO. Others fall outside military control and are better addressed through civil-military cooperation. This example also illustrates how precision in describing the

* The views and opinions expressed in this article are those of the author alone and do not necessarily reflect those of NATO.

composition of cyberspace is essential if military operations in and through cyberspace are to develop into a mature discipline with a solid base of concepts, terminology, techniques, tactics and procedures.

Keywords: *cyberspace, cyberspace operations, cyberspace topography*

1. INTRODUCTION

For all the words that have been written about cyberspace, the lack of a consistent definition and approach to describing it remains one of the biggest obstacles to achieving an effective foundation upon which to advance the state of theory and practice. When the NATO heads of state and government recognized cyberspace as a domain of military operations at the Warsaw Summit in 2016, they managed to do so without actually defining what cyberspace constitutes. While constructive ambiguity might be a useful tool in political negotiations, it becomes an impediment when trying to develop techniques, tactics and procedures for military operations.

The lack of precision in defining what cyberspace comprises undermines the development of effective military responses to its threats and risks because it leads to generalizations that are inaccurate at best and misleading at worst. In a 2015 paper titled "On Cyberwarfare", for example, Fred Schreier postulates five characteristics that make cyberspace unique, including that "the cost of entry into cyberspace is relatively cheap." Because of this, he argues: "The resources and expertise required to enter, exist in, and exploit cyberspace are modest compared to those required for exploiting the land, sea, air, and space domains" (Scheier, 2015). This point about the low cost of entry is often repeated in discussions of cyberspace and its security. For example, the US Army's most recent edition of one of its most basic doctrine publications, Field Manual 3-1, *Operations*, states that:

> Cyberspace is highly vulnerable for several reasons, including ease of access, network and software complexity, lack of security considerations in network design and software development, and inappropriate user activity (US Army, 2017).

The official *NATO Glossary of Terms and Definitions* (AAP-6) does not yet offer a definition of the term "cyberspace". The US Department of Defense issued at least twelve different definitions over the years before issuing its joint doctrine on cyberspace operations in 2013 (Singer, 2014). In its list of cyber definitions, the NATO Cooperative Cyber Defence Centre of Excellence (CCD COE) has collected

29 examples for "cyberspace"– some identical, some similar, some very different (CCD COE, 2017). It is not surprising, then, that as significant a figure as General Michael Hayden, who as Director of the National Security Agency and Director of Central Intelligence was at the center of the initial development of US cyberspace operational capabilities, has written that: "Rarely has something been so important and so talked about with less clarity and less apparent understanding…." (Hayden, 2011).

2. CYBERSPACE, NETWORKS AND CYBERSPACE LITTORALS

One of the basic misunderstandings of cyberspace is the assumption that it is synonymous with the "global grid" of the Internet and public telecommunications networks. By at least three orders of magnitude, the Internet is certainly the largest instance of cyberspace. The Internet Protocol version 6 address space has the capacity to encompass 2^{128} addresses, or something on the order of ten million trillion times the total number of grains of sand on all the beaches in the world. It has also reached many more users than any other network ever developed. It is estimated that, as of mid-2017, over 50% of the world's population are able to access the Internet (World Internet Users and 2017 Population Stats, 2017).

While the Internet is certainly the largest network in cyberspace, it is not the only one. There are still many networks that do not interconnect with the Internet. Closed networks such as classified intelligence, law enforcement and military networks are perhaps the most obvious examples. Others include such closed networks as that operated by the Society for Worldwide Interbank Financial Telecommunication (SWIFT) to provide secure messaging to support international financial transactions. In discussions of the application of international law to military operations in cyberspace, such as the *Tallinn Manual 2.0*, "public, internationally and openly accessible" networks, such as the Internet, are explicitly distinguished from "closed military" networks, in part because this distinction can be important, for example, in determining the appropriate rules of engagement (Schmitt, 2016). Further, as Dror Kenett and his colleagues have written, "In most real-world systems an individual network is one component within a much larger complex multi-level network"– a network within a network of networks (Kenett, et al., 2014).

Each of these networks of networks is an instance of cyberspace. Within a single network there is, at least in principle, the possibility of end-to-end connections: the ability to transfer data, enable transactions, disseminate information, or, from the standpoint of cyberspace operations, create effects. The sum of all the networks that

exist equates to what is referred to as cyberspace in conceptual discussions, but it quickly becomes problematic to make assertions that there are characteristics – such as ease of access – that apply universally across all known networks. Ease of access may be a characteristic of the Internet, but it is certainly not a characteristic of a highly secure network and largely isolated network such as SWIFT.

This distinction between cyberspace as a label for a domain of military operations and individual networks as particular instances of cyberspace is no different from how the term domain has been applied in the context of air, land and maritime operations. While the Earth is wrapped in an atmospheric blanket we refer to as air or aerospace, much of it is divided into airspaces (plural) that are under some level of control – usually national – for such purposes as air safety and national security. Armies concern themselves with land operations, but these must always be tailored to the conditions of a particular location (desert, mountain or jungle). And even the simple distinction between surface and subsurface has profound implications for maritime operations. Indeed, the term "waterspace management" is specifically used for the coordination between submarine and anti-submarine operations.

The need to recognize that cyberspace is more than just the Internet is of critical importance when it comes to planning, organizing and carrying out military operations. In a complex, communications-intensive coalition operation such as that simulated in STEADFAST COBALT – NATO's annual command and control (C2) interoperability exercise – myriad networks, information systems and communications transmission systems are employed. These networks include NATO's unclassified Intranet and its classified network as well as the national equivalents for most of the coalition. The classified networks are then federated through a mission network as a primary interoperability and C2 environment. In addition, the operation will often employ other classified networks handling intelligence or other sensitive data.

The information systems for these operations range from what are termed "core services" – electronic mail, websites, collaboration and office automation – to functional services such as Common Operational Picture and Order of Battle managers. Numerous support applications, such as logistics, movement and spectrum management and external communications tools, such as public affairs, strategic communications and social media, will also be involved. These information systems, along with voice and video traffic, are connected through transmission systems that include both wired and wireless media. Wireless communications span radio frequency bands reaching from VLF (Very Low Frequency) through HF (High Frequency) and VHF (Very High Frequency) to UHF (Ultra High Frequency) and SHF (Super High Frequency). And no military operation today can be carried out without heavy reliance on Positioning, Navigation and Timing (PNT) services such as the Global

Positioning System (GPS), almost entirely carried over portions of a very crowded radio spectrum.

If one looks at the networks at static military facilities, this complexity only increases. The number of networks and information systems in static facilities, as well as the variety of classifications and handling controls of the information they support, typically exceeds that in deployed operations, if only because of the much wider range of functions supported. Some of these are directly connected to the Internet and some are "air-gapped" – isolated from the Internet and other networks through a combination of physical separation, personnel clearances, classification, handling restrictions and encryption. Fewer and fewer military organizations, however, are finding it possible to operate effectively with completely isolated networks, and the pressure to share information is driving them to close the "air gaps" by means of security mechanisms such as guards, gateways, diodes, or encryption, thereby introducing potential vulnerabilities.

Many of the networks, information systems and transmission systems used in deployed operations are anchored through reachback links to these static facilities, which are themselves linked through numerous wide area networks, operating at different levels of classification. Here again, some of these wide area networks are connected to the Internet, directly or indirectly, and some operate over dedicated transmission systems. Because dedicated radio and cable transmission systems tend to play a much smaller role in the interconnection of static facilities than they do in deployed operations, most wide area network connections between static facilities are reliant on commercial leased circuits or tunneled IP services.

Every network also connects to what Paul Withers has termed "cyberspace littorals" – the places where individual instances of cyberspace meet other domains (Withers, 2015). These cyberspace littorals include: the physical infrastructure, including fences, buildings, gates and transportation networks, within which any equipment providing the cyberspace resides; the radio frequency spectrum through which the cyberspace transmissions are carried; the critical infrastructures such as electrical power and water that support the equipment and its supporting personnel; the cyber-physical systems used to control critical infrastructures, force protection systems, industrial systems and even cars and trucks; and finally, the cognitive dimension of decision-making, doctrine, perceptions and even the attitudes shaped through mass and social media.

The term "littoral" should be familiar to military personnel from its use in describing the zone in which the responsibilities of land and maritime forces converge in such operations as amphibious assaults. Applying this term to cyberspace helps to identify those areas in which the responsibilities of cyberspace operators converge with

those of existing military disciplines such as physical security, force protection, area defense, electronic warfare and psychological operations (PSYOPS). It can be useful in better understanding the roles a particular network plays in a military operation and in determining how it can be defended. Indeed, protection of the electromagnetic littoral through spectrum management and electronic countermeasures, for example, can be more critical to the success of a deployed operation that is heavily dependent on radio and satellite communications than any combination of cyber security measures. In the same way, understanding an adversary's cyberspace littorals can help identify effective ways to exploit or disrupt an adversary's use of cyberspace (although this paper does not address offensive considerations).

3. THE TOPOGRAPHY OF ONE INSTANCE OF CYBERSPACE: THE INTERNET

Accurately identifying and understanding the characteristics of any particular network as an instance of cyberspace requires a closer look at its topography – the features of its objects or entities and their structural relationships (Merriam-Webster, 2018). What networks connect to it? Where and how do they connect? How big is it? What types of communications and transactions does it support? And what are the specific features of its littorals? Although the Internet is just one of the networks involved in a military operation, an overview of its topography provides useful insights into how it can be approached in the context of a military operation. It also reveals aspects that military operations are ill-prepared – and arguably ill-suited – to address.

Let us consider, then, the Internet as it might be employed in support of an operation in which a NATO command element and a coalition of forces from NATO and partner nations deploy to an operational theater under the mandate of an operational plan approved by the North Atlantic Council. As with the STEADFAST COBALT exercise, classified networks are still the primary networks employed to support NATO operations. Indeed, for these operations, the reliance on classified networks remains perhaps the single most effective protection against not only conventional military threats, but also threats from the Internet. As standard practice, however, the NATO Unclassified network, which is connected to the Internet through managed gateways hosted in static NATO command structure facilities, is extended to support the NATO command element and eligible parts of national forces. Many nations do much the same, deploying equipment forward to enable access to one or more national networks that are also connected to the Internet.

So, the Internet, the direct and indirect dependencies of his mission on it, and the resulting risks are all considerations for the operational commander. From a

topographical standpoint, every device that can connect to the Internet – directly or indirectly – shares access to a common space defined by an Internet Protocol address (whether version 4 or version 6) and the core Internet link, internet, transport and application protocols (IETF, 1989; IETF, 1989). This is the common plane or elevation (to use a topographical term) on which all Internet-connected devices converge. This is the part of the Internet for which ease of access is indeed its most salient characteristic, and it is understandably the space in which vulnerabilities and attacks that exploit them are most frequently experienced.

As has often been noted, these protocols were designed primarily for fault tolerance and not for trustworthiness or the presence of malicious actors. Consequently, it is also the space where most cyber security efforts are focused. With the growing sophistication of the threats (as one recent Cisco (2016) report puts it: "the time of amateur hackers is long over, and hacking is now an organized crime or state-sponsored event"), however, some in the field of cyber security are arguing that their goal must shift from intrusion prevention to intrusion tolerance – to what has been called the "assume breach" paradigm (Cisco, 2016; Pompon, 2016). While this approach may be new to the Internet, military personnel will recognize it as an example of operating in a contested environment.

Every point of interconnection between information systems supporting military operations and the Internet is a point of exposure to such attacks. Even if such interconnections are minimized or eliminated, these measures do not address the extent to which the Internet has become embedded into most individuals and organizations in the developed world – any of which can, directly or indirectly, represent a dependency for the operation. As Dan Geer has put it, "If […] you are dependent on those who are dependent on the Internet, then so are you" (Geer, 2013).

The risks arising from the use of the Internet in industrial control systems (ICS) and supervisory control and data acquisition (SCADA) systems to manage critical national infrastructure such as electrical power generation and distribution is of growing concern for military operations. Combat and direct support units typically bring their own critical infrastructure in the form of power generation, water treatment, field medical units and other support functions when they deploy. However, this level of autonomy is rare at the reachback command and support facilities to which they are connected, and even the autonomy of deployed units is constrained if this reachback support is disrupted for more than a short time.

If one digs into the Internet below the link layer and looks at the next layers down – what in the Open Systems Interconnection (OSI) model are referred to as the data link layer and the physical layer – ease of access can no longer be taken for granted. Access

to traffic at these layers requires access to the physical transport medium, meaning the radio frequency signal or telecommunication cable carrying the data. It requires the attacker to be within the range of the WiFi access points or to have physical access to the actual cable plant of a local area network or to the cabling carrying traffic across the wide area network through the services of telecommunications providers. The first two – access to WiFi networks and local cable plants – are well within the control of most military commanders. While WiFi vulnerabilities are well known and frequently exploited, so are relatively cheap and effective methods to defend against common threats. However, WiFi availability remains problematic, as WiFi jammers can be easily purchased or manufactured, unless the commander can assure the physical security of all space within jamming range.

Most of the physical transport media carrying Internet wide area traffic, on the other hand, lies outside a commander's control. For short-term deployed operations this is not an issue, because any extension of Internet access to the theater is likely carried over military radio or satellite communications links rather than leased lines. These links are typically protected against a wide range of threats through the use of encryption and anti-jamming mechanisms.

For static facilities, however, the risks arising from dependence on external telecommunications infrastructure are a fact of life, frequently demonstrated through the phenomenon known as "backhoe fade" – damage to underground telecommunications cabling caused by construction equipment. In its *2016 Damage Incident Reporting Tool (DIRT) Analysis and Recommendations Report*, for example, the Common Ground Alliance (2017) reported that nearly 130,000 events (breaks or damage to telecommunications cabling) occurred in the United States and Canada. The potential to exploit or disrupt submarine telecommunications cables is one that has long been known to, and used by, nation states with sufficient technical and operational means (Khazan, 2013).

The Internet also depends on the whole infrastructure of intermediaries involved in any end-to-end communication: foremost, the applications, equipment, facilities and personnel of the Internet Service Providers (ISPs) and Tier 1 (settlement-free interconnection) network providers. The days of the "ISP in the garage" are long past and the vast majority of Internet traffic is carried by a small number of Tier 1 providers. According to the Center for Applied Internet Data Analysis, the top 10 Tier 1 providers support interconnections for over 4.8 billion IPv4 addresses (CAIDA, 2016). In addition, commercial data centers, including those supporting cloud services, have already overtaken the size and capacity of private enterprise on-premise server rooms and data centers, and an increasing number of public and military organizations are shifting applications and services to external data centers and cloud providers.

Finally, this infrastructure also provides much of the intermediary transport media for long-distance telecommunications, which have largely been migrated from circuit-switched to IP transport services.

These providers operate the core physical infrastructure of the Internet that a Belfer Center report recently described as "too connected to fail" – in other words, whose failures could have widespread and potentially global impacts (Snyder, 2017), although these providers have also recognized that high availability and effective physical and personnel security are integral to a viable business model in a highly competitive market. Top-end hyperscale data centers feature security and resiliency measures that equal or exceed those of the most secure military command posts (Branscombe, 2016). These data centers illustrate one of the paradoxes of security on the Internet: while they are protected by many layers of physical security and maintain low profiles to avoid drawing attention to themselves – that is, they fit the profile of a "closed" network facility – many of the services they host are available to anyone with an email address, a valid credit card and access to a device running the essential IP protocol stack – in other words, they host "open" services.

Moving up from the core IP protocol layers of the Internet, one encounters the diverse set of software applications – core and functional services – that play a role in a military operation. Here again, ease of access varies widely and should not be taken as a "one size fits all" measure. For those applications that are available as open source or commercial off-the-shelf, the attack surface and the potential threats tend to be closely related: the more people using an application, the better the chance that attacks have been developed to exploit their vulnerabilities. For the many custom-developed applications employed in military operations, on the other hand, access to source or executable code, development and test documentation, and especially operationally relevant data, is much more limited. However, the simple cost of developing custom military software applications tends to prevent rigorous vulnerability testing.

Finally, moving up from the applications layer in the Internet, one leaves the man-made technical environment and enters what Withers calls the cognitive dimension: decision-making, doctrine, norms, perceptions and attitudes. This is easily the most complex dimension, but it is also not a new consideration for military operations. What is new is the role the Internet plays in enabling access to the cognitive dimension, both through new applications such as social media and streaming video and through new outlets for old applications such as electronic mail, chat, news reporting and psychological operations.

Even in the complex cognitive dimension, however, ease of access is neither universal nor something that can safely be taken for granted. At the simplest level, language is

still an effective barrier to entry. English might be the predominant language on the Internet, but it still ranks behind Mandarin and Spanish in number of native speakers. Context is another: although spearphishing still succeeds in fooling some users to click on links in untrustworthy emails, it would be much more difficult to convince a military operator to trust an email pretending to be a fragmentary order (FRAGO), if only because such communications are usually confined to military message handling systems. Finally, just because there is content on the Internet, it does not mean that anyone is looking at it. With over 1.3 billion websites alone, let alone social media services aimed at mobile users, there are a lot of opportunities to miss the audience.

Revelations about Russian manipulation of social media and its role in the 2016 US presidential election have certainly demonstrated how effective social media can be in advancing state aims. A recent report from Freedom House stated that: "Online manipulation and disinformation tactics played an important role in elections in at least 18 countries over the past year" (Freedom House, 2017). Skillfully positioned and executed, social media can be highly effective. Just six Facebook pages intended by Russian operators to sway US voter perceptions stimulated over 18 million interactions with other Facebook users before being shut down (McCarthy, 2017). As Michael Schmitt, editor of the *Tallinn Manual* and *Tallinn Manual 2.0*, has written, the Russian example illustrates the potential for states to exploit "grey zones" – areas where "international law principles and rules… are poorly demarcated or are subject to competing interpretations" (Schmitt, 2017).

4. THE INTERNET AND THE OPERATIONAL ENVIRONMENT

Part of the task of integrating cyberspace as a domain of military operations is that of fitting into an existing framework of operational doctrine. One aspect of this doctrine is that of the operational environment. NATO's basic doctrine for military operations, AJP-3(B), *Allied Joint Doctrine for the Conduct of Operations*, sets out the operational environment in terms of areas and boundaries. In particular, the Joint Operational Area (JOA) is defined as the "temporary area defined by the Supreme Allied Commander, Europe (SACEUR), in which a designated joint force commander plans and executes a specific mission at the operational level" (NATO, 2011 p. 1-23). While AJP-3(B) recognizes that "the operational environment is expanding, becoming more dispersed and non-linear", the intent of the definition of the JOA remains to ensure that all elements of a joint force "have a common understanding of its principal boundaries" (NATO, 2011 p. 1-22).

AJP-3(B) also establishes the concept of an Area of Interest (AOI), which it defines as "the area of concern to a commander relative to the objectives of current or planned operations, including his areas of influence, operations and/or responsibility, and areas adjacent thereto" (NATO, 2011 p. 1-23). These operational environment constructs have traditionally been defined in geographic terms and are intended to help the commander and operational planners to bound the area within which forces are employed and effects achieved. The operational environment also helps delineate the boundaries of command and control authorities and the rules of engagement.

Taking the topographical overview of the Internet as it relates to a NATO operation as above, there are aspects that fit well within the existing concept of the operational environment. The actual equipment used to access these Internet-connected networks and the troops supporting it in the operational theater – the cyber boots on the ground – clearly fall within the JOA. The equipment is an asset that must be protected as any other physical asset belonging to the forces in theater, and the troops are under the joint force commander's force protection responsibilities. In the same manner, the joint force commander would be expected to exercise operational control to ensure the availability, confidentiality and integrity of the information processed by these assets, whether against kinetic weapons, electronic warfare capabilities or cyber effects. This responsibility also extends to the data link and physical layers described above, so cabling and WiFi signals must be protected as well.

Interconnection to the Internet, however, is a primary reason for deploying this equipment to the theatre, and the gateways in the reachback facilities that provide those interconnections likely fall outside the geographical boundaries of the JOA. These anchor points and gateways may also fall outside the joint force commander's direct operational control. NATO is not alone in assigning the responsibility to run the information systems and networks supporting static military facilities to a civilian organization outside a direct military chain of command. For these reasons, the command and control (C2) arrangements between the joint force commander and the organization(s) providing his reachback support can be complicated and problematic. The commercial service providers responsible for the interconnections between these gateways are certainly both outside the JOA and outside the commander's operational control, as are the vast number of Internet users, devices, applications, data and services and the physical infrastructure supporting them that lie on the other side of the NATO and national static gateways. This also applies to most, if not all, of the Internet-connected critical infrastructures that might be supporting the operation of the static command and support facilities.

Given the prevalence of threats against the Internet and the networks that interconnect with it, it should also be clear that all of these aspects fall within what NATO doctrine

would consider the joint force commander's AOI. Each presents a greater or lesser risk to the success of the operation. Understanding and managing such risks, however, presents a significant challenge for a deployed commander. The already difficult task of situational awareness in cyberspace is further complicated by limitations on bandwidth to the theater and on the tools and expertise of the analysts in theater.

This is one reason why NATO, following the example of numerous nations, is centralizing its support for cyberspace situational awareness and operational planning support in the Cyberspace Operations Center (CyOC). It is far more effective to concentrate the technical, intelligence and operational expertise required for a credible cyberspace situational awareness capability than to attempt to replicate them in one or more operational theaters. However it is organized, this capability – even given the limitations of existing tools, models and data sources – is essential for effective military operations. Another reason is that Internet threats and their risks to operations often arise outside the JOA, not just in terms of geographical boundaries but also in terms of timeframe. Indeed, some of the most significant risks arising from the Internet are those we refer to as advanced persistent threats. Establishing a centralized and persistent situational awareness, planning and coordination capability is perhaps the single most important way in which existing NATO operational doctrine is being adapted to accommodate the unique aspects of cyberspace as a domain.

The delineation of the operational environment geometry also needs to extend to the littorals of the Internet-connected networks supporting an operation. Protection against physical and electronic threats has already been mentioned and is generally within the scope of established capabilities. Likewise, long-standing military practices developed well before the rise of the Internet, such as the use of radio silence, minimize, visual signaling and operational security (OPSEC), can still be of use to mitigate or avoid risks presented by Internet-based threats.

The cognitive dimension, however, still presents challenges. Clearly within the JOA and the commander's operational control are the troops in theater: their decisions, perceptions and actions, and how they communicate them, including over the Internet, are his responsibility. In the same way, he is responsible for how the joint force influences the perceptions of the adversary and affected populations, which is why psychological operations, information operations and strategic communications are integral to military operations. The Internet represents both a medium for conveying his messages and for assessing perceptions among targeted audiences.

As the examples of state-sponsored manipulation of social media demonstrate, however, Internet-based threats are emerging that are difficult to fit into the traditional concept of the operational environment geometry. Indeed, it could be

argued that military operations are not the appropriate mechanisms to target what are purely civilian objects (Harrison-Dinniss, 2015); but the key problem in applying the operational environment geometry is that these threats currently fall into what Schmitt calls the "grey zone," where boundaries of operational control are informed and guided by international law. As Schmitt has put it: "The brighter the redlines of international law as applied to cyber activities, the less opportunity states will have to exploit grey zones in ways that create instability." (Schmitt, 2017) And the easier it will be to delineate how to draw the lines of military responsibility and interest.

The closed networks required to support an operation tend to have far fewer cyber defence considerations for a commander than the Internet. The example of the Internet's topography is offered, however, to illustrate that it is certainly possible to sort these considerations into three rough categories: those within the JOA and under operational control; those within the AOI and within some level of control, if indirect; and those that fall well outside both military authority and the means of any commander to control. By sorting the cyberspace considerations for an operation into these three categories, commanders can begin to identify where effective military response options exist and where they do not.

Those considerations that are within the JOA and within the commander's operational control are those for which existing doctrine is most suitable. Considerations in this category must clearly take first priority for operational planning and situational awareness. This is the area where planners need most to be informed by intelligence about the physical, electronic and cyber threats to be expected in theater. This is also where the commander needs to assess the value of such tried and true practices as the use of radio silence, alternate communications and minimize to mitigate or avoid the risks these threats might present. Finally, this is where the protection – or vulnerability – of cyberspace littorals can have the greatest direct impact on the operation.

The next category covers those considerations that are within the commander's AOI and within some type of C2 arrangement, however problematic. From a planning standpoint, considerations in this category are better addressed by a central and strategically-placed function such as the CyOC for the reasons noted above: theater-based limitations (bandwidth, tools and personnel) and the fact that many of these considerations derive from conditions that are persistent and not tightly coupled to the specifics of the operation, and which likely span multiple operations.

The third category covers those that are within the AOI but outside operational control, even via C2 arrangements. Most of these considerations, such as the protection of critical infrastructures, the security of Tier 1 Internet providers and hyperscale data centers, and state manipulation of social media and other examples of what Schmitt

terms the "grey zone" are wholly outside the military span of control. These challenges can only be addressed through political, diplomatic, legal or regulatory channels. Such liaison falls well outside the current scope of Civil-Military Co-operation (CIMIC), which is typically focused on liaison between the joint force commander and civilian authorities in theater. Another important adaptation of existing doctrine to accommodate cyberspace may be in developing persistent versions of CIMIC between centralized military capabilities like the CyOC and their civil counterparts.

5. CONCLUSIONS

There has been no shortage of sweeping generalizations in much that has been written on cyberspace operations and cyber security. As NATO and national militaries work to establish cyberspace as an operational domain, precision is essential to developing a mature discipline with a solid base of concepts, terminology, techniques, tactics and procedures. One such precision is to recognize that operations in the domain of cyberspace always involve specific networks of networks, of which the Internet is only one. Another is to recognize that the characteristics, threats and risks associated with any particular network vary depending on which aspect of its topography is considered. The ease of access that exists on one plane or elevation, such as the common core set of Internet Protocols, might not characterize another, such as that of submarine telecommunications cables.

This precision is also important to integrating cyberspace into existing doctrine. Cyberspace considerations that fit well within existing constructs such as the JOA and operational control can, for the most part, be addressed by the operational commander in theater. Others are better addressed by a central cyberspace operational planning and situational awareness function such as the CyOC being established in NATO. Finally, there are considerations that either fall clearly outside the scope of military control, or for which such demarcation is still difficult. For these, effective mechanisms for civil-military co-operation need to be established. Such a framework can channel efforts in a practical way and help speed the process not only of implementing cyberspace as an operational domain but of better defending the Alliance against the threats arising from the Internet and the other networks it depends upon.

REFERENCES

Branscombe, M. (2016, November 2). *Inside a hyperscale data center (how different is it?)*. Retrieved from CIO.com: https://www.cio.com/article/3137719/data-center/inside-a-hyperscale-data-center-how-different-is-it.html.

CAIDA. (2016, September 1). *AS Ranking*. Retrieved from Center for Applied Internet Data Analysis (CAIDA): http://as-rank.caida.org/.

CCDCOE. (2017, December 16). *Cyber Definitions*. Retrieved from NATO Cooperative Cyber Defence Centre of Excellence: https://ccdcoe.org/cyber-definitions.html.

Cisco. (2016). *Cisco Global Cloud Index: Forecast and Methodology, 2015–2020*. Retrieved from Cisco.com: https://www.cisco.com/c/en/us/solutions/service-provider/global-cloud-index-gci/white-paper-listing.html.

Common Ground Alliance. (2017, August 11). *2016 Damage Incident Reporting Tool (DIRT) Analysis and Recommendations Report*. Retrieved from Common Ground Alliance: http://commongroundalliance.com/media-reports/dirt-report-2016.

Freedom House. (2017, November). *Freedom on the Net 2017: Manipulating Social Media to Undermine Democracy*. Retrieved from Freedom House: https://freedomhouse.org/report/freedom-net/freedom-net-2017.

Geer, D. (2013, April 26). Resolved: the Internet Is No Place for Critical Infrastructure. *ACMQueue, 11*(4). Retrieved January 1, 2018, from http://queue.acm.org/detail.cfm?id=2479677.

Harrison-Dinniss, H. (2015, March). The Nature of Objects: Targeting Networks and the Challenge of Defining Cyber Military Objectives. *Israel Law Review, 48*(1), 39-54.

Hayden, M. V. (2011, Spring). The Future of Things Cyber. *Strategic Studies Quarterly, 5*(1), 3-7.

IETF. (1989, October). *RFC (Request for Comments) 1122: Requirements for Internet Hosts - Communication Layers*. Retrieved from Internet Engineering Task Force (IETF): https://tools.ietf.org/html/rfc1122.

IETF. (1989, October). *RFC (Request for Comments) 1123: Requirements for Internet Hosts - Application and Support*. Retrieved from Internet Engineering Task Force (IETF): https://tools.ietf.org/html/rfc1123.

Kenett, D. Y., et al. (2014). Network of Interdependent Networks: Overview of Theory and Applications. In G. D'Agostino, et al. (*SCALA, Network of Networks: The Last Frontier of Complexity* (pp. 3-13)). Cham: Springer International Publishing.

Khazan, O. (2013, July 16). The Creepy, Long-Standing Practice of Undersea Cable Tapping. *The Atlantic*. Retrieved January 1, 2018, from https://www.theatlantic.com/international/archive/2013/07/the-creepy-long-standing-practice-of-undersea-cable-tapping/277855/.

McCarthy, T. (2017, October 14). *How Russia used social media to divide Americans*. Retrieved from The Guardian: https://www.theguardian.com/us-news/2017/oct/14/russia-us-politics-social-media-facebook.

Merriam-Webster. (2018, January 1). *Topography - definition of topography*. Retrieved from Merriam-Webster: https://www.merriam-webster.com/dictionary/topography.

NATO. (2011, March 16). *Allied Joint Publication (AJP) 3(B), Allied Joint Doctrine for the Conduct of Operations*. Retrieved from NATO Standardization Office: http://nso.nato.int/nso/zPublic/ap/ajp-3(b).pdf.

Pompon, R. (2016). *IT Security Risk Control Management: An Audit Preparation Plan*. New York City, NY, USA: Apress.

Scheier, F. (2015). *On Cyberwarfare (DCAF Horizon 2015 Working Paper No. 7)*. Retrieved from Geneva Centre for the Democratic Control of Armed Forces (DCAF): https://www.dcaf.ch/cyberwarfare.

Schmitt, M. N. (Ed.). (2016). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. New York City, NY, USA: Cambridge University Press.

Schmitt, M. N. (2017, August 8). *Grey Zones in the International Law of Cyberspace (2017 James Crawford Lecture on International Law)*. Retrieved from https://ore.exeter.ac.uk/repository/handle/10871/27563.

Singer, P. W. (2014). *Cybersecurity and Cyberwar: What Everyone Needs to Know*. New York City, NY: Oxford University Press.

Snyder, C. (2017). *Too Connected to Fail*. Harvard Kennedy School, Cyber Security Project. Cambridge, MA, USA: Belfer Center. Retrieved January 1, 2018, from https://www.belfercenter.org/publication/too-connected-fail.

US Army. (2017). *Field Manual 3-0, C1: Operations*. Washington, DC, USA.

Withers, P. (2015, Spring). What is the Utility of the Fifth Domain? *Air Power Review, 18*(1), 126-150.

World Internet Users and 2017 Population Stats. (2017, June 30). Retrieved December 2017, from Internet World Stats: http://www.internetworldstats.com/stats.htm.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Net Neutrality in the Context of Cyber Warfare

Kim Hartmann
Conflict Studies Research Centre
Oxford, United Kingdom
kim.hartmann@conflictstudies.org.uk

Keir Giles
Conflict Studies Research Centre
Oxford, United Kingdom
keir.giles@conflictstudies.org.uk

Abstract: Real or potential connections between infrastructure of different security levels, from relatively unprotected individual users up to interfaces with critical national infrastructure, have made cyberspace a highly contested and congested domain. But operating conditions within this domain strongly favour malicious actors over legitimate operators seeking to provide security and protect systems and information. Technical capabilities to establish dominance and cause damage in this domain are widely distributed, but legal and ethical constraints prevent legitimate actors from using them to their full potential.

Within this context, net neutrality presents a limiting factor on the capability of legitimate actors to respond to harmful activity in cyberspace whose common aim is to install and uphold a technical imbalance. Under the principle of net neutrality, each data packet must be transmitted with equal priority, irrespective of its source, destination, content or purpose. This is disadvantageous to cyber defence. Comparisons to jungle or arctic warfare, where operating conditions are neutral and degrade the performance of each combatant side equally, are invalid, as malicious operators are capable of technically manipulating data traffic to their favour. While both malicious and legitimate actors may have comparable capabilities, legitimate actors are bound to legal and political restrictions, making them immobile in several cyber warfare scenarios. Transferring the principles of net neutrality to real life scenarios corresponds to depriving military, police and emergency operators from any privilege that allows them to respond to an incident – in effect, depriving them of their blue lights and emergency powers even in severe incidents targeting critical infrastructure that may threaten civilian lives.

This paper investigates the potential opportunities and challenges of an adjustment to the principle of net neutrality to facilitate defensive action by legitimate actors; how this adjustment could contribute to regaining control in congested cyber domains

in the case of national or international cyber incidents; and the risks associated. The different ways of dealing with net neutrality in cyber defence situations in the EU, UK and Russia are compared. Particular focus is put on the organisations and capabilities needed to establish technical sovereignty in multi-domain networks, including consideration of the acceptability of outsourcing the task of upholding cyber sovereignty to external institutions.

Keywords: *net neutrality, cyber defence, cyber security, net regulation*

1. INTRODUCTION

The long-running debate over net neutrality gained unprecedented prominence in public attention during the autumn of 2017 as United States Communications Commission (FCC) chairman Ajit Pai proposed the repeal of policies dating from 2015 that safeguarded net neutrality in the US. The public discussion on net neutrality was primarily concerned with potential abuse and the prospect of forming and protecting positions within specific markets such as the telecommunications sector; a situation exacerbated in the United States in particular by limited consumer choice resulting from a small number of major telecommunications companies already enjoying near-monopoly status.[1] This threat would not only affect the telecommunications market and its service providers, but also any other market or services depending on communication through Internet Service Providers (ISPs) – in effect, any area of modern business. The most prominent and intensively discussed examples of services which faced severe disruption were social media and streaming platforms, both of which derive clear benefits from neutral treatment of Internet traffic because of their data-heavy nature and vulnerability to any increase in the cost of data transfer.

It is likely that the involvement of these platforms in the debate, augmented by their substantial presence in everyday civil life, ignited the mainly emotion-driven debate on the 'freedom of the Internet'. This topic rapidly eclipsed the technical aspects of net neutrality overhaul. Comparisons were often made to regulations on water and electricity prices. The suggestion that Internet access is an essential service, and therefore should be protected from open market forces, illustrated how net neutrality discussions focus on matters of principle while neglecting technical aspects that challenge a universally connected, digital society.

The concept of net neutrality has predominantly been associated with constraining ISPs from throttling transmission rates and limiting Internet access for end-users. However,

1 Brian Fung, 'FCC plan would give Internet providers power to choose the sites customers see and use', *Washington Post*, November 21, 2017.

this article will consider how net neutrality influences the way data is transferred in cyberspace in a number of other ways. The abolition of net neutrality principles in one country or more provides both opportunities and challenges, affecting the nature of both offensive and defensive computer network operations (CNO) during peacetime as well as overt hostilities.

Real or potential connections between infrastructures of different security levels are established through networking devices and the individuals or organisations that own them. Security levels ranging from relatively unprotected Internet of Things (IoT) appliances, through individual user devices and interfaces up to critical national infrastructure may easily and unnoticeably become interconnected, rendering cyberspace a highly contested and congested domain. But operating conditions within this domain strongly favour malicious actors over legitimate operators, especially as security standards may be legally binding but not technically enforced. This is also observable for net neutrality principles: it is common practice to provide an equal level of Internet service availability to end-users by ISPs and legislation may require compliance with according policies, but there is no technical enforcement. Consequently, malicious actors can abuse net neutrality principles through different attack vectors and use it to hide their actions. While the technical capabilities to establish dominance in the cyber domain are widely distributed, legal and ethical constraints prevent legitimate actors from utilising them to their full potential.

A key common aspect to many CNO attacks is establishing, maintaining and protecting privileged access to systems or processes. Cyber attacks can seek to establish an imbalance between the attacker and the defenders in terms of prioritised access to data, components or networks. As such, net neutrality presents a limiting factor on the capability of legitimate actors to respond to harmful activity in cyberspace. Under the principle of net neutrality, each data packet should be transmitted with equal priority, irrespective of its source, destination, content or purpose. This means that cyber defence, or responses to critical incidents, will not receive any prioritisation over 'normal' traffic, and consequently present an advantage to an attacker seeking to isolate the target of the attack. However, the ability to respond to cyber attacks from any location is crucial to efforts by NATO member states to set up cyber defence units capable of cooperating in live cyber operations.[2] Officials must be aware that net neutrality principles may compromise this effort unless other methods are established to uphold cyber dominance among allies. Examples of such alternative methods may range from dedicated private networks, through hidden network entry points, to organisational and administrative measures.

In effect, interdiction of remote cyber defence efforts by an attacker poses an analogous problem in cyberspace to hostile actors seeking to isolate areas of planned operations

2 NATO, 'Cyber Defence', December 14, 2017, https://www.nato.int/cps/en/natohq/topics_78170.htm.

by means of advanced anti-access and area denial (A2AD) systems, preventing access by NATO reinforcements seeking to defend them. But while in air, sea or land operations, friendly forces can take advance steps to ensure privileged access in time of crisis,[3] in cyberspace the principles of net neutrality prevent any such pre-emption.

While communications transferred through separate networks independent of civilian ISPs are unlikely to be affected (such as would be expected in military operations), CNO against critical infrastructure and cyber espionage have already been conducted through the public Internet, open for access to all.[4] With critical infrastructure a likely target in cyber warfare, legitimate cyber actors must be capable of effectively and remotely counteracting sophisticated cyber attacks.[5] This remote access to attacked network components could be enabled by physically separate communication lines as physical backdoors to the network (economically unfeasible in almost all cases) or allowing data traffic to be tunnelled. However, the latter does not guarantee that communication is possible in a congested domain as components and routes may be inoperative or compromised. Prioritising traffic through ISPs, by contrast, could allow network administrators to identify the tunnelled communication and install in advance packet-based rules that enable critical communication even during attacks.

Comparisons to jungle or arctic warfare, where operating conditions are neutral and degrade the performance of each combatant side equally, are invalid since the operating conditions in cyberspace can be adapted by one side or the other. Skilled cyber actors are capable of ensuring that their data traffic is prioritised or that the opponent's traffic is downgraded or blocked. Additionally, the opponent in a cyber warfare scenario may not only target military components but also potentially attack civilian critical infrastructures, forcing governments to respond immediately to ensure the safety of their citizens and prevention of crippling or catastrophic damage. Therefore, transferring the principles of net neutrality to real life scenarios would rather correspond to depriving military, police and emergency operators of any privilege that allows them to respond promptly to an incident – in effect, taking away their blue lights and emergency powers even in military operations or severe incidents targeting critical infrastructure that may threaten civilian lives. A more appropriate analogy would be a car chase where criminals can run red lights and set up roadblocks, but the police must still observe traffic rules and speed limits.

Net neutrality is currently not technically enforced, nor has it ever been. There are no central authorities capable of monitoring and enforcing net neutrality on global networks. Additionally, even when legislation demands the enforcement of net neutrality policies, no guarantees can be given once traffic is routed outside national

3 Daniel Fiott, 'Towards a "military Schengen"?', EU Institute for Security Studies, November 2017, https://www.iss.europa.eu/sites/default/files/EUISSFiles/Brief%2031%20Military%20Schengen.pdf.

4 National Cyber Security Centre (NCSC), last access: January 7, 2018, https://www.ncsc.gov.uk/index/alerts-and-advisories.

5 Thomas A. Johnsson (Ed), *Cybersecurity: Protecting Critical Infrastructures from Cyber Attack and Cyber Warfare*, 1st Edition, CRC Press, April 16, 2015, ISBN: 978-1482239225.

borders. The management of data traffic has always been the responsibility of telecommunication organisations and network administrators. Routing rules based on packet origins, content, frequency and general network load are common practice in most networks. This has not been a problem as long as fast communication appeared cheap and unlimited, and large-scale cyber attacks remained the preserve of science-fiction novels or far-fetched 'cyber Pearl Harbor' predictions. While some corporate entities may very plausibly have the intention of abusing the new regulatory situation in the United States for financial benefit, there is also a need for a rational and problem-oriented discussion on how to handle network traffic management in the future with the rising challenges of cyber warfare in mind.

Hence the remainder of this paper investigates the potential opportunities and challenges of an adjustment to the principle of net neutrality to facilitate defensive action by legitimate actors; how adjustments may allow actors to gain control in congested cyber domains in the case of national or international cyber incidents; and risks associated with weakening of net neutrality principles. The different ways of dealing with net neutrality in the EU, UK and Russia are considered. Particular focus is put on the organisations and capabilities needed to establish technical sovereignty in multi-domain networks, including consideration of the acceptability of outsourcing the task of upholding cyber sovereignty to external institutions.

2. NET NEUTRALITY IN THE EU, UK AND RUSSIA

This section explores principles under which ISPs may legitimately interfere with network traffic by technical means in order to illustrate the opportunities and challenges of weakening net neutrality overall. Three different regulatory environments (the EU, UK and Russia) are compared to illustrate the wide variations in philosophy and enforcement between different jurisdictions.

A. Net Neutrality

In simplistic terms, net neutrality means that network providers must treat all network traffic equally and may not interfere with data traffic in a way that affects the traffic of selected parties only. Net neutrality is a set of principles, not a technical implementation. In fact, due to the need of modern networks to be able to cope with data transmission errors and delays, most communication protocols are designed to deal with limitations without end-users noticing. In other words, their design renders them capable of hiding net neutrality violations. This is part of what opens network communications to abuse in hidden cyber operations and creates the huge imbalance between legitimate actors bound to net neutrality on the one hand, and malicious actors with no effective constraint by the rule of law on the other.

Computer networks consist of components, which in turn have physical and logical entities, all of which can communicate between themselves. In order to be able to connect components of completely different architectures, purposes, languages and communication types, the ISO OSI standard was developed.[6] This is a conceptual model that defines how 'data' is organised and communicated on different abstraction layers, moving from physical representations to logical units. An ISP provides the core physical components within a network[7] and as a result has access to the complete OSI stack. ISPs are capable of interfering with traffic on any layer: cutting the physical connection, dropping packets, filtering for services and (unencrypted) content in data, and more.

Net neutrality advocates have been concerned with ISPs throttling down transmission rates, while their opponents put forward counter-arguments of innovation of better bandwidth distribution techniques and networking technologies that are incompatible with net neutrality principles. It is currently impossible to predict how ISPs will handle traffic in the future if net neutrality principles are weakened, but the status quo leads to educated guesses on future network management techniques, such as:

- The pure 'throttling' of data traffic based on origin or destination is commonly associated with dropping packets. By dropping packets, the quality of the single connection may go down, while the overall bandwidth is improved: the ISP regains some of its bandwidth by not servicing one of its customers.
- Another way of gaining bandwidth is by queuing packets. Packets are not 'lost' but take longer to be delivered as they are not forwarded immediately. Again, the ISP gains bandwidth by reducing processing time.

Selection of which traffic to interfere with may be based on packet, service or content information. Depending on the type of information chosen, the interference is performed on different layers of the network stack and may require additional methods such as deep packet inspection (DPI). DPI has been associated particularly with Internet censorship,[8] but is also a common tool for cyber forensics and network administration.

However, methods that alter bandwidth distribution merely by dropping or queuing are not suitable to guarantee privileged data transmissions for selected customers or services, as solutions exist to avoid dropping, queuing and DPI. The most prominent example known to be adopted to avoid censorship (which is usually also based on these methods) is the use of virtual private networks in combination with so-called

6 Andrew S. Tanenbaum, David J. Wetherall, *Computer Networks*, 5th Edition, Pearson, January 9, 2010, ISBN: 978-9332518742.

7 Barry Raveendran Greene, Philip Smith, *Cisco ISP Essentials*, Cisco Press – Networking Technology Series, April 16, 2002, ISBN: 978-1587050411.

8 Ralf Bendrath. 'Global technology trends and national regulation: Explaining Variation in the Governance of Deep Packet Inspection.' International Studies Association Annual Convention. Vol. 15. No. 18. 2009.

'onion routing' networks such as the Tor network.[9] It is therefore more than likely that alternative methods will be used.

The relevance to cyber warfare lies in the fact that, in addition to simple destructive potential, cyber attacks commonly serve the purpose either of gathering information or of exerting power through the medium of the Internet. This can be through achieving and demonstrating interdiction or malfunctioning of networks and their associated services. While current attacks tend to aim at specific network components, it is likely that future attacks will be directed against bandwidth distribution technologies.

Several already-common attack types include methods that abuse net neutrality principles to ensure a larger portion of bandwidth is available to the attacker. This provides a number of secondary effects for any botnet or distributed attack. It allows an attacker to undertake further activities in parallel, unaffected by the ongoing attack itself; it demonstrates power in the domain; it creates an impression of omnipresence of the attacker; it hijacks the bandwidth of legitimate actors; it disables the attacked components; and finally, and most significantly for the current discussion, it hampers external interference by legitimate cyber defence actors as the attacked components may become inaccessible.

B. EU

In September 2013, the European Commission published a draft set of regulations for the telecommunications single market. This draft was heavily criticised for not sufficiently addressing net neutrality regulations and for introducing differentiation between 'communications access' and 'specialised services access' without specifying these services adequately. The draft was adjusted and approved by the EU parliament in April 2014.[10]

The adjusted draft specifically declares that Internet service access:

> 'means a publicly available electronic communications service that provides connectivity to the Internet in accordance with the principle of net neutrality, and thereby connectivity between virtually all end points of the Internet, irrespective of the network technology or terminal equipment used'.[11]

9 The Tor project, https://www.torproject.org/, last access: January 8, 2018. See also McCoy, D., Bauer, K., Grunwald, D., Kohno, T. & Sicker, D. 'Shining light in dark places: Understanding the Tor network'. In *Proceedings of the 8th International Symposium on Privacy Enhancing Technologies* (pp. 63-76). Springer, Berlin, Heidelberg, July 2008.

10 EU Parliament, 'Draft on the proposal for a regulation of the European Parliament and of the Council laying down measures concerning the European single market for electronic communications and to achieve a Connected Continent, and amending Directives 2002/20/EC, 2002/22/EC, and Regulations (EC) No 1211/2009 and (EU) No 531/2012', March 20, 2014.

11 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0281+0+DOC+XML+V0//EN.

Net neutrality is defined as the:

> 'principle according to which all Internet traffic is treated equally, without discrimination, restriction or interference, independently of its sender, recipient, type, content, device, service or application'.[12]

Specialised services are allowed for that are:

> 'provided over logically distinct capacity, relying on strict admission control, offering functionality requiring enhanced quality from end to end, and that is not marketed or usable as a substitute for Internet access service'.[13]

In other words, specialised services are considered as supplementary offers to Internet access services. Examples of such services could be real time applications, sensory data aggregations or distributed computing services.

Following heated discussion, the 2014 draft was further adjusted and approved in November 2015 as EU Regulation 2015/2120.[14] The guidelines for implementation of the April 2014 draft no longer included the term 'net neutrality'. ISPs are still required to follow the 'best effort' principle, requiring all packets to be treated equally (in other words, a core aspect of net neutrality). However, permission for 'zero rating' and a specification of 'sufficient data traffic management' methods have both been criticised. Although violations of net neutrality principles through the use of DPI is possible, several ISPs in EU states are known to use DPI in varying contexts. DPI is known to be carried out by governments and their legitimate actors. The inspection results are used for further processing, prosecution and surveillance.

Zero rating refers to the practice of not imposing additional costs for access to selected online services, while all others incur such charges. The application of this approach varies widely across Europe. The Netherlands enforced a strict net neutrality policy, but at the other extreme, in Portugal ISPs offer a strictly limited connection service with additional charges for access to a wide range of common applications.[15] These charges are usually in the form of purchasing specific packages, named for example 'social' or 'music', which include services selected by the ISP; the criteria

12 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML%2BAMD%2BA8-2015-0300%2B014-024%2BDOC%2BPDF%2BV0%2F%2FEN.

13 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0281+0+DOC+XML+V0//EN Chapter 1, Article 2 (15) in reference 10.

14 Official Journal of the European Union, Regulation 2015/2120 of The European Parliament and Council, November 25, 2015, http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32015R2120&rid=2, last access: January 8, 2018.

15 *The Guardian*, 'Net neutrality enshrined in Dutch law', June 23, 2011, https://www.theguardian.com/technology/2011/jun/23/netherlands-enshrines-net-neutrality-law, last access: January 7, 2018; Alex Hern, 'Net neutrality: why are Americans so worried about it being scrapped?', *The Guardian*, 22 November 22, 2017, https://www.theguardian.com/technology/2017/nov/22/net-neutrality-internet-why-americans-so-worried-about-it-being-scrapped, last access: January 7, 2018.

for selection include the profitability of each service, since the service providers pay a sum to the ISP for inclusion. The 'social' package can, therefore, include Facebook and WhatsApp, while all other social media platforms are not available. The less profitable services cannot be blocked by the ISP, since this would clearly violate net neutrality principles, but they can be de facto excluded by pricing policies. This could, for example, take the form of imposing an indirect cost penalty on users of Telegram by ensuring that data transferred via that app counts against the user's strictly limited 'free' quota, while WhatsApp data has a much higher limit as part of a package.

The 'traffic management' stipulation means that ISPs may adjust data flow rates, for instance to avoid service disruption due to traffic overload. ISPs are reported to have throttled throughput during evening hours (when most customers use their streaming services) to 'encourage' users to stagger demand. Consequently, instead of all customers starting to stream video at, for example, 8 p.m. they do so earlier or later. This allows the ISP to avoid specific traffic peaks, and therefore economise on investment in new hardware that would otherwise be necessary only during a once-a-day data throughput peak. However, this form of management has been criticised as potentially offering a back door to abandoning net neutrality by preferring specific services or traffic.

C. UK

In direct contrast to current developments in the US, the UK government has taken a regulatory approach to ensuring that all UK homes and businesses should have a minimum standard of access to high-speed Internet by 2020.[16] This in itself, however, does not currently prevent the UK's leading ISPs from filtering and blocking Internet content.

In 2014, the *Enemies of the Internet* annual report published by Reporters Without Borders (RSF) listed the UK among the top 14 states where data traffic is monitored, blocked or manipulated.[17] Yet in its 2017 report to the European Commission on compliance with net neutrality regulations, the UK communications regulator Ofcom claimed that 'there are no major concerns regarding the openness of the Internet in the UK.'[18] Those areas identified were minor concerns related primarily to choice of end-users' terminal equipment and zero rating. This apparent contradiction derives from limitations in the EU regulations. In addition to introducing 'sufficient data traffic

16 Paul Sandle, 'Britons will have legal right to high-speed broadband by 2020', Reuters, December 20, 2017, https://uk.reuters.com/article/uk-britain-broadband/britons-will-have-legal-right-to-high-speed-broadband-by-2020-idUKKBN1EE0RS.

17 Reporters Without Borders, annual Report 'Enemies of the Internet 2014', 12 March 12, 2014. See also James Vincent, *The Independent*, 'UK Branded an "Enemy of the Internet" for the first time by Reporters Without Borders', March 17, 2014, https://www.independent.co.uk/life-style/gadgets-and-tech/uk-branded-an-enemy-of-the-internet-for-the-first-time-by-reporters-without-borders-9196571.html, last access: January 7, 2018.

18 'Monitoring compliance with the EU Net Neutrality regulation: A report to the European Commission', Ofcom, June 23, 2017, p. 2, https://www.ofcom.org.uk/__data/assets/pdf_file/0018/103257/net-neutrality.pdf.

management' and 'specialised services', the EU also leaves decisions on whether actions are compliant with the regulations with national courts. As a result, while the Commission may have drafted a regulation on the telecommunications single market that seems to prohibit general filtering, blocking and monitoring of data packets due to net neutrality considerations, in practice implementation of these regulations depends on national jurisdiction. In other words, varying standards of net neutrality can be applied that are still compliant with the EU Regulation and with national law. While Ofcom followed the Commission's regulatory guidelines, RSF applied an ideal image of net neutrality not defined by the EU.

The fact that the landing points of several of the submarine cables that form the backbone of the Internet, especially between Europe and the US, are in the UK is particularly noteworthy. If European net neutrality standards are not carried across into UK law on the withdrawal of the UK from the EU, this will mean that the UK is free to apply its own standards to a substantial proportion of the data that passes between the United States and the EU. Unlike internal developments in the US, this could have a direct effect on the uninterrupted throughput of packets intended for delivery to Europe.

D. Russia

Russia has taken a significantly different approach to net neutrality and to privileging defensive measures compared to the UK, Europe or the US.[19] Most Russian ISPs provide clients with cost-free access to certain websites and services, such as Facebook, Vkontakte, Odnoklassniki, LiveJournal and Yandex Maps.[20] But in addition, governmental privilege is a significant factor in determining access. Many government websites are free to access by law,[21] and by contrast the government has the legal and technical power to disrupt or entirely block access to other Internet resources. According to Russian prosecutor-general Yuriy Chaika, by 2017 around 1,200 websites had been officially blocked under this legislation.[22]

In March 2017 legislation was reported to be under preparation under which Russian courts would be able to punish both domestic and foreign corporations for failing to comply with Russian law by ordering that access to their websites be slowed down.[23] The storage of Russian users' data on Russian servers by foreign Internet companies has been required by law since September 2015, when Law No. 242-FZ,

19 Roman Mirov, 'Конец нейтралитета: как США проиграли битву за интернет,' *Lenta.ru*, January 3, 2018, https://www.gazeta.ru/tech/2018/01/03/11551418/no_net_neutrality.shtml.

20 Sergey Vorniches, 'Всё, что нужно знать о сетевом нейтралитете,' *Apparat.cc*, February 27, 2015, https://apparat.cc/world/about-net-neutrality/.

21 'Доступ к 122 сайтам Рунета сделают бесплатным,' *Известия*, February 24, 2015, https://iz.ru/news/583390.

22 'Russian Police Have Blocked 1,200 Websites Since 2014,' *The Moscow Times*, January 12, 2017, https://themoscowtimes.com/news/1200-russian-websites-blocked-since-2014-56794.

23 Anastasia Golitsyna, 'Для интернет-компаний придумали наказание—замедлять доступ к их сайтам,' Ведомости, March 13, 2017, https://www.vedomosti.ru/technology/articles/2017/03/13/680827-zamedlit-skorost-dostupa.

adopted in 2014, came into force. Compliance with this localisation requirement by Twitter[24] and Snapchat[25] has been claimed by the Russian communications regulator Roskomnadzor but denied by the companies themselves, while Facebook is not yet compliant and consequently is regularly threatened with a nationwide ban.[26] According to a November 2017 survey, Google, Apple, Alibaba, Viber, Gett, Uber and Microsoft all rent Russian data centre space for the purpose of compliance.[27]

All of these measures are in accordance with a predominant view among Russian government agencies, especially those concerned with national security, that the Internet presents more of a threat than an opportunity. In April 2014, President Vladimir Putin remarked that the Internet 'came about as a special project of the CIA' and implied that it continued to be a tool of the US government, and consequently dangerous for Russia.[28] In contrast with Western assumptions, Russian information security preoccupations focus on the role not only of hostile code such as cyber attacks, but also hostile content such as opinions or information which are detrimental to the Russian state.[29] President Putin has personally praised Chinese-style censorship and defended it against criticism from digital rights advocates.[30]

But Russia's plans to protect itself from the Internet go even further, and extend to consideration of operating without access to global Internet services at all.[31] This scenario is variously presented by Russian government officials as either a voluntary withdrawal by Russia – 'pulling the plug' – or being disconnected by the hostile West, which according to one persistent Russian view, controls the Internet.[32] President Putin's adviser on Internet affairs, German Klimenko, is a particular advocate of Chinese-style Internet restrictions and preparing for possible total net withdrawal.[33]

24 Alec Luhn, 'Moscow Says Twitter Ready to Store Data of Users on Russian Servers Despite Concerns Over Surveillance,' *The Telegraph*, November 8, 2017, http://www.telegraph.co.uk/news/2017/11/08/moscow-says-twitter-ready-store-data-users-russian-servers-despite/.

25 Marina Galperina, 'Oops, Snapchat Accidentally Ended Up on a Russian Government Snitch Registry,' *Gizmodo*, August 10, 2017, https://gizmodo.com/oops-snapchat-accidentally-ended-up-on-a-russian-gover-1797721574.

26 'Роскомнадзор пригрозил Facebook блокировкой,' *РБК*, September 26, 2017, https://www.rbc.ru/own_business/26/09/2017/59ca1e899a7947351acdf385.

27 Galina Boyarkova, 'Все терабайты в гости к нам,' *Фонтанка*, November 12, 2017, https://www.fontanka.ru/2017/11/10/144/.

28 'Путин заявил, что интернет - это проект ЦРУ,' *BBC Russian Service*, April 24, 2014, http://www.bbc.com/russian/rolling_news/2014/04/140424_rn_putin_csi_Internet.

29 This contrast is examined in detail in Keir Giles, 'Russia's Public Stance on Cyberspace Issues', in *C. Czosseck, R. Ottis, K. Ziolkowski (Eds.), 2012 4th International Conference on Cyber Conflict.*

30 'Не стоит критиковать китайский вариант ограничений в Интернете – Путин,' *Звезда*, April 3, 2017, https://tvzvezda.ru/news/vstrane_i_mire/content/201704031346-9yin.htm.

31 Grigory Naberezhnov and Darya Luganskaya, 'Кремль прокомментировал сообщения об отключении России от интернета,' *РБК*, September 19, 2014, https://www.rbc.ru/politics/19/09/2014/5704225a9a794760d3d419b6.

32 'Клименко: Россия должна быть готова к отключению от мирового интернета,' *ТАСС*, December 29, 2016, http://tass.ru/obschestvo/3914882.

33 'Советник президента предложил ограничить интернет в России,' *Дождь*, January 26, 2017, https://tvrain.ru/news/Internet-426274/.

In March 2018, Klimenko announced that, after lengthy preparations, Russia was now technically capable of removing itself from the global Internet.[34]

Russia's security-driven approach to managing the Internet stands in stark contrast to the Euro-Atlantic community, and the difference is instructive. We argue in this paper that net neutrality as currently understood by the West is a potential handicap for ensuring security and responding to cyber warfare actions. In Russia, this challenge is well recognised and bound up with the perceived threat of free flow of information across national borders, which for the West is an inalienable element of how the Internet works. The result is that Russia has circumvented the net neutrality challenge by changing the entire basis for Internet access, and making it conditional on state interest. Any solution this extreme would be unpalatable and unworkable in Western liberal democracies, being incompatible both with principles of freedom of expression and with the greater independence of commercial entities including ISPs outside Russia.

3. NET NEUTRALITY AND CYBER WARFARE

Recent net neutrality discussions have centred on censorship, Internet access and traffic limitations. However, these discussions are too narrow and must be expanded to more general considerations on data traffic management, which should be perceived as a core element in future cyber warfare.

A. Net Neutrality in Attack Vectors

Malicious actors can abuse net neutrality to establish dominance through different attack vectors, including DDoS, DrDoS and SYN-flood attacks.

DDoS-attacks use the fact that all incoming traffic is treated equally to create an advantage for the attacker. All IT components have a limit to their processing capabilities, and when legitimate requests to a component compete on an equal basis with a flood of malicious traffic from bots, the component is overloaded and becomes unable to reply. While this principle is a standard tactic, there are many different ways of carrying out a DDoS-attack. In a distributed reflected DoS-attack (DrDoS), the attacker hijacks (spoofs) the IP-address of its target and sends service requests to servers (such as the DNS), asking them to reply to the spoofed IP. What follows is a DDoS-attack with no attribution being possible and, depending on the servers involved, that is impossible to block without self-inflicted damage. One of the largest DDoS-attacks recorded to date was observed during March 2018 against Github, causing a record-breaking data transfer rate of 1.35 Terabits per second using

34 'Советник Путина: Россия готова к отключению от мирового интернета', *RFE/RL*, March 5, 2018, https://www.svoboda.org/a/29079358.html.

a modified DrDoS.[35] In this scenario the attacker also relies on the fact that the target will treat all data packets equally, even when not useful, not requested or identified as potentially harmful.

One of the most basic, yet highly imbalanced methods to attack a network component is a SYN flood attack. SYN flood attacks belong to the group of DoS-attacks that abuse both the equal treatment of packets at the target's side and the TCP handshake protocol. To establish a TCP connection to the target (server-side) from the attacker (client-side), a three-way handshake is initiated. The client sends a SYN-request (synchronise) to the server, the server replies with a SYN-ACK (SYN-acknowledge) and allocates resources for the awaited TCP connection. Usually, the client replies with another ACK, which establishes the TCP connection, however, a malicious client can withhold the final ACK. This leads to the server keeping the resources allocated blocked until a timeout is reached. Depending on the servers' configuration, the allocated resources may make up a considerable proportion of the resources available and the timeout may be excessively long. If this attack is combined with a distributed approach, or if many SYN requests are started in parallel, the result is a DoS.

B. Imbalance of actors

Techniques for malicious actors to circumvent the legitimate control and regulation of data are publicly available and used. Legitimate actors, by contrast, cannot demand more bandwidth or privileged access from ISPs to create a power balance between themselves and sophisticated attackers. In fact, even direct responses to an ongoing attack may be problematic as in many cases attribution has to be examined and verified by juridical institutions to make any actions against the source legitimate. Legitimate actions therefore often focus on re-routing mechanisms or involve large redundancy set-ups to cope with outages. However, these fail-safe environments are necessarily limited and bound to the number of fall-back components integrated.

Currently, net neutrality places still further constraints on the technical capabilities of legitimate cyber actors. When considered strictly, net neutrality principles prevent live monitoring of suspicious traffic and hinder any attempts of attribution through the ISP, even though the ISP is often the first to notice unusual cyber activities. Traffic blocking is also against net neutrality standards, even if it is obvious to the technical expert that the traffic is involved in an ongoing attack. To resolve this issue, ISPs have begun to attempt to contact the initiators of such traffic; a tedious, costly and potentially fruitless venture.[36]

35 Lily Hay Newman, 'Github survived the biggest DDoS attack ever recorded', *Wired Security*, March 3, 2018, https://www.wired.com/story/github-ddos-memcached/, last access: March 16, 2018.

36 Michael Kan, 'Amid cyberattacks, ISPs try to clean up the Internet', *CSO Online*, February 23, 2017, https://www.csoonline.com/article/3173274/security/amid-cyberattacks-isps-try-to-clean-up-the-Internet.html, last access: January 7, 2018.

Discussing net neutrality in terms of traffic management and control inevitably leads to the insight that net neutrality protects both ordinary users and actors with hostile intent. While the rights and protection of innocent users should not be reduced unnecessarily, methods should be developed to empower legitimate over malicious actors.

C. Cyber Actions

The effects net neutrality has on cyber warfare scenarios can be divided into three distinct categories, based on the type of cyber action: cyber defence, proactive cyber defence and offensive cyber operations.

While cyber defence generally describes actions taken in the aftermath of cyber attacks and passive methods to deter or prevent the attack, proactive cyber defence allows an active response during and, to a degree, prior to cyber attacks taking place. Offensive cyber operations may range from aggressive, conflict-initiating operations, to supportive actions among allies during defensive cyber scenarios, but are generally directed against the attacker or its associated components.

Long-term defensive measures include log analysis, system hardening, redesigning of networks, training of personnel and developing incident response strategies. Immediate defensive techniques are especially those that are used to prevent further damage and neutralise the ongoing attack by measures taken at the victim's end only. Typical examples are the shutdown of servers, network components or infected devices and the blocking of traffic and services associated with the attack. These methods generally do not conflict with net neutrality principles if coordinated through legitimate law enforcement units or if immediate action is needed to prevent further damage to the ISP. However, immediate action through cyber units or proactive approaches through ISPs to prevent damage in foreign networks are currently limited.

One possible resolution of this conflict of interest would be that legitimate actors should be limited to defensive techniques to minimise contravention of net neutrality principles. However, purely defensive techniques are often of limited utility if the attacker's motivation is to cause the unavailability of services or devices. This is commonly seen in the various forms of denial-of-service attacks (DDoS). Furthermore, defensive strategies may also be considered too insecure if more sophisticated attacks are expected that may remain unnoticed for longer periods of time. These types of attacks are typically associated with espionage or information warfare, and it is these cyber activities in particular that are protected by current net neutrality standards. Although ISPs may be able to deduce that traffic is suspicious based on heuristics (i.e. without violating net neutrality), net neutrality would prevent further investigation

and action against the initiator unless authorised by law enforcement and judicial authorities.

Proactive cyber defence allows defensive methods to be combined with more aggressive monitoring and filtering rules. The line between defence and proactive defence is often blurred and depends on the specific technologies used. Firewall rules may be proactive and not compatible with net neutrality standards and DPI, which allows analysis on the content of the data packet passing and is often used to enforce Internet censorship. DPI is not compatible with net neutrality principles when applied to certain packets only.

Offensive cyber actions may vary greatly depending on the assets and technologies used. Any type of offensive strategy that aims at limiting, blocking, monitoring or manipulating specific traffic has to be considered as violating net neutrality principles. Whether legitimation can be given and under which circumstances has to be considered by the judiciary. It appears questionable whether it can be demanded of ISPs that they participate in military or governmental operations violating agreed telecommunication standards, such as net neutrality. But if they do not, this would imply a need for legitimate cyber actors to reroute traffic to their own network components to bypass ISPs in the context of offensive cyber activities to avoid limitations introduced by those ISPs during the operation.

If applied strictly to all traffic, demanding and enforcing the equal treatment of all data packets would prohibit the use of several cyber defence techniques. Those considered proactive would be particularly affected, since they rely on traffic being monitored based on origin, destination or content. If carried out by ISPs, these measures are not in line with net neutrality principles. Offensive cyber actions too may need the permission or active involvement of ISPs, which raises questions of legitimacy, particularly if this includes violations of agreed telecommunication standards.

D. Cyber Power

Actors in cyberspace are represented by their data and traffic. Controlling either data or traffic corresponds to controlling the actor. Limiting the capabilities of legitimate actors to legally interfere with malicious traffic is a digital form of unilateral disarmament, and as a consequence has the capability to destabilise cyber sovereignty.

As described above, net neutrality places limits on the whole range of legitimate actions in cyberspace, reducing both offensive and preventive measures. However, these limitations again only apply to actors bound by restrictions, while illegitimate actors can choose to circumvent or disregard them. The limitation of preventive measures plays a major role not only in constraining defence against future attacks,

but also in helping attackers conceal their activities and avoid prosecution. This is because net neutrality prevents ISPs from collecting only selected data from the traffic they forward. Paradoxically, this has often been a contributory factor to the adoption of general telecommunications data retention (e.g. in Germany). The irony is that from the point of view of net neutrality, if you collect data on everybody this is legal and acceptable, but only collecting data on traffic that appears suspicious is not.

Overall, strict application of net neutrality principles contributes to an unbalanced cyberspace. Legitimate actors are being deprived of rights granted in non-digital circumstances, while the community is unable technically to enforce net neutrality on the attackers' side as well. This gives rise to a substantial mismatch in the distribution of cyber power among actors.

4. OPPORTUNITIES AND CHALLENGES

If net neutrality principles are weakened, ISPs will need to reserve bandwidth and develop reliable methods to identify privileged customers and services without introducing additional physical media in order to guarantee high transmission rates for these customers and services; the mere throttling of 'unprivileged traffic' is insufficient. It is likely that both channelling and protocol developments will take place. Additional hardening of access to these channels may help to ensure that only legitimate users have access to the channel. Creating privileged channels contributes to restoring a balance between legitimate cyber actors and attackers in cyberspace. Currently, attackers have the ability to simply allocate bandwidth and to technically enforce prioritised processing, while the options of legitimate actors are severely limited.

Cyber defence support among allies could be affected positively by weakening net neutrality principles and installing prioritised channels. Establishing privileged high-speed connections may prove valuable in scenarios where remote access to networks under attack is needed. This occurs when network administration personnel are faced with sophisticated cyber attacks for which they are insufficiently prepared. In such cases, remote access could be established, even in scenarios including a denial of service, by technically enforcing processing of data received by the prioritised channels through networking rules and interrupt handling strategies. Such methods could be implemented easily in Software Defined Networks (SDNs), however, standards should be defined that ensure these measures conform to our democratic norms. This would in turn not only allow remote support during cyber incidents but facilitate forensic activities during and after the attack.

Prioritised channels could also be used to uphold a minimal service availability if, for example, critical infrastructure is being targeted. The use of prioritised channels allows the separation of critical traffic from common or public traffic. While a smaller number of sophisticated attacks should be expected to target the prioritised channels, the larger portion of less sophisticated and limited attacks will target the public traffic channels, which in turn may be processed on less prioritised components with limited device access. Although this may appear unfair at first, current security standards attempt to enforce precisely this by network virtualisation and service encapsulation. However, due to their high abstraction layer, several vulnerabilities arise within solutions based on virtualisation and the attack surface is even enlarged.[37] These vulnerabilities are not to be expected on lower abstraction layers, which is why we would envision low layer solutions.

Although several benefits could be expected from weakening net neutrality principles and establishing prioritised traffic through ISPs, new attack vectors must also be expected. As bandwidth and transmission rates are high-value assets in cyberspace, attackers are likely to work on ways to obtain access to prioritised traffic. Therefore, the development of such technologies and the definition of adequate standards should not be left to the free market only. It must also be guaranteed that democratic values and standards are not being undermined. However, this is an obligation of Western democracies that should not only apply for legitimate actors, but must also be enforced for malicious actors threatening the cyber domain.

5. OUTLOOK

This article has explored net neutrality and networking principles from both strategic and technical views. The handling of net neutrality and traffic equality within the EU, UK and Russia were compared and discussed. Particular attention was given to the influence the different approaches have in the uprising congested and contested cyber domains as expected in cyber warfare scenarios.

Russia's distinct approach to net neutrality and network regulations in general was explored, highlighting the measurements taken and scheduled to prevent the destabilising effect net neutrality has on cyber power and sovereignty. While several of the technologies and regulations established within Russia are not acceptable by Western standards due to their limitation of individual rights, the deployed methods show Russia's sensibility to the arising threats and an awareness of the cyber power imbalance.

37 Candid Wueest, 'Threats to virtual environments', *Symantec Security Response*, August 12, 2014; European Union Agency for Network and Information Security (ENISA), *Security aspects of virtualization*, February 2017, ISBN 978-92-9204-211-0.

The EU is currently struggling with enforcement of the approved Regulation on the telecommunications single market. The Regulation allows national judicial interpretation which leads to different implementations of net neutrality within the EU. This condition is unsatisfactory as it creates an imbalance between EU members both in terms of market regulations and cyber power. This limits joint cyber operations, as cyberspace is not limited by national borders, but data traffic is treated according to national jurisdiction, possibly hindering prosecution depending on the national networking regulations.

The UK has made a step forward in terms of providing broadband access to all consumers, however, it has also been considered as one of the 'enemies of the Internet' by the RSF since 2014. The UK is known for its surveillance capabilities, which can also be applied through local ISPs. It is noteworthy that the UK plays a major role in building the transatlantic backbone of the Internet, especially between the United States and EU. Severe limitations of net neutrality must be expected to follow the withdrawal of the UK from the EU unless regulatory and technical enforcement are developed.

Discussions on net neutrality are discussions on traffic management. There is a requirement to define standards and policies that regulate when and how legitimate actors may demand assistance by ISPs to either prioritise their own traffic or limit the traffic of potentially malicious actors. As blocking or reducing malicious traffic may result in unjust penalisation of unaware end-users, this paper advocates the prioritisation of governmental (or governmentally legitimated) cyber actors. The aim of any legitimate action in cyberspace must be to protect civilian users while defending networks and services and to establish cyber sovereignty and power.

While there are good reasons to weaken net neutrality principles, this should be done in a controlled manner and monitored by independent authorities. As demonstrated in the case of the United States both before and immediately following the 2017 easing of net neutrality constraints, uncontrolled outsourcing to private companies bears the risk of abusive methods that not only influence the end users of telecommunication services but may also limit free market growth and lead to monopolies.

Net neutrality regulations should consider the protection of individual rights and equality among civilian end-users but must also ensure stability in cyberspace and equality among actors. This is of particular importance in cyber war scenarios where some states are less constrained in their legitimate cyber activities than others. There are two possible choices: either to technically enforce net neutrality (which has already been proven impractical in the face of botnets or distributed cyber attacks as the attribution of cyber actions remains an unsolved task) or to define regulations that

allow legitimate actors to rebalance cyber power and regain control over congested networks during cyber incidents to uphold sovereignty in cyberspace.

AKNOWLEDGEMENT

The authors are grateful for research assistance from Lincoln Pigman in the preparation of this paper.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

The Cyber Decade: Cyber Defence at a X-ing Point

Robert Koch
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
robert.koch@unibw.de

Mario Golling
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
mario.golling@unibw.de

Abstract: As a consequence of the numerous cyber attacks over the last decade, both the consideration and use of cyberspace has fundamentally changed, and will continue to evolve. Military forces all over the world have come to value the new role of cyberspace in warfare, building up cyber commands, and establishing new capabilities. Integral to such capabilities is that military forces fundamentally depend on the rapid exchange of information in order for their decision-making processes to gain superiority on the battlefield; this compounds the need to develop network-enabled capabilities to realize network-centric warfare. This triangle of cyber offense, cyber defence, and cyber dependence creates a challenging and complex system of interdependencies. Alongside, while numerous technologies have not improved cyber security significantly, this may change with upcoming new concepts and systems, like decentralized ledger technologies (Blockchains) or quantum-secured communication.

Following these thoughts, the paper analyses the development of both cyber threats and defence capabilities during the past 10 years, evaluates the current situation and gives recommendations for improvements. To this end, the paper is structured as follows: first, general conditions for military forces with respect to "cyber" are described, including an analysis of the most likely courses of action of the West and their seemingly traditional adversary in the East, Russia. The overview includes a discussion of the usefulness of the measures and an overview of upcoming technologies critical for cyber security. Finally, requirements and recommendations for the further development of cyber defence are briefly covered.

Keywords: *cyber war review, cyber defence implications, cyber defence recommendations, cyber defence requirements, future technologies, cyber war*

1. INTRODUCTION

As a consequence of the cyber attacks on Estonia in 2007, both the consideration and use of cyberspace by the military has fundamentally changed and will continue to do so. Over the years, such attacks have effectively demonstrated how significant impacts can be wrought by supposedly trivial and low-key means. On the other hand, military forces also depend strongly on the rapid exchange of information for their decision-making process so their forces can gain battlefield superiority, which enforces the need for network-enabled capabilities (NEC) [1] to realize network-centric warfare (NCW) [2]. This creates a challenging and complex system of interdependencies, opening a broad spectrum of possible attack vectors. Therefore, operations in cyberspace can be used to generate effects not only in cyberspace itself, but also in the physical environment, which is an attractive new capability for military commanders. Indeed, armed forces worldwide now highly value the new role of cyber, and are building cyber commands and establishing new operational capabilities, and the asymmetric nature of cyber warfare can give the advantage to armed forces otherwise in possession of comparatively smaller weaponry. However, the complexity of sophisticated cyber attacks like Stuxnet can also imply the opposite. Thus, cyber defence is also of enormous importance. And while numerous technologies proposed over recent years have not improved cyber security significantly, this may change with upcoming new concepts and systems. Blockchains, quantum-secured communication, mathematically verified software microkernels, and trusted hardware platforms are likely to be key elements for new, more secure systems. Along with the armaments industry itself developing a better understanding of cyber threats, this should lead to better and more resilient weapon systems.

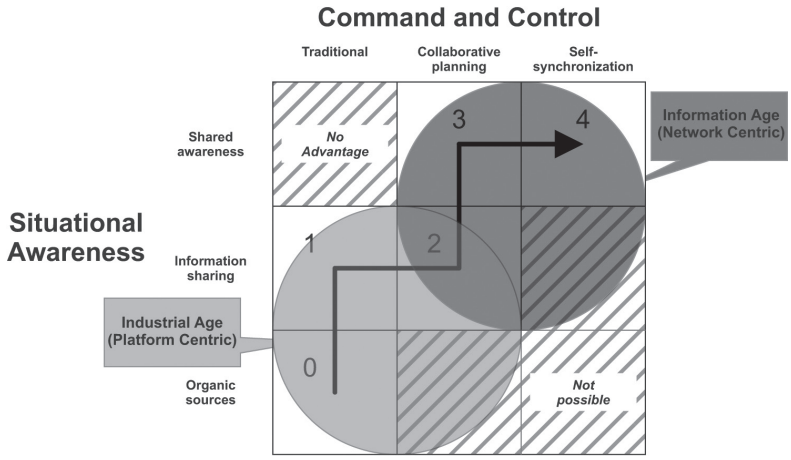
In light of these thoughts, the paper analyses the development of both cyber threats and defence capabilities over the past 10 years from 2007 to 2017, evaluates the current situation and gives recommendations for further development. The paper is structured as follows: first, general conditions for military forces with respect to “cyber” are described and dependencies and requirements are highlighted. Second, a brief overview of the development of cyber threats and defence capabilities during the past ten years is given, including a discussion of the usefulness of the measures. Upcoming technologies which are important for cyber security are briefly discussed to analyse opportunities for more secure systems. Finally, the conclusions of the paper are summarized and requirements for the further development of cyber defence are derived.

2. DETERMINING FACTORS

For all the millennia of warfare that have passed, the tools and tactics of how armies fight have evolved as military technologies have evolved [2]. However, recent years have seen fundamental changes come to affect the very character of war [2]. Military forces worldwide are increasingly capitalizing on the advances and advantages of information technology to facilitate radical changes in the way they structure and deliver offensive and defensive capabilities [1]. The US Navy was among the first to investigate how to use Information and Communication Technology (ICT) to increase the efficiency and efficacy of their forces on the 21st Century battlefield [3], the main consequence being the increased integration of individual, hitherto autonomously acting systems, thus a fundamental shift from what is called platform-centric warfare to network-centric warfare (NCW). NCW harnesses network technology to facilitate radical improvements in the shared awareness of disposition and intent, together with a capability for rapid reconfiguration, and synchronization of operations [1] and thus improves both the efficiency and effectiveness of military operations [4].

As such, NCW creates superiority in war by harvesting information from a network of reconnaissance systems and enabling its analysis and use by command and control centres, as well as use in weapons systems. Hence military superiority across the entire range of military operations, i.e. full spectrum dominance, is achieved. The vision for NCW is to provide seamless access to timely information at every echelon in the military hierarchy and enable all elements to share information within a single, coherent, complete, and dynamically accurate picture of the battlefield. It is intended that NCW will produce an improved understanding both of the intent of higher command and of the operational situation at all levels of command, with every element better able to tap into the collective knowledge and reduce the “fog and friction” [4] of war, and enable the optimal use of resources. Although the transformation towards NCW is not finalized completely, even not by the United States [1, 4], NCW is anticipated to be one of the greatest revolutions in military operations in the past 200 years (see Figure 1).

FIGURE 1. NCW ROUTE MAP [5, 6]



However, within such an integrated system lies a greater vulnerability: attacking the weakest link could compromise the entire system and lead to catastrophic consequences, in the worst case rendering an entire military force incapable of action.

A. On Multinational Coalitions

In addition to the increased use of information technology, the aspect of cooperative, multinational participation in conflicts is of great importance. Military operations today are almost always multilateral in complexion. With regard to NATO, since 1990, there has been a significant increase in the number of military operations requiring NATO member states to contribute forces to some multinational coalition or alliance [7]. Moreover, the range of mission types has broadened to include peacekeeping, peace support, and humanitarian operations [7]. Corresponding challenges with such a force are, for example, what the agreed operational concepts are, different intelligence requirements and structures, the diverse capabilities and qualities of the various formations as well as command, control, communication, and intelligence (C4I)/cyber interfaces that have to be developed and integrated [8, 7].

Increased defence cooperation, such as "Smart Defence" (NATO), "Pooling & Sharing" (European Union), or the "Framework Nations Concept", in theory increases sustainability and helps to preserve key military capabilities [9]. Smaller armies can plug their remaining capabilities into an organizational backbone provided by a larger, "framework" nation [9]. In practice, however, this theory has yet to fully prove itself, and the extent to which those well-understood obstacles to defence cooperation can be overcome remains to be seen [9, 10]. Deeper cooperation also calls for reliability among the different partners [9]. In terms of ICT, cyber is always a potential risk.

Therefore, it can be said that despite the undisputed advantages of multi-national coalitions, a military force made up of numerous divergent parts can see the overall system's cyber defences be compromised, which can easily play into the hands of the attacker and hamper one's own defence.

B. On Russia – Analysis of Russia's Course of Action

Western countries follow the assumption that economically prosperous democracies are less likely to wage war against each other. Therefore, the EU operates a "Europeanization policy" aimed at democratization and economic liberalization, particularly in its eastern domain [11]. This basic principle of foreign policy, however, is by no means *a priori* transferable to all states. In light of Russia's annexation of the Crimea in Ukraine and the war in the Donbas, unresolved territorial conflicts on the eastern borders of the EU have gained international attention and concern. From Russia's perspective, the West's approach is flawed on several fronts. In context, shortly after World War Two, the Kremlin sought to protect the USSR by establishing a *cordon sanitaire* between itself and its major nemeses, the Western powers [13], with the occupation and coercive support of eastern European states under the Warsaw Pact. Although this buffer zone disintegrated over 1989-1991, as did the USSR itself, with the Kremlin believing its borderlands could slip under the aegis and control of the West, Moscow created the concept of "Frozen Conflicts" to weaken, divide, and ultimately prevent these countries (Moldova, Georgia, and Ukraine, for example) from drifting far from their eastern orbit of Russia [13]. Russia did so through the manipulation of nationalist impulses among border populations [13], encouraging minorities to think of themselves as distinct from the majority population [13]. Unwilling to risk more than limited open military intervention, the Russians enhanced hybrid warfare (which has its origins in 1938), using the presence of its peacekeepers and its diplomatic powers to keep these conflicts in a "no war, no peace" situation (i.e., Frozen Conflicts) that perpetuates a Russian role in its borderlands [13].

As much as Russia profits more from enabling if not inciting temporary and regionally-limited "skirmishes" to justify its own intervention, it is important to prevent the West from being drawn into these conflicts. Hence the importance of the concept of "Escalation Dominance" within every domain, including cyber, which imparts the ability to create a credible deterrence to outside forces involving themselves. Like any offensive or defensive capability, this will only work as a deterrent if the host nation shows it has the appropriate means, and the will to use them. As a conclusion, it can be stated that Russia – unlike the West – benefits more from regionally limited Frozen Conflicts and, for reasons of Escalation Dominance, also in terms of cyber, might feel the need to demonstrate Cyber Dominance to hamper other nations engaged or seeking to engage in those conflicts. Correspondingly, this increases the likelihood of cyber attacks.

3. A DECADE OF CYBER THREATS AND DEFENCE

Making hard decisions in the area of cyber security requires a comprehensive understanding of cyber security threats and developments. What follows then is an analysis of the last 10 years in the evolution of cyber threats and defence from 2007 to 2017, the starting point being the Distributed Denial-of-Service (DDoS) attacks on Estonia, which marked a step-change in the onset of cyber warfare.

A. Development of Security Incidents

CyCon X signifies 10 years of conferences dealing with legal, strategic, conceptual, and technical challenges of cyber conflict. Motivated by the consequences of the DDoS attacks on Estonia [14], which affected broad parts of everyday life in a country that had already highly digitized systems of infrastructure and governance, the conferences sought to explore and discuss numerous aspects of cyber security.

2007-2009: The DDoS attacks on Estonia were not the only remarkable event in 2007. DDoS attacks are themselves a relatively simple method of attack, where vast amounts of data requests are directed towards a target with the aim of exhausting the target's means of providing data, and legitimate traffic is blocked out in a simple but effective method. But this is just one case and while any number of digital assaults may be ostensibly quite primitive in format, it is the failure to anticipate them that enables their effectiveness, and significant, material impacts can be delivered. For example, the US Department of Energy ran the so-called Aurora experiment in their Idaho Labs in 2007 [15], showing how an attack on a power generator's control system could led to the generator's destruction. These incidences, both actual and theoretical, brought the issue of vulnerabilities in modern critical infrastructures into the public domain for the first time.

Meanwhile, a remarkable military operation was undertaken by the Israeli Air Force (IAF). During Operation Orchard, the IAF executed a pre-emptive strike against Syria's plutonium-powered nuclear reactor Al Kibar shortly before it became active [16]. Highly successful, no plane was lost, with not a single Syrian missile fired. Some reports said this was because Syria's air-defence systems were blinded by standard electronic scrambling tools [16], but some analyses highlighted the use of either special software or a backdoor in the adversary's systems as more likely explanations for their failure to fire [17].

The notorious worm Downadup (also known as Conficker) appeared in October 2008. While worm attacks had already been declining for some years, Downadup manifest itself as one of the most widespread threats seen in some time [18]. It combined several techniques to spread itself and hide within systems, and defend itself against

attacks. Even by 2014, over a million machines were still infected, highlighting the difficulties of removing this malware [19]. The ability to observe the defenders and adapt the code underlined the sophistication of the hackers [20]. While attribution of the precise origins of the attack is still not clear, with the creators of Downadup remaining unknown, sources were traced to Ukraine and China. Only the last version of the worm carried a malicious payload and it was a version that deleted itself after a month. This may indicate that Downadup was more of a test run by a still unknown source rather than a directed attack by cyber criminals.

Also in 2008, manipulated credit-card readers were found in UK supermarkets. The devices were fitted with wireless equipment and could transmit stolen data once a day or go dormant to avoid detection [21]. This was a remarkable attack on the country's retail supply chain and its customers.

The rising threat towards critical infrastructures was seen when the US Federal Aviation Administration's computer systems were hacked in 2009 [22], endangering not only commercial air traffic but military operations as well. In February, the (in-) famous Downadup malware together with poor cyber hygiene grounded French naval aircraft [23], and in December, the US military realized that Iraqi insurgents had used the $26 software "SkyGrabber" to capture video feeds from US drones that had been transmitted via satellite links [24]. Despite what newspapers reported at the time, there was no "hacking" involved, only installing the software, aligning the antenna, and starting the record: the transmissions themselves were unencrypted.

2010: Some serious incidents affected the Internet in early 2010. Apparently, a configuration error made the I-root instance of the Domain Name System (DNS) root servers visible outside of China. I-root does not give correct address resolutions for all queries because of online censorship in China. Suddenly it was being used by computers outside of China, which unintentionally fell into this censorship [25]. Only two weeks later, a small Chinese ISP called IDC China Telecommunications Corporation, that had normally sourced about 40 prefixes, announced nearly 37,000 unique prefixes for about 15 minutes. Because of that, approximately 10 per cent of Internet traffic was rerouted through China, including traffic from providers like Deutsche Telekom and AT&T [26]. The incident highlighted how a good understanding of structures and protocols can be used to generate simple and effective attacks.

An important incident was discovered in October 2010 by the Belarusian company VirusBlokAda: Stuxnet. While the complexity of the malware sample challenged the security companies (resulting in some incorrect analysis), eventually it was determined that the malware attacked the Iranian enrichment facility in Natanz, interfering with the enrichment process and finally destroying centrifuges [27]. While it was not, as

reported at the time, the first cyber attack to result in physical damage, it definitely was a game changer, clearly demonstrating the new opportunities thrown up by a globalized, interconnected world. Even more, it marked the start of a new area of cyber ambitions from numerous countries around the world.

2011-2012: A sophisticated spear-phishing attack in 2011 obtained data used to compromise network security company RSA's SecureID technology, which was then used to attack Lockheed Martin [28].

In 2012, the media reported on Chinese hackers stealing classified information about Lockheed Martin's F-35 Joint Strike Fighter (JSF), as revealed by documents obtained by the NSA whistle-blower Edward Snowden (whose own story is a testament to how vast, top-security IT systems can be compromised by one person with a USB-stick, see below) [29]. Comparing Lockheed Martin's JSF and China's Shenyang J-31 fighter, David Majumdar has said: "On the surface, the J-31 looks very much like a twin-engine F-35 clone – and there are plenty of reasons to believe that the Chinese jet was based on stolen JSF technology – and could eventually be more or less a match for the American jet" [30].

Another controversial discussion was about a hardware backdoor in the Microsemi ProASIC3 processor – a chip used in numerous high performance aircraft, ranging from USAF fighters to the Boeing 777 Dreamliner, as well as military applications like encryption devices. While the researchers found some processor commands on-board the chip which could be used as a backdoor [31], industry argued that these functions were only undocumented debugging functionality to be used by the chip developers for testing purposes. On the one hand this may be true, especially as modern processors contain thousands of undocumented commands and features [32], but on the other hand, for a sensitive or classified military application, it was a dangerous attack vector.

2013-2014: The power of relatively simple hacks when executed by an agent with a strong understanding of a system and its dependencies was once again demonstrated in April 2013, when a fake Tweet sent by the Syrian Electronic Army via the Associate Press's Twitter feed caused a temporary crash of the New York Stock Exchange, costing US $136 billion. The content of the tweet said "Breaking: Two Explosions in the White House and Barack Obama is injured" [33]. Of course, it was quickly realized that there had not been an attack and the index recovered quickly; nevertheless, knowing (or executing) such a ploy can result in a lot of money being lost, or at least, changing hands when otherwise it might not.

Another major event that should profoundly change the importance with which cyber security is viewed was the Snowden Leaks. The whistle-blower Edward Snowden worked as a system administrator for the NSA until May 2013. He passed on top secret, classified information about surveillance projects to the world's press. The range of the revelations was vast, from the eavesdropping of Internet links, the introduction of hardware as well as algorithmic backdoors, to techniques for bridging the air-gap [34].

The Snowden Leaks came as part of a growing tide of stories about incidents of high level breaches of data and hacking. Even so, another breach in 2014 is of particular note, with the US Office of Personnel Management (OPM) targeted [35]. The severity of the breach stemmed from the business engaged in by the companies concerned, namely KeyPoint Government Solutions, the contractor for OPM, doing security clearance background investigations. Thus, the nature of the data was highly critical, not only because of personally identifiable information like Social Security numbers and addresses, but because of the risk of interference with and blackmailing potential of actual employees, with information heisted from such background checks.

In October 2012, NATO identified a comprehensive espionage campaign [36] that was attributed to Russia, and was found to have been going for five years already, additionally targeting institutions of the EU and the Ukrainian government. As is often the case, it was very difficult to give a close estimate as to quantity of data stolen. For example, logging data is often available only for short periods of time and is limited by legal regulations, which confounds the chances of getting a complete picture of what has happened. Hence, identifying the extent of the damage, and by that the scale and detail of potential hazards thereafter faced, is highly challenging.

There were also breaches identified and intensified in the energy sectors in the United States and across Europe. Hackers from the "Dragonfly" group, also known as "Energetic Bear", and traced to Eastern Europe, successfully hacked IT systems run by energy grid providers, electricity generation firms, petroleum pipeline operators, and industrial equipment providers in the US, Spain, France, Italy, Germany, Turkey, and Poland [37]. While the primary objectives were espionage and persistent access, there also remained the capability to carry out acts of sabotage [37].

2015: A highly "visible" attack occurred in April 2015, when 12 channels of the broadcasting station TV5 Monde went off air. While a defacement displayed IS propaganda online, an investigation identified the Russian hacker group APT28 as the source of the attack. It was a well-prepared assault and possibly sought to destroy the television station, but greater damage was prevented by the serendipitous presence on site of many more technicians than usual due to a new channel going on air the

same day of the attack [38]. Another attack resulting in actual physical consequences was demonstrated in western Ukraine in December. A long-prepared cyber attack on multiple electricity distribution stations caused power outages that affected approximately 225,000 customers. In parallel, phone DoS attacks were carried out on call centres to prevent customers from contacting the power company under assault [39].

2016: Early 2016 saw the beginning of the end for old-school methods of bank robbing, i.e. masked men with guns telling everyone to get on the floor, as new high-tech methods introduced themselves to the world stage [40]. An attacker group named "Lazarus", traced to North Korea, stole a total of over US $100 million, mainly from the Bangladesh Bank, among others, by penetrating the Alliance Access software used by the Society for Worldwide Interbank Financial Telecommunication (SWIFT) networks, which carries worldwide financial transactions in a (up to that point) secure and standardized way. In the same year, North Korean hackers also looted 235 GB of sensitive documents from South Korea's defence data centre, including blueprints for a joint-US plans for war on the peninsula and scenarios for removing the North Korean leader, Kim Jong-un [41].

During the breach of the Philippines Commission on Elections, personal information from all of the country's 55 million registered voters, including fingerprint data, passport numbers, and expiry dates, was exposed online and fully searchable [42], while the designs of India's Scorpion submarines was leaked from the French shipbuilder DCNS [43]. Other data breaches that came to the public's attention involved the casual dating website AdultFriendFinder, with the details of 412 million users exposed to the world [44], and even the NSA's own hacking tools were stolen by the hacker group "The Shadow Brokers" [45].

But while such incidents have grown in number and severity, the methods deployed in the attacks techniques are still often quite simple, with DDoS attacks achieving disruption of services ranging from Amazon and Netflix to the PlayStation Network – nearly one decade after the attacks on Estonia.

2017: Numerous cyber security incidents were seen in 2017. In May, the WannaCry ransomware campaign hit enterprises and institutions all over the world [47], with impacts including the taking offline of 61 National Health Service hospitals in the UK and leading to production at numerous Renault factories in France stopping. By using the ETERNALBLUE vulnerability stolen from the NSA in 2016 and published by the Shadow Brokers in April 2017, the malware was very virulent. While patches had been made available by Microsoft for supported systems in March 2017, the run affected especially older Windows XP/8/Server 2003 systems for which no patch

had been published. In light of the outbreak, Microsoft took the "extraordinary" and "unusual" step of providing an emergency update for the aforementioned systems [48]. While the attacks elicited little by way of ransom, the financial impact can be enormous. An attack by the NotPetya ransomware later in 2017 on Maersk cost the Danish shipping giant up to US $300 million [49].

Some details about stolen data from the NSA were also published. A contractor for the organization had without authorization copied data and stored it on his computer at home. Russian hackers then compromised that computer and raided the files. According to *The Wall Street Journal*, the files had been identified by the Russian attackers through the contractor's use of a popular antivirus software made by the Russia-based company Kaspersky Lab [50]. In addition that year, a new series of classified documents was leaked, this time from the CIA [51]. The material called Vault 7 and 8 showed the activities of the CIA in detail, including compromising cars, Smart TVs, and smartphones, and the CIA's capability to conduct cyber warfare [52].

Already by this selection of cyber security incidents of the past 10 years, it seems that the situation has not been improving. To better understand the underlying problems, as well as new opportunities for cyber defence, a quick look at security-related developments during the decade in question follows.

Technological Development

Looking back to 2007, the Canadian company D-Wave Systems, Inc. presented their first commercial 16-qubit quantum annealing processor. Annealing is not universal quantum computing (the most powerful form of quantum computing), and is really only able to solve optimization problems. But D-Wave was the first company using quantum effects for building new kinds of processors. A publication in 2015 [53] led to heated debates over the statement that a calculation by an annealing-based system was carried out "100 million times faster than [that of a] PC", but the comparison was not fair; the problem was greatly optimized for that demonstration and only slightly reflected real-world problems. Moreover, it was easy solvable by certain cluster-detecting algorithms, which were not used for comparison in the paper [54]. Anyway, while no application has been found yet where quantum annealing notably outperforms classical simulation approaches, the benefits of quantum annealing are becoming better understood and speed advantages have been demonstrated [55]. Further steps towards a universal quantum computer have been made, e.g., IBM presented a 50-qubit quantum processor in late 2017 [56]. While quantum key distribution (QKD) enables mathematically provable secure connections, and for which commercial systems have been offered since 2003, there have also been attacks on these systems that target weaknesses in their implementation. For example, Liu

and Sauge demonstrated a hack of QKD back in 2009, while the following year Xu et al. demonstrated the "phase remapping" attack.

In 2008, a paper by an author who called himself Satoshi Nakamoto was published, describing a peer-to-peer electronic cash system [57]. While the elements of the concept, cryptographic signatures, Merkle Chains, and P2P-networks, had already been known, the author was able to solve the double-spending problem by combining them within a distributed, trustless consensus system. The further success of Bitcoin is well-known, but the underlying concept of Blockchains is much more powerful, as it is able to guarantee the integrity of arbitrary data and enable different applications in the area of cyber security [58].

In 2009, a search engine well-known among security researchers was founded: Shodan [59]. Unlike previous systems, Shodan scans the Internet for connected devices, looking at services and collecting all provided information [60]. Different techniques for handling data evolved, especially in the area of big data and cognitive systems. For example, IBM celebrated a great success for cognitive systems in 2011, when IBM's Watson computer won the Jeopardy! challenge [61]. Since then, Watson has been deployed in more and more areas, e.g., cancer treatment, financial planning, or advanced cyber threats and defence.

Much progress has also been observed in the area of Artificial Intelligence (AI). Going beyond the search in problem spaces or behaviour-based approaches, AI is opening up more and more fields, in creativity and even in consciousness [62]. For example, AI is already able to paint new art based on original drawings [63], or compose new music [64]. Of course, there have also been hurdles. Microsoft's Twitter chatbot "Tay" had to be shut down in 2016 after less than 24 hours because it began using racist language [65], while a team from MIT's Computer Science and Artificial Intelligence Laboratory tricked Google's AI into misidentifying pictures of turtles as weapons [66]. Nevertheless, the well-disposed AI program Sophia was granted citizenship in Saudi Arabia in 2017. Now, Sophia is calling for women's rights [67]. Also, Google's already well-known AlphaGo AI system was enhanced even further in a very interesting and powerful away, being no longer constrained by the limits of human knowledge, but learning tabula rasa from itself and outperforming all previous systems [68].

From a military perspective, in 2013 the Chief of the General Staff of the Russian Federation Armed Forces, General Valery Gerasimov, published an article highlighting the asymmetrical possibilities offered by cyberspace and the necessity of perfecting activities in this information space [69]. In summary, the approach is guerrilla, and waged on all fronts with a range of actors and tools – for example, hackers, media,

businessmen, leaks and, yes, fake news, as well as conventional and asymmetric military means. Chaos is the strategy the Kremlin pursues, Gerasimov specifies that the objective is to achieve an environment of permanent unrest and conflict within an enemy state [70]. In November 2014, the US Secretary of Defense Chuck Hagel announced the "Defense Innovation Initiative" [71], with the aim being to "pursue innovative ways to sustain and advance our military superiority for the 21st Century" [71]. To stop the erosion of American dominance in key domains in waging war, it is necessary, he argued, to find "new and creative ways to sustain and in some areas expand our advantages even as we deal with more limited resources" [71]. While this sounds quite challenging, it is historically motivated: "The US changed the security landscape in the 1970s and 1980s with networked precision strike, stealth, and surveillance for conventional forces. We will identify a third offset strategy that puts the competitive advantage firmly in the hands of American power projection over the coming decades" [71].

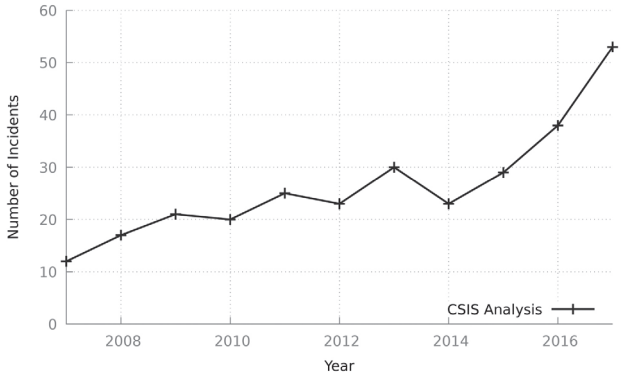
As new and challenging technologies are emerging with increasing pace, many of which are part and parcel of cyber security, a closer look at the root causes of the incidents between 2007-2017 incidents is necessary.

B. Attacker vs. Defender

Today, hundreds of cyber security systems are available on the market. Already back in 2004, the market research company International Data Corporation (IDC) coined the term "Unified Threat Management" (UTM). Basically, UTM is the evolution of firewall techniques into a comprehensive security solution, containing areas like control usage and policy enforcement, and therefore, combining techniques like content filtering, intrusion detection, and prevention, DDoS mitigation and antivirus applications. However, in spite of so broad and extensive an approach to cyber security, cyber security incidents are on the rise in number and gravity. Indeed, as with antivirus software being the conduit for hackers being able to expose data on a NSA contractor's laptop, we are repeatedly seeing how products intended to protect the system have become the gateway for attackers (e.g., see [72]). It is not unsurprising then that in 2015 Netflix chose to discard its antivirus systems [73].

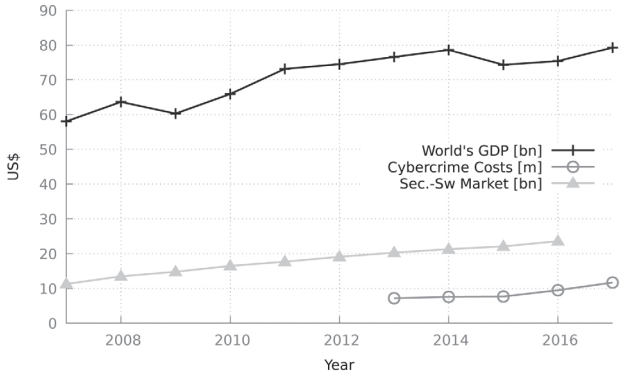
As a definition of an "incident" we should exclude any "ping" or an attempted connection from an unknown machine, as they generate huge numbers, but for the most part are of no greater significance. An attempt was made but failed. Far better then to concentrate on events where huge amounts of personal data or confidential files or even money have been stolen, or where physical damage has been wrought. The Center for Strategic and International Studies (CSIS) has recorded incidents [74], with Figure 2 charting occurrence since 2007.

FIGURE 2. NUMBER OF SIGNIFICANT CYBER INCIDENTS SINCE 2007 AS RECORDED BY CSIS [74]



It is particularly noticeable how the number of significant cyber security incidents has risen since 2015. However, a major variable is the efficacy of protective measures, which can be affected by numerous factors, including falling investment in cyber security. Hence figures for investment in cyber security are included in Figure 3, which shows investment in cyber security has consistently risen [75–77], even in the years of global economic crises when overall GDP has contracted [78].

FIGURE 3. EVOLUTION OF THE GLOBAL GDP, THE CYBER SECURITY MARKET REVENUE AND THE AVERAGE COST OF A CYBER INCIDENT BASED ON [75-78]



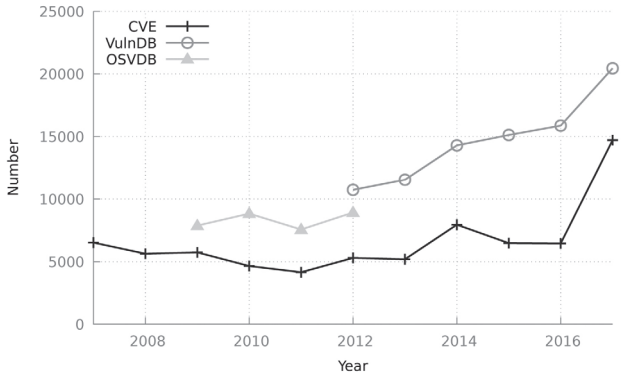
Estimating the net loss generated by cybercrime is a challenging task. Official numbers published by government or non-governmental bodies are a weak indicator, as only those cases filed with them are included. Additional data can be harvested from companies engaging in surveys on the matter, but this may still only be scratching the surface. Various public and private sources produce reports on a regular basis, but even when comparing the same periods under review, there is no consistent picture regarding cyber-attack statistics. For example, IDG's summary of PwC's Global State of Information Security Survey 2018 [79], published on October 18th, 2017, states

that “The number of security incidents detected continues to drop, along with the average financial loss due to cyber security attacks. However, the financial loss per incident continues to climb” [80]. In contrast, the 2017 Cost of Cyber Crime Study published by Accenture on September 26th, 2017, highlights a 27.4% *increase* in the average annual number of security breaches [81]. This underlines how the presented numbers cannot be generalized and how difficult it is to estimate the total damage, with the lack of reliable data being a core issue [82]. Recent studies by McAfee and CSIS throw some light on the subject by estimating the economic impact [83] and the global cost of cybercrime [84]. The most recent report suggests that the global cost of cybercrime is now US $600 billion, which includes gains to criminals and costs to companies for recovery and defence [84]. Within the studies, McAfee highlights the importance to include certain additional indirect costs, such as reputational damage, and this is also emphasized by Anderson et al. [85].

Thus, the global damage has increased sharply since the calculations for the period 2013-2014 where the estimation was US $400 billion [86]. The data is on a par with calculations from the British insurance company Lloyd’s. Anderson ultimately concluded, “that we should perhaps spend less in anticipation of computer crime (on antivirus, firewalls etc.) but we should certainly spend an awful lot more on catching and punishing the perpetrators” [85].

While there are only a few studies dealing with net losses, a large number of cyber security reports are released. For example, Verizon’s Data Breach Investigations Report (DBIR) 2017 indicates that 75% of data breaches are perpetrated by outsiders and 25% involved internal actors [87]. For the tactics used, Verizon surmised that 62% of breaches featured hacking, 51% included malware, and 43% had been social engineering attacks [87]. Still, such numbers are too abstract to identify underlying problems. For example, Figure 4 shows the evolution of known vulnerabilities, as seen by Common Vulnerabilities and Exposures (CVE) [88], Open Sourced Vulnerability Database OSVDB [89], and Vulnerability Database VulnDB respectively [90] (OSVDB was discontinued in 2012, the same year VulnDB started). The range of very different identified vulnerabilities per year is striking. This may be due to some databases also including additional, non-publicly available information in their statistics. As vulnerabilities are the gateway for attackers, one would assume that an evaluation of this data brings light into the cyber security darkness. However, an in-depth analysis by Rory McCune showed that the technical evaluations of various reports are “built on faulty data at best” [91, 92]. As the used data is heavily biased, evaluations are not representative of real-world challenges and by that, are not the strongest of foundations upon which to base any counter action or cyber defence strategies.

FIGURE 4. DEVELOPMENT OF IDENTIFIED VULNERABILITIES AS SEEN BY CVE AND OSVDB/ VULNDB [88-90]



The publicly known vulnerabilities may also be biased [93]: publications strongly depend on the interest and knowledge of some researchers, e.g., the pattern of "local privilege escalation", where vulnerability numbers followed an expected pattern based on the knowledge of researchers and their activities [94]. Therefore, analysing the vulnerability databases is not enough to obtain a comprehensive and accurate picture of the scale of hazards and events encountered.

Overall, then, one can see that neither the array of figures and cases of cyber security breaching incidents, nor the evaluation of vulnerabilities or malign programs, can suffice to comprehensively address all the challenges posed by cyber security issues. Even as investment in cyber security constantly rises, so too are overall net losses growing, and strongly so. An analysis of the root causes is challenging, due to the inherent limits of available data, how it is collated, how incidents are defined, and other flaws and biases innate to any study. All these factors complicate the question then of what is to be done, what kind of effective measures can be deployed in defence. For all of that, efforts need to be made to construct a macro-level investigation of the global situation involving all actors and agents, to best identify the scale of the problems and what can be done about them.

C. Conclusions from 10 Years of Cyber (In)security

Looking back at 10 years of cyber security incidents and technical circumstances and development, a number of trends can be identified:

Trivial vs. Sophisticated Attacks: Although the techniques of attackers are becoming more advanced, it is often the more relatively trivial attacks that the media hypes up. As highlighted, the "hacking" of military drones in Iraq was nothing more than recording and displaying what was arguably accessible information. Still, it is the

technically simple assaults that can generate the most extensive effects, as seen by the DDoS attacks on Dyn. Despite infecting more than 300,000 systems in 150 countries and having dramatic consequences, WannaCry was also, at a technical level, rather rudimentary, in that it only exploited already known vulnerabilities.

On the other hand, it can be seen how there is a growing number of ever more sophisticated attacks. Stuxnet marked the beginning of a new class of cyber attacks, in that a cyber weapon was deployed that led to significant material damage. The same can be seen in the attacks on the power grid in Ukraine, which were long in preparation and affected various kinds of software and systems, including the manipulation of firmware of Industrial Control Systems (ICS). Also, our partial knowledge of the surveillance structures and tools of the NSA and CIA shows they are highly sophisticated, as revealed by the respective leaks.

The trend towards more sophisticated attacks and their development on the timeline leads to the next findings:

Preparation of the Battlefield: A variety of actions in recent years reveals how ever more comprehensively engagement on the cyber battlefield is being prepared. The number as well as the quality of attacks on critical infrastructures is rising, as are infiltration campaigns aimed at installing backdoor access. The preparation of access opportunities also can be seen by different attacks on the supply chain, introducing malevolent hardware that can manipulate whatever software is installed upon it. At the same time, comprehensive cyber espionage activities can be identified, focusing on military systems and developments, as well as blueprints for the development and testing of new cyber weapons being used in the field, e.g., like in the cases of TV5 Monde or the Ukrainian power grid.

Glassy Humanity: The amount and quality of breached data reaches a level that can severely affect many areas of life, but especially people in security-critical tasks and functions. The OPM breaches presented a very severe incident, including data relating to personnel security background checks, while the breaches of the casual dating platforms Ashley Madison and AdultFriendFinder contained very detailed, personal data. Also, medical and personal devices and trackers are collecting more and more data and are often poorly secured, putting them well within the range of hostile online forces.

Further, newly available services like satellite surveillance offered by the company Planet Labs Inc. [95] will make surveillance capabilities previously reserved for the military and states available for almost everybody.

Theoretical vs. Practical Security: Another aspect already visible in the real world and growing in importance is the tension between theoretical and practical security. While having new systems based on mathematically provable security systems like QKD, complex technical implementations open up almost inestimable possibilities for side channel attacks. Being a very powerful and evolving instrument, AI can support cyber security, but at the same time such systems may also produce unpredictable and unwanted results, based on their complexity and "black box" character.

Demonstration of Cyber Power: Finally, some cases of the demonstration of cyber power can already be identified. Eventually, Stuxnet turned out in the demonstration of cyber power, based on the change of the code and attacker behaviour, as well as the too intense public discussions and statements. In part the attack on TV5 Monde and the activities of The Shadow Brokers can be seen as demonstrations of cyber power.

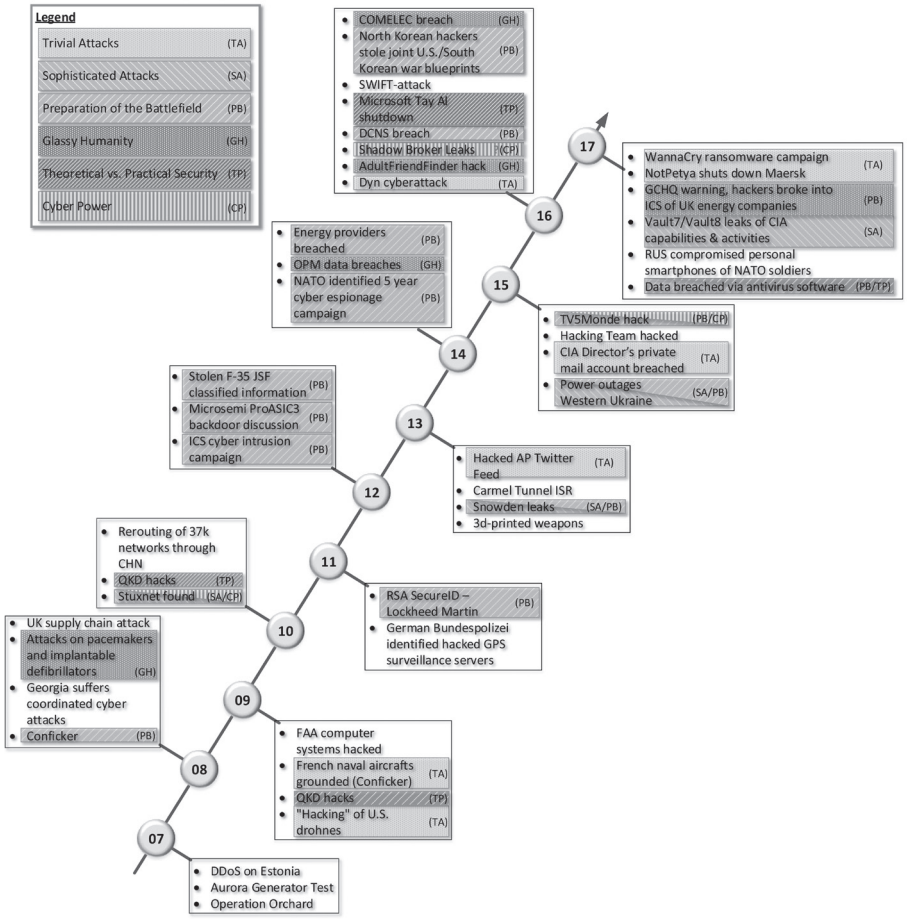
Figure 5 highlights the coherences between cyber security incidents and the derived characteristics.

D. The next 10 Years: A look into the Crystal Ball

Technological evolution is exponential, and IT improvements grow at an even super-exponential rate over long time spans [96]. This is something hard to cope with for human beings in daily life, and often decisions are taken based on a "linear feeling". Having a look at current research programs and activities coupled with various developments and announcements over the last few years provides glimpse of a picture of what we may expect.

The "Defense Innovation Initiative" already mentioned earlier is a good starting point to figure out how tomorrows technical world may look like. By having a look at related programs setup by DARPA, and state-of-the-art research, the following aspects can be identified:

FIGURE 5. COHERENCES OF IMPORTANT CYBER SECURITY INCIDENTS FROM 2007 TO 2017



First, the wide use of practically unlimited storage like 5d Glass discs and DNA storage [97] will challenge encryption security systems. Being able to store everything until one can decrypt it requires stronger encryption methods, and renders traditional concepts of proposed key lengths for certain periods of time as insufficient.

Second, Quantum supremacy will be achieved. While applications like QKD are already available, new techniques for secure communications will become ready for use. More so, universal quantum computers will open new opportunities for simulation, prediction, and security of systems, in the process supplanting and surpassing traditional security concepts. Recent research published by the University of Cambridge [98], IBM, and Intel shows tremendous progress in this area.

Robots and the “Soldier 4.0” concept (technical and bio enhancements to soldiers) will be much more powerful, but to the same degree, much more dependent on IT – ranging from the use of exoskeletons [99] to smart bandages for the faster healing of wounds or selectively erasing memories of trauma from the brain [100]. Of course, technology in \itself is morally inert – there is as much scope for misuse and abuse as there is for beneficent impacts benefiting all mankind.

Self-X technologies will find their way into products used in the real-world. While DARPA’s Cyber Grand Challenge in 2016 demonstrated the potential for self-defending systems to analyse attacks and patch themselves up, the setup was based on tiny, very limited operating systems. Anyway, it shows the future of cyber security, and related programs like “System Security Integrated Through Hardware and Firmware” (SSITH) [101] will raise the bar for attackers.

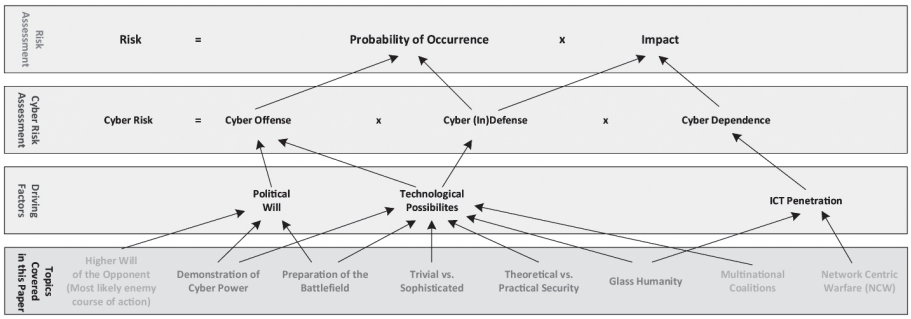
Finally, AI will only increase in power and come to pervade all areas of life, with algorithms achieving ever more superior performance with no human input. Together with more and more powerful and specialized hardware, e.g., self-learning neuromorphic chips that mimic brain functions [102], this will enable completely autonomous systems to operate independently in hostile environments, and much faster than any system reliant on human input in their loop.

Summing up these aspects highlights key elements of tomorrow’s forces: autonomous and collectively mission-executing systems that are produced cheaply and mobile via 3D-printing, that can self-destruct or dissolve in air so the technology will not fall into the hands of the adversary. While this can enable future supremacy on the battlefield even in denied environments (A2AD), the core requirement remains the same – strong cyber security, not only to protect the systems of tomorrow, but also to protect the research, design, and production that led to them.

4. CONCLUSION

Looking back on 10 years of cyber security, the situation seems to be becoming more and more challenging. Cyber is a popular tool for numerous reasons and many players. Upcoming technologies enforce hard and timely decisions, but the opportunities exist for a sustained improvement in cyber security. Figure 6 summarizes the paper and thus visually establishes a relationship between the Sections 2 and 3.

FIGURE 6. MAPPING OF THE IDENTIFIED COHERENCES AND DEDUCTIONS IN CYBERSPACE TO A RISK ASSESSMENT



Basically, risk can be seen as a mathematical product of the factors "probability of occurrence" and "impact of the damage". With respect to cyber, this equation is often extended to a three-factor equation:

Cyber Risk = Cyber Offense x Cyber Defence x Cyber Dependence

Section 2 has outlined how Russia's political will to use cyber weapons has increased. On that point, we have used examples of (i) economic subversion, and (ii) use of cyber attacks in Ukraine and Georgia. Technical developments are manifold and can be subdivided into the categories outlined. As per the actual realization of political will and technical capabilities in conducting cyber warfare, we have seen only the tip of the iceberg and, in the future, we will see cyber powers demonstrated far more often. Many of the attacks thus far could be ascribed to the notion of the "Preparation of the Battlefield". To be able to survive in a cyber war tomorrow, you have to do your homework today, and thus "prepare your opponent". It has to be assumed that countries such as Russia or China have quite different weapons at their disposal. Anyone who believes that these nations find cyber vulnerabilities "by accident" is wrong. Systematic preparation means deliberately finding and exploiting vulnerabilities on your opponent's side, if not indeed actively installing them in the soft- or hardware they may have sourced from you, and not waiting for "luck" to lean in your favour.

For the West, this means we have to think about cyber security more holistically and system-wide, especially in our military forces, and we need more innovative concepts with shorter procurement cycles. The topic of whether or not Western nations need a "critical security industry" is also an issue that needs to be discussed.

For the military, the power of future assertiveness means using NCW and autonomous systems. Fast decision-making requires information superiority and that in turn requires

ICT. Nonetheless, parallel to networking, greater autonomy and decentralization must be given greater consideration. To realize this complex task, sometimes less is more: in order not to end up in the "complexity trap", we should rather stick to the keep-it-simple approach, instead of looking for a vast, single, super solution. This means, in particular, the use of cost-effective systems, which are built to be mission-specific and on time using additive production methods and which are able to fulfil their missions on the basis of AI and swarm behaviour, even under A2AD conditions. Multi-billion dollar, high-value systems intended for use over decades, are only needed to a small extent as part of an overall strategy.

Furthermore, critical systems like weapon or crypto systems need verifiably secure designs. Trusted hardware for selected and highly-critical components as well as verified microkernels like seL4 are ways to realize this.

It is important to realize that the preparation of tomorrow's battlefield is happening now, resulting in backdoors in today's design and production. Therefore, better security along the supply line is required quickly and can be pushed by, for example, the use of Blockchain technologies.

Finally, disruptive technologies can have a huge impact on cyber security. For example, quantum computers will have a huge and immediate impact on cyber security when they are finally realized and deployed on a wholesale, real-world scale. Therefore, preparation is essential, even in the unlikely case that quantum computing does not get beyond the experimental lab stage. Thus, systems must be highly adaptive; for example, algorithms must be exchangeable quickly and comprehensively, but also structures and organizations must be flexible, being able to control and implement the required administrative processes.

REFERENCES

1. Ferbrache, D. (2003). "Network enabled capability: concepts and delivery". *Journal of Defence Science*. Vol. 8, No. 3, pp. 104-107.
2. Cebrowski, A. K., and Garstka, J. J. (1998). "Network-centric warfare: Its origin and future". In *US Naval Institute Proceedings*. Vol. 124 No. 1, pp. 28-35.
3. Alberts, D.S. (2002). "Information age transformation: getting to a 21st century military". *Command and Control Research Program (CCRP)*. Available at: http://www.dodccrp.org/files/Alberts_IAT.pdf [Accessed 10 Apr. 2018].
4. Wilson, C. (2007). "Network centric operations: background and oversight issues for congress". *Congressional Research Service*. Available at: http://www.au.af.mil/au/awc/awcgate/crs/rl32411.pdf [Accessed 10 April 2018] p. 1-55.
5. Lloyd, M. (2004). "Commanding mission groups: a speculative model". *9th International Command and Control Research and Technology Symposium (ICCRTS)*. Available at: http://dodccrp.org/events/9th_ICCRTS/CD/papers/122.pdf [Accessed 10 April 2018].

6. Alberts, D. S., Garstka, J. J., Hayes, R. E., and Signori, D. A. (2001). "Understanding information age warfare". *Command and Control Research Program (CCRP)*. Available at: http://www.dodccrp.org/files/Alberts_UIAW.pdf [Accessed 10 April 2018].
7. Febbraro, A. R., McKee, B., and Riedel, S. L. (2008). "Multinational Military Operations and Intercultural Factors". *NATO Research and Technology Organisation*. Available at: https://pdfs.semanticscholar.org/d67a/090784c6224fb8a3b230b628448e4b67ef0d.pdf [Accessed 10 April 2018].
8. Palin, R. (1995). "Multinational military forces: Problems and prospects". *The Adelphi Papers*. Vol. 35 No. 294.
9. Major, C., and Mölling, C. (2014). "The Framework Nations Concept: Germany's contribution to a capable European defence". *German Institute for International and Security*. Available at: https://www.swp-berlin.org/fileadmin/contents/products/comments/2014C52_mjr_mlg.pdf [Accessed 10 April 2018].
10. Glatz, R., and Zapfe, M. (2017). "Ambitious Framework Nation: Germany in NATO", *German Institute for International and Security* Affairs, Available at: https://www.swp-berlin.org/fileadmin/contents/products/comments/2017C35_glt_zapfe.pdf [Accessed 10 April 2018].
11. Pogodda, S., et al. (2014). "Assessing the impact of EU governmentality in post-conflict countries: pacification or reconciliation?". *European Security*. Vol. 23 No. 3, pp. 227-249.
12. Sanders, K. (2014). "Did Vladimir Putin call the breakup of the USSR 'the greatest geopolitical tragedy of the 20th century?' ". *Politifact.com*. Available at: http://www.politifact.com/punditfact/statements/2014/mar/06/john-bolton/did-vladimir-putin-call-breakup-ussr-greatest-geop/ [Accessed 10 April 2018].
13. Coyle, J. J. (2017). *Russia's Border Wars and Frozen Conflicts*. Cham: Springer.
14. Lesk, M. (2007). "The new front line: Estonia under cyberassault". *IEEE Security & Privacy*. Vol. 5 No. 4, pp. 76-79.
15. Meserve, J. (2007). "Sources: Staged cyber-attack reveals vulnerability in power grid". *CNN*. Available at: http://edition.cnn.com/2007/US/09/26/power.at.risk/ [Accessed 10 April 2018].
16. Makovsky, D. (2012). "The Silent Strike". *New Yorker*. Vol. 17, pp. 34-40.
17. Adee, S. (2008). "The hunt for the kill switch". *IEEE Spectrum*. Vol. 45 No. 5, pp. 34-39.
18. Nahorney, B. (2009). "The Downadup Codex - A comprehensive guide to the threat's mechanics". *Symantec Security Response*. Available at: https://www.symantec.com/connect/sites/default/files/the_downadup_codex_ed1_0.pdf, [Accessed 10 April 2018].
19. Asghari, H., Ciere, C. M., and Van Eeten, M. J. (2015). "Post-mortem of a zombie: Conficker cleanup after six years". *Proceedings of the 24th USENIX Conference on Security Symposium*. Available at: https://www.usenix.org/system/files/conference/usenixsecurity15/sec15-paper-asghari.pdf, [Accessed 10 April 2018].
20. Hypponen, M. (2009). "The Conficker Mystery". *BlackHat*. Available at: http://www.blackhat.com/presentations/bh-usa-09/HYPPONEN/BHUSA09-Hypponen-ConfickerMystery-PAPER.pdf [Accessed 10 April 2018].
21. Gorman, S. (2008). "Fraud Ring Funnels Data From Cards to Pakistan". *Wall Street Journal*. Available at: https://www.wsj.com/articles/SB122366999999723871 [Accessed 10 April 2018].
22. Marks, P. (2011). "Air traffic system vulnerable to cyber attack". *New Scientist*. Vol. 211 No. 2829, pp. 22-23.
23. Willsher, K. (2009). "French fighter planes grounded by computer virus". *The Telegraph*. Available at https://www.telegraph.co.uk/news/worldnews/europe/france/4547649/French-fighter-planes-grounded-by-computer-virus.html [Accessed 10 April 2018].
24. Arthur, C. (2009). "SkyGrabber: the $26 software used by insurgents to hack into US drones". *The Guardian*. Available at: https://www.theguardian.com/technology/2009/dec/17/skygrabber-software-drones-hacked [Accessed 10 April 2018].
25. Zmijewski, E. (2010). "Accidentally importing censorship". *Renesys Blog*. Available at: https://dyn.com/blog/fouling-the-global-nest/ [Accessed 10 April 2018].
26. Toonk, A. (2010). "Chinese ISP hijacks the Internet". *BGPMON*. Available at: https://bgpmon.net/chinese-isp-hijacked-10-of-the-internet/ [Accessed 10 April 2018].
27. Langner, R. (2013). "To kill a centrifuge: A technical analysis of what Stuxnet's creators tried to achieve". *The Langner Group*. Available at: http://www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf [Accessed 10 April 2018].
28. Hirvonen, T. (2013). "How RSA Was Breached". *PWC*. Available at: https://www.pwc.dk/da/arrangementer/assets/cyber-timohirvonen.pdf [Accessed 10 April 2018].
29. Gary, F. S. (2015). "New Snowden Documents Reveal Chinese Behind F-35 Hack". *The Diplomat*. Available at: https://thediplomat.com/2015/01/new-snowden-documents-reveal-chinese-behind-f-35-hack/ [Accessed 10 April 2018].

30. Majumdar, D. (2015). “America’s F-35 Stealth Fighter vs. China’s New J-31: Who Wins”. *The National Interest*. Available at: http://nationalinterest.org/blog/the-buzz/americasf-35-stealth-fighter-vs-chinas-new-j-31-who-wins-13938 [Accessed 10 April 2018].
31. Skorobogatov, S. and Woods, C. (2012). “Breakthrough silicon scanning discovers backdoor in military chip”. *International Workshop on Cryptographic Hardware and Embedded Systems*. Available at: https://www.cl.cam.ac.uk/~sps32/ches2012-backdoor.pdf [Accessed 10 April 2018].
32. Domas, C. (2017). “Breaking the x86 ISA”. *Blackhat*. Available at: https://www.blackhat.com/docs/us-17/thursday/us-17-Domas-Breaking-The-x86-Instruction-Set-wp.pdf [Accessed 10 April 2018].
33. Peter, F. (2013). “Bogus AP tweet about explosion at the white house wipes billions off US markets”. *The Telegraph*. Available at: https://www.telegraph.co.uk/finance/markets/10013768/Bogus-AP-tweet-about-explosion-at-the-White-House-wipes-billions-off-US-markets.html [Accessed 10 April 2018].
34. Macaskill, E., and Dance, G. (2013). “NSA files decoded: Edward Snowden’s surveillance revelations explained”. *The Guardian*. Available at https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded [Accessed 10 April 2018].
35. Koerner, B. I. (2017). “Inside the Cyberattack that Shocked the US Government”. *Wired*. Available at: https://www.wired.com/2016/10/inside-cyberattack-shocked-us-government/ [Accessed 10 April 2018].
36. Ejinsight (2014). “Russians hack NATO, EU and Ukraine in 5-year espionage”. *Ejinsight.com*. Available at: http://www.ejinsight.com/20141014-Russians-hack-NATO,-EU-and-Ukraine-in-5-year-espionage/ [Accessed 10 April 2018].
37. Symantec Security Response (2014). “Dragonfly: Western Energy Companies Under Sabotage Threat”. *Symantec*. Available at: https://www.symantec.com/connect/blogs/dragonfly-westernenergy-companies-under-sabotage-threat [Accessed 10 April 2018].
38. Corera, G. (2016). “How France’s TV5 Was Almost Destroyed by ‘Russian Hackers’ ”. *BBC News*. Available at http://www.bbc.com/news/technology-37590375 [Accessed 10 April 2018].
39. Lee, R. M., Assante, M. J., and Conway, T. (2016). “Analysis of the cyber attack on the Ukrainian power grid”. E*lectricity Information Sharing and Analysis Centre*. Available at: https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf [Accessed 10 April 2018].
40. Symantec Security Response (2016). “SWIFT attackers’ malware linked to more financial attacks”. *Symantec*. Available at: https://www.symantec.com/connect/blogs/swift-attackersmalware-linked-more-financial-attacks [Accessed 10 April 2018].
41. Sang-Hun, C. (2017). “North Korean Hackers Stole US-South Korean Military Plans, Lawmaker Says”. *New York Times*. Available at: https://www.nytimes.com/2017/10/10/world/asia/north-korea-hack-war-plans.html [Accessed 10 April 2018].
42. Boyd, C. (2016). “COMELEC breach data released online, fully searchable”. *Malwarebytes*. Available at: https://blog.malwarebytes.com/cybercrime/2016/04/comelec-breach-data-released-online-fully-searchable/ [Accessed 10 April 2018].
43. Evans, G. (2016). “Hacking the sting out of Scorpene: DCNS leak exposes secrets”. *Navaltechnology*. Available at: http://www.naval-technology.com/features/featurehackingthe-sting-out-of-scorpene-dcns-leak-exposes-secrets-5645820/ [Accessed 10 April 2018].
44. Whittaker, Z. (2016). “AdultFriendFinder network hack exposes 412 million accounts”, *Zdnet.com*. Available at: https://www.zdnet.com/article/adultfriendfinder-network-hack-exposes-secrets-of-412-million-users/ [Accessed 10 April 2018].
45. Solon, O. (2016). “Hacking group auctions ‘cyber weapons’ stolen from NSA”. *The Guardian*. Available at: https://www.theguardian.com/technology/2016/aug/16/shadow-brokers-hack-auction-nsa-malware-equation-group [Accessed 10 April 2018].
46. Hilton, S. (2016). “Dyn analysis summary of Friday October 21 attack”. *Dyn*. Available at: https://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack/ [Accessed 10 April 2018].
47. Chen, Q., and Bridges, R. A. (2017). “Automated Behavioral Analysis of Malware: A Case Study of WannaCry Ransomware”. *International Conference on Machine Learning and Applications (ICMLA)*. Available at: https://arxiv.org/pdf/1709.08753.pdf [Accessed 10 April 2018].
48. Lawler, R. (2017). “Microsoft patches Windows XP to fight ‘WannaCrypt’ attacks (updated). *Engadget*. Available at: https://www.engadget.com/2017/05/13/microsoft-windowsxp-wannacrypt-nhs-patch/ [Accessed 10 April 2018].
49. Thomson, I. (2017). “NotPetya ransomware attack cost us $300m - shipping giant Maersk”. *Forbes*. Available at: https://www.forbes.com/sites/leemathews/2017/08/16/notpetya-ransomware-attack-cost-shipping-giant-maersk-over-200-million/#7e0f16a34f9a [Accessed 10 April 2018].
50. Lubold, G., and Harris, S. (2017). “Russian Hackers Stole NSA Data on US Cyber Defense”. *Wall Street Journal*. Available at: https://www.wsj.com/articles/russian-hackersstole-nsa-data-on-u-s-cyber-defense-1507222108 [Accessed 10 April 2018].

51. MacAskill, E., Thielman, S., and Oltermann, P. (2017). "WikiLeaks publishes 'biggest ever leak of secret CIA documents' ". *The Guardian*. Available at: https://www.theguardian.com/media/2017/mar/07/wikileaks-publishes-biggest-ever-leak-of-secret-cia-documents-hacking-surveillance [Accessed 10 April 2018].
52. WikiLeaks, (2017). "Vault 7: Projects". *Wikileaks*. Available at: https://wikileaks.org/vault7/ [Accessed 10 April 2018].
53. Denchev, V. S., Boixo, S., Isakov, S. V., Ding, N., Babbush, R., Smelyanskiy, V., Martinis, J., and Neven, H. (2016). "What is the computational value of finite-range tunneling?" *Physical Review X*, Vol. 6 No. 3, p. 031015.
54. Mandrà, S., Zhu, Z., Wang, W., Perdomo-Ortiz, A., and Katzgraber, H. G. (2016). "Strengths and weaknesses of weak-strong cluster problems: A detailed overview of state-of-the-art classical heuristics versus quantum approaches". *Physical Review A*. Vol. 94 No. 2, p. 022337.
55. King, J., Yarkoni, S., Raymond, J., Ozfidan, I., King, A. D., Nevisi, M. M., Hilton, J. P., and McGeoch, C. C. (2017). "Quantum annealing amid local ruggedness and global frustration", arXiv preprint arXiv:1701.04579. Available at: https://arxiv.org/abs/1701.04579 [Accessed 10 April 2018].
56. Moore, S. K. (2017). "IBM Edges Closer to Quantum Supremacy with 50-Qubit Processor". *IEEE Spectrum*. Available at: https://spectrum.ieee.org/tech-talk/computing/hardware/ibm-edges-closer-to-quantum-supremacy-with-50qubit-processor [Accessed 10 April 2018].
57. Nakamoto, S. (2008). "Bitcoin: A peer-to-peer electronic cash system". *Bitcoin.org*. Available at: https://bitcoin.org/bitcoin.pdf [Accessed 10 April 2018].
58. Piscini, E., and Kehoe, L. (2018). "Blockchain & Cyber Security. Let's Discuss". *Deloitte*. Available at: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-blockchain-and-cyber-security-lets-discuss.pdf [Accessed 10 April 2018].
59. Shodan (2018). "Shodan.io". Available at: https://www.shodan.io/ [Accessed 10 April 2018].
60. Bodenheim, R., Butts, J., Dunlap, S., and Mullins, B. (2014). "Evaluation of the ability of the Shodan search engine to identify Internet-facing industrial control devices". *International Journal of Critical Infrastructure Protection*. Vol. 7 No. 2, pp. 114-123.
61. Gabbatt, A. (2011). "IBM computer Watson wins Jeopardy clash". *The Guardian*. Available at: https://www.theguardian.com/technology/2011/feb/17/ibm-computer-watson-wins-jeopardy [Accessed 10 April 2018].
62. Dehaene, S., Lau, H., and Kouider, S. (2017). "What is consciousness, and could machines have it?". *Science*. Vol. 358 No. 6362, pp. 486-492.
63. Higginbotham, S. (2015). "Who drew this? A computer... or Van Gogh?". *Fortune*. Available at: http://fortune.com/2015/08/31/ai-vangogh/ [Accessed 10 April 2018].
64. Goldhill, O. (2016). "The first pop song ever written by artificial intelligence is pretty good, actually". *Quartz*. Available at: https://qz.com/790523/daddys-car-the-firstsong-ever-written-by-artificial-intelligence-is-actually-pretty-good/ [Accessed 10 April 2018].
65. Hunt, E. (2016). "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter". *The Guardian*. Available at: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter . [Accessed 10 April 2018].
66. Staff, R. (2017). "How to Trick Google's AI into Thinking a Turtle is a Gun". *Robotics Business Review*. Available at: http://www.roboticstrends.com/article/how_to_trick_googles_ai_into_thinking_a_turtle_is_a_gun/Artificial_Intelligence [Accessed 10 April 2018].
67. Galeon, D. (2017). "Saudi Arabia Made a Robot a Citizen. Now, She's Calling For Women's Rights". *Futurism*. Available at: https://futurism.com/saudi-arabia-maderobot-citizen-calling-womens-rights/ [Accessed 10 April 2018].
68. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton A., (2017). "Mastering the game of go without human knowledge". *Nature*. Vol. 550 No. 7676, pp. 354-359.
69. Gerasimov, V. (2016). "The value of science is in the foresight: New challenges demand rethinking the forms and methods of carrying out combat operations". *Military Review*. Vol. 96 No. 1, p. 23.
70. Mckew, M. K. (2017). "The Gerasimov Doctrine". *Politico*. Available at: https://www.politico.com/magazine/story/2017/09/05/gerasimovdoctrine-russia-foreign-policy-215538 [Accessed 10 April 2018].
71. Hagel, C. (2014). "The Defense Innovation Initiative". *Department of Defense*. Available at: http://archive.defense.gov/pubs/OSD013411-14.pdf, [Accessed 10 April 2018].
72. Ormandy, T. (2016). "How to Compromise the Enterprise Endpoint". *Google*. Available at: https://googleprojectzero.blogspot.be/2016/06/how-to-compromise-enterprise-endpoint.html [Accessed 10 April 2018].

73. Fox-Brewster, T. (2015). “Netflix is dumping anti-virus, presages death of an industry”. *Forbes*. Available at: https://www.forbes.com/sites/thomasbrewster/2015/08/26/netflix-and-death-of-anti-virus/#7d88b11d18a5 [Accessed 10 April 2018].
74. Lewis, J. A. (2013). “Significant cyber incidents since 2006”. *Center for Strategic and International Studies*. Available at: https://www.csis.org/programs/cybersecurity-and-governance/technology-policy-program/other-projects-cybersecurity [Accessed 10 April 2018].
75. Landesman, M. (2017). “A Brief History of Malware”. Lifewire. Available at: https://www.lifewire.com/brief-history-of-malware-153616 [Accessed 10 April 2018].
76. Forni, A. A., and van der Meulen, R. (2016). “Gartner Says Worldwide Security Software Market Grew 3.7 Percent in 2015”. *Gartner*. Available at: https://www.gartner.com/newsroom/id/3377618 [Accessed 10 April 2018].
77. Deshpande, S. (2017). “Market Share: Security Software, Worldwide, 2016”. *Gartner*. Available at: https://www.gartner.com/doc/3698417/market-share-securitysoftware-worldwide [Accessed 10 April 2018].
78. Statista - Das Statistik-Portal (2018). “Weltweites Bruttoinlandsprodukt (BIP) in jeweiligen Preisen von 2007 bis 2017 (in Billionen US-Dollar)”. Available at: https://de.statista.com/statistik/daten/studie/159798/umfrage/entwicklung-des-bip-bruttoinlandsprodunkt-weltweit/ [Accessed 10 April 2018].
79. Price Waterhouse Coopers (2018). “The Global State of Information Security Survey 2018”. *PWC*. Available at: https://www.pwc.com/us/en/cybersecurity/information-security-survey.htm [Accessed 10 April 2018].
80. IDG Communications Inc. (2017). “2018 Global State of Information Security Survey”. *IDG*. Available at: https://www.idg.com/tools-for-marketers/2018-global-stateinformation-security-survey/ [Accessed 10 April 2018].
81. Ponemon Institute LLC (2017). “Cost of cyber crime study 2017 insights on the security investments that make a difference”. *Accenture*. Available at: https://www.accenture.com/t20170926T072837Z__w__/usen/_acnmedia/PDF-61/Accenture-2017-CostCyberCrimeStudy.pdf [Accessed 10 April 2018].
82. Armin, J., Thompson, B., Ariu, D., Giacinto, G., Roli, F., and Kijewski, P. (2015). “2020 cybercrime economic costs: No measure no solution”. *10th International Conference on Availability, Reliability and Security*, pp. 701-710.
83. Lewis, J., and Baker, S. (2013). “The economic impact of cybercrime and cyber espionage”. *McAfee*. Available at: https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/60396rpt_cybercrime-cost_0713_ph4_0.pdf [Accessed 10 April 2018].
84. McAfee (2018). “Economic Impact of Cybercrime - No Slowing Down”. *McAfee*. Available at: https://csis-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf [Accessed 17 April 2018].
85. Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., Moore, M., and Savage, S. (2013). “Measuring the cost of cybercrime”. In Böhme, Rainer ed., *The Economics of Information Security and Privacy*. Münster: Springer.
86. McAfee, (2014). “Estimating the global cost of cybercrime”. *McAfee*. Available at: https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/attachments/140609_McAfee_PDF.pdf [Accessed 10 April 2018].
87. Verizon (2017). “2017 Data Breach Investigations Report - 10th Edition”. *Verizon*. Available at: https://www.verizonenterprise.com/verizon-insights-lab/dbir/ [Accessed 10 April 2018].
88. The MITRE Corporation (1999). “Common Vulnerabilities and Exposures”. Available at: https://cve.mitre.org/ [Accessed 10 April 2018].
89. OSVDB (2018). “Open Sourced Vulnerability Database”. Available at: https://blog.osvdb.org/ [Accessed 10 April 2018].
90. Risk Based Security (2018). “Vulnerability Statistics”. Available at: https://vulndb.cyberriskanalytics.com/#statistics [Accessed 10 April 2018].
91. Raesene (2015). “Some potential problems extrapolating from data in security”. Available at: https://raesene.github.io/blog/2015/04/17/some-potential-problemsextrapolating-from-data-in-security/ [Accessed 10 April 2018].
92. Jerichoattrition (2015). “A Note on the Verizon DBIR 2015, ‘Incident Counting’, and VDBs”. Available at : https://blog.osvdb.org/2015/04/23/a-note-on-theverizon-dbir-2015-incident-counting-and-vdbs/ [Accessed 10 April 2018].
93. Christey, S., and Martin, B. (2013). “Buying into the bias: Why vulnerability statistics suck”. *BlackHat*. Available at: https://media.blackhat.com/us-13/US-13-Martin-Buying-Into-The-Bias-Why-Vulnerability-Statistics-Suck-WP.pdf [Accessed 10 April 2018].
94. Jerichoattrition (2017). “The Duality of Expertise: Microsoft”. *OVSDB*. Available at: https://blog.osvdb.org/category/vulnerability-statistics/ [Accessed 10 April 2018].

95. Planet Labs Inc. (2018). “Welcome to the insights economy”. Available at: https://www.planet.com [Accessed 10 April 2018].
96. Bui, Q. M., Nagy, B., Farmer J. D., and Trancik, J. E. (2013). “Statistical basis for predicting technological progress”. *PloS One*. Vol. 8 No. 2, p. e52669.
97. Birney, E., Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., and Sipos, B. (2013). “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA”. *Nature*. Vol. 494 No. 7435, pp. 77-80.
98. Cambridge Core (2018). “World’s first complete design of a silicon quantum computer chip”. *Cambridge University Press*. Available at: https://www.cambridge.org/core/journals/mrs-bulletin/news/world-s-first-complete-design-of-a-silicon-quantum-computer-chip [Accessed 10 April 2018].
99. Ackerman, E. (2015). “DARPA Tests Battery-Powered Exoskeletons on Real Soldiers”. IEEE Spectrum. Available at: https://spectrum.ieee.org/video/robotics/military-robots/darpa-tests-batterypowered-exoskeletons-on-real-soldiers [Accessed 10 April 2018].
100. Adler, K., Hu, J., Ferguson, L., Farah, C. A., Hastings, M. H., Sossin, W. S., and Schacher, S. (2017). “Selective erasure of distinct forms of long-term synaptic plasticity underlying different forms of memory in the same postsynaptic neuron”. *Current Biology*. Vol. 27 No. 13, pp. 1888-1899.
101. DARPA (2017). “Baking Hack Resistance Directly into Hardware”. *Defence Advanced Research Projects Agency*. Available at: https://www.darpa.mil/news-events/2017-04-10 [Accessed 10 April 2018].
102. Mayberry, M. (2017). “Intel’s New Self-Learning Chip Promises to Accelerate Artificial Intelligence”. *Intel*. Available at: https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/ [Accessed 10 April 2018].

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Aladdin's Lamp: The Theft and Re-weaponization of Malicious Code

Kārlis Podiņš
CERT Latvia
Riga, Latvia

Kenneth Geers
Comodo Group
Toronto, Canada

Abstract: Global superpowers do not have a monopoly on cyber warfare. Software thieves can steal malware written by more advanced coders and hackers, modify it, and reuse it for their own purposes. Smaller nations and even non-state actors can bypass the most technically challenging aspects of a computer network operation – vulnerability discovery and exploit development – to quickly acquire world-class cyber weapons. This paper is in two parts. First, it describes the technical aspects of malware re-weaponization, specifically the replacement of an existing payload and/or command-and-control (C2) architecture. Second, it explores the implications of this phenomenon and its ramifications for a range of strategic concerns including weapons proliferation, attack attribution, the fog of war, false flag operations, international diplomacy, and strategic miscalculation. And as with Aladdin's magic lamp, many malware thieves discover that obtaining a powerful new weapon carries with it risks as well as rewards.

Keywords: *malware, cyberwar, re-weaponization, false flag, attribution*

1. INTRODUCTION: STEALING CYBER WEAPONS

In *Arabian Nights*, a poor but clever Aladdin finds a magic lamp offering power, wealth, and love. However, the acquisition of these benefits also carried a burden of risk and responsibility. This parable offers lessons for aspiring cyber armies. The theft of advanced malware facilitates a similar shortcut to increased power on digital national security terrain. Computer code written by the Great Powers, including the United States, Russia, China, and Israel, can be acquired, reverse-engineered, and re-weaponized by small nations and even non-state actors.

Malware is a weapon unlike old-fashioned tanks and planes, and it is not necessary to break into a top-secret malware vault to steal it. Rather, compiled and fully-functioning cyber weapons can be found every day, by a careful observer, within network traffic and even on most email servers. And just as with Aladdin's magic lamp, these tools can be quickly repurposed for new operations, entirely distinct from what the malware was originally intended to do. Such malware theft can save thousands of hours of time and effort.

When Sir Isaac Newton said, "if I have seen further, it is by standing on ye shoulders of giants," [1] he was also presaging this phenomenon. Indeed, not just malware but all of today's software benefits from the millions of coders and hackers who came before. Precious little code today is written entirely from scratch. Instead, existing code is customized and/or has new features added to it. And this is only one example of the way in which IT has changed both the nature of power and the way in which power is transferred between people, organizations, and nations. This is true not only for source code, but also in the case of malware samples, where only access to executable code is available.

We know for a fact that malware re-weaponization is possible because we often see it within academic research[1] [2] [3] and in capture-the-flag (CTF) hacker competitions [4]. However, we have also seen reflections of it in real-world computer network operations by nation-states [5] [6]. Cyber actors and campaigns with names like DarkHotel, Lazarus, and TigerMilk have been seen throughout Asia, reusing attack code such as NetTraveler and Decafett in ways that also appear to incorporate false flags intended to cast blame on others during cyber operations [7].

One of the most prominent recent cases of malware source code theft involved the U.S. National Security Agency (NSA), from which code was allegedly stolen and released by the "Shadow Brokers" via the website *Wikileaks* in 2016. Reportedly, an NSA exploit named EternalBlue was leveraged in May 2017 to facilitate the WannaCry ransomware attack that targeted Windows computers and demanded Bitcoin payments. A month later, EternalBlue was used again to propagate the Petya ransomware, primarily against Ukraine. In March 2017, the Shadow Brokers also released malware allegedly developed by the CIA, again via *Wikileaks* [8].

What is a "cyber weapon"? To be sure, this term has been abused and exaggerated by analysts, journalists, and politicians, even when describing some well-known case studies [9]. And strangely, in some long-standing international conflicts, there seem to have been no known examples of cyber-attacks at all [10]. Part of the challenge in defining cyber-attacks and "cyber war" is the novelty of this new conflict domain.

1 The Bao paper cited here discusses an "automatic system" for identifying and replacing outer shellcode. Our discussion in this paper goes deeper and examines the escalation of privilege exploits, as well as a C2 replacement technique that appears perfect for false flag attacks.

On March 23, 2018, noted security researcher "The Grugq" explored this question in depth during a Black Hat conference keynote entitled, "A Short Course in Cyber Warfare." The Grugq referred to "Cyber" as the "5th Domain" of warfare, which is "literally a new dimension" and "much more complicated than anything we know." He explained that cyber-attacks comprise "Active," "Passive," "Physical," and "Cognitive" elements that can be employed in unique ways every time, making the next cyber-attack painfully hard to predict – and sometimes even to understand.[2]

For the purposes of this paper, the authors consider that a cyber-attack can be any information-based or kinetic operation designed to compromise the confidentiality, integrity, or availability of an IT system. In a national security context, such an operation must cause sufficient harm that it rises to the attention of national decision makers. It is this latter criterion that contributes to the definition controversy, as a final determination is subjective and open to political or business opportunism; however, this is a problem that certainly predates the Internet. Finally, the authors share the opinion that the malware sample analysed in this paper more than meets the requirement for a cyber weapon, as it contains two rare "zero-day" exploits and is specifically designed to give an attacker full remote-access to a target computer.

Here is what current U.S. policy states about "computer network operations": "Cyberspace is the most affordable domain through which to attack the United States. Viruses, malicious code, and training are readily available over the Internet at no cost. Adversaries can develop, edit, and reuse current tools for network attacks." [11].

The concept of malware theft via executable code manipulation (i.e. no access to source code) has also been addressed directly. In an August 2017 speech to a U.S. Department of Defense Intelligence Information Systems (DoDIIS) conference, Defense Intelligence Agency (DIA) Director Lt Gen Vincent Stewart said, "Once we've isolated malware, I want to reengineer it and prep to use it against the same adversary who sought to use it against us." [12].

Within the context of NATO, there is ample evidence that computer network operations have already risen to the highest level of importance. In 2016, NATO promised to defend allied cyberspace as it has land, sea, and air since the end of World War II. Further, it is now officially integrating cyber operations into its military plans [13] with the explicit goal of trying to deter cyber-attacks like those that have occurred in Estonia, Georgia, Ukraine, and the United States [14] [15].

The theft and re-weaponization of malware samples, in which hackers steal each other's executable code, swap existing payloads for custom munitions, and/or

2 As an example of an "Active" cyber-attack, The Grugq cited Israel's manipulation of the Palestine Liberation Organization's online financial resources; for "Passive" he cited China's "Operation Aurora" vs. Google in 2009; for "Physical" he cited Stuxnet; and "Cognitive" includes the doxing of the U.S. Democratic National Committee in 2016.

replace its command-and-control (C2) functionality, will increase the number of actors, attacks, and complexities on the cyber battlefield, and will negatively impact deterrence, diplomacy, and arms control in cyberspace.

This paper is divided into two primary sections: 1) a description of the technical aspects of malware re-weaponization, and 2) an exploration of its strategic implications.

2. MALWARE RE-WEAPONIZATION: TECHNICAL ASPECTS

In this section, the authors will examine the first part of their argument: that malware analysis is not "rocket science" and that executable code of any type can be captured, reverse-engineered, and repurposed with relative speed and ease. We will look at a genuine malware sample that was detected on a live network in 2017.[3] We believe that this malicious program was used by a nation-state with the specific intent of breaching a well-defended computer network. By any measure, it is advanced code, in part due to the fact that the program leverages no fewer than two "zero-day" exploits.[4]

The key takeaway from this short analysis is that the most technically challenging part of a cyber-attack's lifecycle – its vulnerability discovery and exploit development – can simply be stolen from another cyber actor. A malware thief (or cyber army) can then reconfigure and repurpose the code, adding unique functionality and/or control data, and then launch a high-grade cyber weapon in any direction they choose.

FIGURE 1. RUSSIAN DOLL [16]

3 The authors do not go into sufficient detail to allow the reader to create a live weapon. Specific technical details such as exact byte offsets are omitted.

4 "Zero-day" exploits target computer vulnerabilities that are yet unknown to software makers and security researchers; an exploit ceases to be a zero-day once specific patches are available.

A. Malware and its Russian doll design

Computer programs, including malware, are characterized by a layered structure that can be compared to a Russian matryoshka doll. With malware, most of the layers form a benign skeletal structure, while others (some of which can be hidden or encrypted) are designed to subvert computer security, hijack communications, or steal data.

1. Human layer
    - The outermost layer is that which humans see and understand, such as a Microsoft (MS) Office document. Our sample was an MS Word file sent via email. For an infection to begin, the email recipient simply had to open the attached file which had been expertly crafted by a phishing specialist.
2. Image file
    - Once opened, the MS file loaded an Encapsulated PostScript (EPS) image file that contained hidden, encrypted computer instructions in hexadecimal format[5] [17].
3. Shellcode
    - The decrypted code exploited a vulnerability in the Office EPS engine CVE-2017-0261 and executed shellcode that was embedded within the EPS file in order to open a command window through which an attacker could try to access the target computer.
4. Dropper
    - The shellcode was obfuscated (packed) and contained a Portable Executable (PE) file to be launched on the victim's computer. The executable file performed privilege escalation (CVE-2017-0263, individual exploits for 32- or 64-bit OS) and wrote a payload executable to disk.[6]
5. Payload
    - Once sufficient privileges were gained on the target computer, a "payload" was run, which was a fully-fledged remote administration tool that could perform a range of malicious actions such as stealing, blocking, and/or manipulating data. In our sample, the payload had been encrypted and compressed as an additional way to delay and complicate malware analysis.
6. Command-and-Control (C2)
    - After successful installation, the malware tried to "phone home" to a malicious C2 domain somewhere on the Internet in an attempt to report for duty, seek updates, and await further instructions. These communications were encrypted to help protect them from the prying eyes of network defenders.

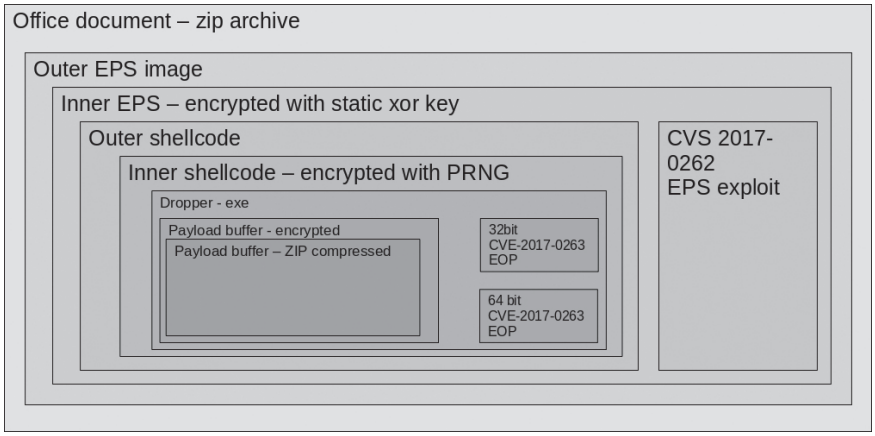
This level of malware analysis is not difficult and is available to any nation. Powerful tools such as code disassemblers and debuggers can perform decryption, de-

5 Researchers recently reported that multiple online threat actors, including Russian cyber espionage groups, have been leveraging EPS files and zero-days against European diplomatic and military entities.

6 These exploits took advantage of Common Vulnerability and Exposure (CVE) 2017-0263.

obfuscation, and unpacking of malware samples. In our case study, we employed numerous techniques. Some aspects of the design, such as the ".zip" algorithm and the "XOR cipher", are well-known to most malware researchers. Others, such as a string obfuscation algorithm for the C2, were custom-made by the malware's author, and required in-depth reverse-engineering.[7]

FIGURE 2. MALWARE SAMPLE ARCHITECTURE



B. Re-weaponization

Malware dissection at this level of detail already yields sufficient understanding for redesign and re-weaponization purposes. This section describes two ways to re-weaponize malware: 1) C2 replacement, and 2) payload replacement. Once either modification is performed, the malware thief simply reverses the steps taken in the malware's analysis, layer-by-layer, for the entire software package – just like a Russian doll.

The authors successfully tested both C2 replacement and payload replacement on this sample. They also wrote user-friendly command-line-interface scripts whereby even non-technical personnel, without any reverse-engineering knowledge, could perform the entire process.

1) Command and Control (C2) replacement

The quickest way to re-weaponize a malware sample is simply to replace its C2 components, such as by giving it a new domain that is under the malware thief's control. In fact, malware authors often reuse C2 architectures over time, even for

7 This effort required knowledge of the C programming language, as well as some luck. For example, one algorithm was symmetric, i.e. encrypt = decrypt. Asymmetric encryption could be defeated as well, but we would need to use a new encryption key and therefore the re-weaponized sample would be different from the original.

different exploits and malware campaigns. This typically serves to simplify ongoing operations which can grow in complexity over time. However, this characteristic also helps cyber defenders and malware thieves to analyse and reverse engineer how an attacker's C2 architecture works, both tactically and strategically.

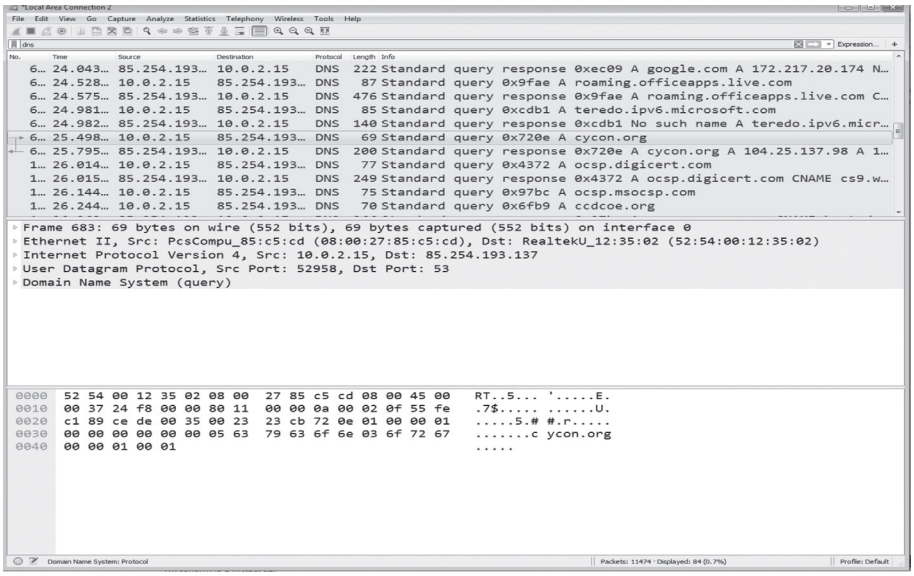
Replacing the C2 requires an intermediate level of technical expertise in software coding, reverse engineering, and network communications. But with the aid of disassembler software, this task can be accomplished relatively quickly, even by a small team or a lone expert. There can be technical limitations, such as with the length of the domain name. However, in practice, such limitations are easily overcome with some level of flexibility and creativity on the part of the malware thief.[8]

Finally, C2 replacement offers malware thieves an additional, tantalizing opportunity: the possibility of running easy false-flag operations. First, a re-weaponized malware sample is virtually indistinguishable from the original. Second, the malware thief can use the same service providers (including certificate issuers, hosters, DNS registrars, etc.) to make a new operation simply blend in with the campaign that the original attacker was already running, providing instant anonymity, or at least plausible deniability.

In Figure 3, below, the authors have written a small (120 lines of code) script to demonstrate the simplicity of C2 replacement. Here, there is just one command line parameter: the new C2 domain (cycon.org). All the necessary steps to replace the C2 domain in the malicious EPS file have been automated in an easy-to-use script. Running "python changeCnC.py cycon\.org [epsOutputFile" produces a malicious EPS file that can be included in a Word document. Once the malicious Word document is opened, malware infects the computer and connects to the modified C2 domain (cycon.org, as seen in the example screenshot). The primary challenge regarding C2 replacement is that one needs to reverse-engineer the C2 communication protocol and write server-side software to support this protocol.

8 For example, there are many ways to generate a short domain name, and to verify that it works, before an attack is launched.

FIGURE 3. C2 REPLACEMENT TO CYCON.ORG



2) Payload replacement

A second option for a would-be malware thief is to replace the payload with a tailored munition of their choice. For many scenarios, this is in fact the preferred option for a malware thief, such as:

1. when the thief already possesses custom agent and server software;
2. time constraints do not allow for C2 reverse-engineering; or
3. a proposed operation has easily achievable objectives such as wiping all data on the victim's machines.

Payload replacement is more invasive than C2 replacement and requires more malware expertise. As with C2 replacement, there can be some technical limitations, such as payload size. However, these can also be overcome with some flexibility and creativity after which the attacker can download additional malware modules via the Internet.

3. STRATEGIC IMPACT

In the previous section, the authors established that, even with limited time and expertise, a malware thief can reverse-engineer advanced malware, replace its C2 architecture, or replace its payload with a tailored munition, and launch an entirely new attack. In this section, we will explore the ramifications of this phenomenon for cyber defenders and for national security decision-makers[9] [18]. We will cover six strategic consequences in order from the logically most urgent and compelling to address to the least:

1. Proliferation
2. Attribution
3. Fog of War
4. False Flags
5. Diplomacy
6. Miscalculation

A. Proliferation

The first and most obvious challenge posed by malware re-weaponization is proliferation. Arms control, as a discipline, seeks to reduce the size of military arsenals that are capable of inflicting harm on humanity. But recycling malware means that the same vulnerabilities and exploits can be used by Country A against Country B, Country C against Country D, Country E against Country F, and so on. Furthermore, smaller nations and even non-state actors will sometimes be able to employ truly world-class digital weapons that would have been almost impossible for them to develop on their own.

So far, the cyber battlefield has seemingly been dominated by the Great Powers, such as the United States, Russia, and China, as well as regional powers with ongoing conflicts like Israel, Iran, and North Korea. Further, one experienced national security and cyber security specialist, James Lewis, recently argued that non-state actors are simply incapable of launching "massive and damaging" cyber-attacks [19]. But we suspect that most governments are, at the very least, leveraging computer network operations for cyber espionage in support of their core national security interests. We contend that malware theft and re-weaponization will only make this more common.

9 The Leitzel paper cited here, "Cyber Ricochet: Risk Management and Cyberspace Operations," uses the phrase "cyber ricochet" to denote denial-of-service attacks where the attacker does not directly communicate with the target but instead sends packets to intermediate nodes with spoofed source/destination addresses. The authors of this paper feel that the term "cyber ricochet", along with the label "reflection attack" which is used to describe a common hacker technique, imply that the malware thief is not directly controlling the operation and that an attack with unexpected consequences could result. However, when the payload or C2 infrastructure is wholly replaced, as we describe here, the attacker is in full control.

Above all, re-weaponization can save an aspiring cyber power significant time and money. IT and hacker talent are expensive. A credible cyber-attack program requires software developers, vulnerability analysts, exploit developers, malware testers, bot herders, and much more. In 2010, noted hacker and former NSA employee Charlie Miller told a CyCon audience that an effective cyber army would cost about $45 million per year with almost one-quarter of that sum spent on vulnerability analysts and exploit developers [20]. Thus, malware reuse offers a substantial reduction in cost for the most technically challenging parts of any operation: vulnerability discovery and exploit development.

B. Attribution

Increased cyber weapons proliferation means that there will be more armies on the cyber battlefield which in turn will increase the challenge of attribution. The digital battlefield has always been difficult for humans to see, understand, and contextualize. And three of the primary goals of a cyber-attacker are stealth, anonymity, or plausible deniability. Most cyber-attacks are closer to a covert operation than a traditional military operation. The laws of war state that soldiers should wear national uniforms with proper insignia, in part to bolster accountability for actions taken. However, hackers take advantage of the labyrinthine architecture of the Internet to obscure their true location.

The question of finding who is sitting at a remote keyboard is therefore fundamental to enhancing not only cyber security but also national security including deterrence, diplomacy, arms control, prosecution, and/or retaliation.[10] For computer network operations, this has been true since at least the mid-1980s.[11] Following the Cold War, and especially after the terrorist attacks of 9/11, law enforcement and counterintelligence agencies have invested considerable resources in cyber-attack attribution, but the size of the Internet and the dynamic nature of cyberspace have ensured that this will remain a vexing challenge for the foreseeable future.[12] Attribution is an art as well as a science, and a cyber-attack must usually cross a high threshold in terms of damages before sufficient resources will be allocated to its success [21].

Today, cyber defense is a professional discipline, and attribution is typically based on a wide range of observable tactics, techniques, and procedures (TTP).[13] However, in many cyber-attack investigations, there has been a singular, most valuable attribution

10 For example, in the 1990s, there were numerous cases in which the U.S. Government believed that a cyber-attack had been launched by a nation-state only to discover that it was a teenage student.

11 In the 1980s, Cliff Stoll, a system administrator at the University of California, Berkeley, spent a year tracking likely Russia-backed hackers who were targeting U.S. national laboratories, a tale recounted in *The Cuckoo's Egg*.

12 More recently, commercial firms have gotten into the attribution game. However, without the benefit of other sources of intelligence available to nation-states, such as human (HUMINT) and signals intelligence (SIGINT), they remain at higher risk of making mistakes in attribution.

13 Robust attribution relies on many pieces of evidence, including MD5 hashes, "diff" results, payloads, IP addresses, C2 infrastructure, domain names, digital certificates, network searches, exfiltrated data, source code, time zones, algorithms, encryption, current events, and more.

indicator: the malware "signature". Cyber actors have traditionally been associated with particular "families" of malware. Malware theft and re-weaponization therefore threatens to wreak havoc on the attribution process as we know it if an increasing number of players are simply using the same hacker tools that tend to be tightly controlled by their creator, and only accessible to others by malware reuse.

C. Fog of War

If already-challenging attribution becomes harder, national security decision-making will suffer from a thicker "fog of war". Sun Tzu famously wrote that "all warfare is based on deception" [22], but in the age of cyberwar, this dictum has never been more true. The problem is exacerbated by the fact that so many cyber-attacks take place during peacetime, either as cyber espionage or preparation of the battlespace for some future war that may never take place. Thus, in many ways, what we call "cyberwar" has no beginning – and no end.

The risks that cyber-attacks pose to our national critical infrastructures is high. Their integrity rests on the proper functioning of IT. This is true for everything from electricity to elections. Examples abound: in 2007, Syrian air defense personnel were apparently blinded by a cyber-attack that preceded an assault by Israeli warplanes; in 2015, foreign hackers are believed to have turned out the lights in Western Ukraine; and in 2016, Russian hackers were blamed for interfering in the U.S. Presidential election.

Malware theft and re-weaponization will increase the fog of war precisely because it increases weapons proliferation and hinders attack attribution. If all nations have access to roughly the same arsenal of vulnerabilities and exploits, who is to say that a third party is not playing *agent provocateur* in an ongoing conflict between two other nations? And how does any nation know when its cybersecurity has been compromised to the point that a traditional military invasion – or a coup d'état – is imminent? The chances for misunderstanding and miscalculation in cyberspace loom large indeed, especially in a conflict domain where time is of the essence.[14]

D. False Flags

Potential cyber-attackers know that the fog of war is thicker than ever. This fact will tempt many of them to engage in "false flag" operations that involve an effort to pin the blame on a third party. Such tactics long preceded the Internet, as pirate ships used to hoist false flags in an effort to prevent their targets from readying their defenses or evading the threat [23]. Modern spies also carry counterfeit passports, wear disguises, and lie about their true intentions.

14 Especially considering that the latest craze in both cyber-attack and defense is artificial intelligence (AI).

Malware theft and re-weaponization will tempt national-level decision-makers to engage in this type of behavior across the open Internet. False flag operations can be tricky to run as there are so many details to get right and so many ways that an operation can go wrong. But in cyberspace, the chances of success are higher, and the penalty for getting caught less severe than for a traditional military or intelligence operation. For most cyber operations, anonymity is not required, as plausible deniability will suffice.

Cyberspace is vast, and growing more crowded by the day, with students, soldiers, spies, and statesmen all living and working in the same space. There are 193 sovereign member states of the U.N., but there are 255 Internet country code top-level domains (ccTLD)[15]. This gives cyber-attackers the chance to be whomever they want, and suggests that malware reuse will increase the number of false flag political and military operations we see.

E. Diplomacy

If malware reuse is so helpful from an attacker's perspective, those who would seek to counter these advantages – law enforcement, counterintelligence, and diplomats – will have a more arduous road before them. Within the realm of international relations, the management of negotiations, treaties, and tension fall under the rubric of diplomacy. However, the rise of the Internet and cyberspace has complicated our understanding of both national security and diplomacy. There is only one Internet, and one cyberspace, and all nations are struggling to retain their traditional concepts of national sovereignty and law enforcement jurisdiction within it.

In 2018, diplomatic tensions over information security could hardly be higher. In cyber espionage, there are continuing reverberations over the Snowden revelations.[16] In propaganda, Russian interference in the U.S. electoral process has led to efforts throughout Europe to protect social media from information operations emanating from Moscow. And in nuclear diplomacy, cyber-attacks have been used by both sides on the Korean peninsula to improve their odds of victory in a real war.

Cyberwar is of special significance to diplomats for four reasons. First, cyber-attacks typically fall below the threshold of the use of force, so will be publicly addressed by diplomats more often than by soldiers. Second, most cyberwar occurs in peacetime when diplomacy takes priority over military operations. Third, diplomats are prime targets of an adversary's cyber espionage and influence operations. Fourth, alliance members risk getting dragged into a cyber conflict which they did not approve or even know about.

15 Internet country code top-level domains (ccTLD) encompass not only countries but also dependent territories.

16 For example, governments in Europe and South America have discussed building a new undersea cable in the Atlantic Ocean that could avoid direct digital contact with the United States.

Success or failure in diplomacy can have life-or-death consequences. Malware theft and re-weaponization will complicate cyber-related diplomacy, because of the expected rise in the number of actors, frequency of attacks, and the level of complexity of many cyber operations.

F. Miscalculation

History is littered with national security-related mistakes, from invading Russia to bombing Pearl Harbor, made by those who trusted in hope. It is human nature to be overly optimistic. And the theft of world-class malware is no different, carrying as it does risks for any malware thief. Attribution is difficult, but ultimately not impossible. It is easy to imagine that smaller nations, without sufficient political and military strength, will use such a weapon rashly and prematurely, and suffer disproportionate retaliation, in what could be a miscalculation of strategic proportions.

The fact that computer network operations are often time-sensitive only adds to this risk. When an attacker is able to pair an exploit (even a zero-day) with a discovered vulnerability, it is understood that the window of opportunity will not be open forever. A system administrator or software company can update, patch, or harden the target network, operating system, or application at any time. Malware signatures are constantly updated. And a malware thief has the added pressure of knowing that at least one other party knows about the exploit and vulnerability.

Even the possession of powerful malware does not mean that an attacker can properly execute all facets of a complex computer network operation. Part of it they may get right and others wrong. Hackers are routinely caught during any phase of a cyber-attack, from reconnaissance, to lateral movement on a network, during data exfiltration, and so on – sometimes even long after an attack is over. Incident response is always improving, and if done correctly, it will incorporate traditional intelligence analysis sources and methods as part of its attribution determination.

A final consideration involves stolen malware that has been backdoored, trojaned, or watermarked (potentially with malware theft in mind). Unless a hacker has written a computer program from scratch, it is hard to know whether it contains undiscovered, hidden functionality. For example, in 2013, the Syrian government allegedly targeted non-governmental organizations in Syria by encouraging them via social media to download Freegate, a common Virtual Private Network (VPN) client used to circumvent censorship. The Syrian government had reportedly trojaned this version of Freegate, precisely to target domestic opposition [24]. Thus, the desire for a quick, cheap cyber-attack can lead a malware thief into a trap.

4. CONCLUSIONS

For Aladdin, the acquisition of a magic lamp brought both benefits and risks. The theft and re-weaponization of malware is no different. Smaller nations, and even non-state actors, can obtain powerful digital weapons almost for free. As a result, there will be more armies on the cyber battlefield, more cyber-attacks, and a higher overall level of complexity for cyber defense. This phenomenon will have ramifications for weapons proliferation, attack attribution, the fog of war, false flag operations, international diplomacy, and strategic miscalculation.

If a malware thief asks too much of the magic lamp, however, there may be serious repercussions and unintended consequences. All cyber thieves must ask themselves whether they have the traditional political and military might to absorb a potential response. In this light, reliable attribution might still tend toward traditionally strong military powers – states that in any case may be less concerned with unforeseen consequences.

In terms of mitigating the potential impact of malware theft and re-weaponization, governments are likely to consider a wide range of options, including enhanced vulnerability disclosure,[17] watermarking digital weapons to keep closer track of them, the use of blockchain to enhance attribution, and even the signing of non-aggression pacts for cyberspace.[18] More research is needed on mitigation strategies.

In the longer term, it is possible that an increased awareness of this phenomenon will slow down the current pace of cyber operations worldwide, so that nations can better safeguard their code and operations. Potentially, this will serve to decelerate the prevailing level of conflict and instability in cyberspace, since every nation is now home to an abundance of cyber vulnerabilities. Advanced cyber powers might be wise to consider more carefully the potential fallout from approving reckless digital operations so that they do not lose control of the magic lamp.

5. REFERENCES

[1] I. Newton, "Letter from Sir Isaac Newton to Robert Hooke (1675)," Historical Society of Pennsylvania, 2017. Available: https://discover.hsp.org/Record/dc-9792/Details [Accessed March 28, 2018].

[2] R. W. Y. S. D. B. T Bao, "Your Exploit is Mine: Automatic Shellcode Transplant for Remote Exploits," IEEE Symposium on Security and Privacy, San Jose, 2017. Available: https://www.ieee-security.org/TC/SP2017/papers/579.pdf [Accessed March 28, 2018].

[3] Y. S. R. W. C. K. G. V. D. B. Tiffany Bao, "How Shall We Play a Game? A Game-theoretical Model for Cyber-warfare Games," Carnegie Mellon University, p. 1. Available: https://users.ece.cmu.edu/~youzhib/paper/bao2017csf.pdf [Accessed March 28, 2018].

17 Cyber commands already weigh the operational value of vulnerabilities and exploits against their national exposure to them.

18 There might be voluntary limits on cyber espionage, which is understood to be a natural precursor to cyber-attack.

[4] W. K. J. Yam, "Strategies used in capture-the-flag events contributing to team performance," Naval Postgraduate School, March 2016, pp. 2-3. Available: https://calhoun.nps.edu/bitstream/handle/10945/48498/16Mar_Yam_Jerel.pdf?sequence=1&isAllowed=y [Accessed March 28, 2018].
[5] J. Cox, "The CIA Allegedly 'Borrows' Code From Public Malware Samples," *Motherboard*, March 7, 2017. Available: https://motherboard.vice.com/en_us/article/3dyd53/the-cia-allegedly-borrows-code-from-public-malware-samples [Accessed March 28, 2018].
[6] G.-S. C. R. Juan Andrés, "Walking in Your Enemy's Shadow: When Fourth Party Collection Becomes Attribution Hell," in *Virus Bulletin*, Madrid, 2017. Available: https://cdn.securelist.com/files/2017/10/Guerrero-Saade-Raiu-VB2017.pdf [Accessed March 28, 2018].
[7] K. Zetter, "Masquerading Hackers are Forcing a Rethink of How Attacks are Traced," *The Intercept*, October 4, 2017. Available: https://theintercept.com/2017/10/04/masquerading-hackers-are-forcing-a-rethink-of-how-attacks-are-traced/ [Accessed March 28, 2018].
[8] M. R. A. W. L. Scott Shane, "WikiLeaks Releases Trove of Alleged C.I.A. Hacking Documents," *New York Times*, March 7, 2017. Available: https://www.nytimes.com/2017/03/07/world/europe/wikileaks-cia-hacking.html [Accessed March 28, 2018].
[9] M. D. Cavelty, "The Militarisation of Cyberspace: Why Less May Be Better," in *4th International Conference on Cyber Conflict*, NATO CCD COE, Tallinn, 2012. Available: https://ccdcoe.org/publications/2012proceedings/2_6_Dunn%20Cavelty_TheMilitarisationOfCyberspace.pdf [Accessed March 28, 2018].
[10] R. C. M. Brandon Valeriano, *Cyber War Versus Cyber Realities: Cyber Conflict in the International System*, New York, NY: Oxford University Press, 2015, p. 2. Available: http://www.brandonvaleriano.com/uploads/8/1/7/3/81735138/cyber_war_versus_book_review_itp.pdf [Accessed March 28, 2018].
[11] "Strategic Cyberspace Operations Guide," United States Army War College, June 1, 2016, p. 7. Available: https://info.publicintelligence.net/USArmy-StrategicCO.pdf [Accessed March 28, 2018].
[12] I. Thomson, "US Military Spies: We'll Capture Enemy Malware, Tweak it, Lob it Right Back at Our Adversaries," *The Register*, 15 Aug 2017. Available: https://www.theregister.co.uk/2017/08/15/us_government_wants_to_reverseengineer_malware_to_fight_back/ [Accessed March 28, 2018].
[13] NATO, "NATO Defence Ministers Agree to Adapt Command Structure, Boost Afghanistan Troop Levels," November 9, 2017. [Online]. Available: https://www.nato.int/cps/en/natohq/news_148722.htm. [Accessed December 2017].
[14] T. E. Ricks, "NATO's Little Noticed but Important New Aggressive Stance on Cyber Weapons," *Foreign Policy*, December 7, 2017. Available: http://foreignpolicy.com/2017/12/07/natos-little-noticed-but-important-new-aggressive-stance-on-cyber-weapons/ [Accessed March 28, 2018].
[15] C. Bing, "Russia-linked Hackers Impersonate NATO in Attempt to Hack Romanian Government," *Cyberscoop*, May 11, 2017. Available: https://www.cyberscoop.com/dnc-hackers-impersonated-nato-attempt-hack-romanian-government/ [Accessed March 28, 2018].
[16] W. Commons, "Floral matryoshka set 2 smallest doll nested," [Online]. Available: https://commons.wikimedia.org/wiki/File:Floral_matryoshka_set_2_smallest_doll_nested.JPG. [Accessed December 2017].
[17] A. L. A. B. B. R. D. K. G. M. Genwei Jiang, "EPS Processing Zero-Days Exploited by Multiple Threat Actors," FireEye, May 9, 2017. Available: https://www.fireeye.com/blog/threat-research/2017/05/eps-processing-zero-days.html [Accessed March 28, 2018].
[18] B. Leitzel, "Cyber Ricochet: Risk Management and Cyberspace Operations," Center for Strategic Leadership, U.S. Army War College, 2012. Available: http://www.dtic.mil/dtic/tr/fulltext/u2/a568619.pdf [Accessed March 28, 2018].
[19] J. Lewis, "Fighting the Wrong Enemy, aka the Stalemate in Cybersecurity," *The Cipher Brief*, November 26, 2017. [Online]. Available: https://www.thecipherbrief.com/column_article/fighting-the-wrong-enemy-aka-the-stalemate-in-cybersecurity. [Accessed March 9, 2018].
[20] C. Miller, "Kim Jong-Il and Me: How to Build a Cyber Army to Defeat the U.S.," oral presentation at *CyCon and DEF CON 18*, Tallinn and Las Vegas, 2010.
[21] B. B. Thomas Rid, "Attributing Cyber Attacks," *The Journal of Strategic Studies*, vol. 38, no. 1–2, p. 4–37, 2015.
[22] S. Tzu, "Wikiquote," Creative Commons, [Online]. Available: https://en.wikiquote.org/wiki/Sun_Tzu [Accessed March 28, 2018].
[23] L. deHaven-Smith, *Conspiracy Theory in America*, University of Texas Press, Austin, 2013, p. 225.
[24] M. M.-B. John Scott-Railton, "A Call to Harm: New Malware Attacks Target the Syrian Opposition," The Citizen Lab, Toronto, June 2013. Available: https://citizenlab.ca/2013/06/a-call-to-harm/ [Accessed March 9, 2018].

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Cyber Law and Espionage Law as Communicating Vessels

Dr. Asaf Lubin
Post-Doctoral Cyber Research Fellow
Fletcher School of Law and Diplomacy
Tufts University
Medford, MA, United States

Abstract: Existing legal literature would have us assume that espionage operations and "below-the-threshold" cyber operations are doctrinally distinct. Whereas one is subject to the scant, amorphous, and under-developed legal framework of espionage law, the other is subject to an emerging, ever-evolving body of legal rules, known cumulatively as cyber law. This dichotomy, however, is erroneous and misleading. In practice, espionage and cyber law function as communicating vessels, and so are better conceived as two elements of a complex system, Information Warfare (IW). This paper therefore first draws attention to the similarities between the practices – the fact that the actors, technologies, and targets are interchangeable, as are the knee-jerk legal reactions of the international community. In light of the convergence between peacetime Low-Intensity Cyber Operations (LICOs) and peacetime Espionage Operations (EOs) the two should be subjected to a single regulatory framework, one which recognizes the role intelligence plays in our public world order and which adopts a contextual and consequential method of inquiry. The paper proceeds in the following order: Part 2 provides a descriptive account of the unique symbiotic relationship between espionage and cyber law, and further explains the reasons for this dynamic. Part 3 places the discussion surrounding this relationship within the broader discourse on IW, making the claim that the convergence between EOs and LICOs, as described in Part 2, could further be explained by an even larger convergence across all the various elements of the informational environment. Parts 2 and 3 then serve as the backdrop for Part 4, which details the attempt of the drafters of the *Tallinn*

Manual 2.0 to compartmentalize espionage law and cyber law, and the deficits of their approach. The paper concludes by proposing an alternative holistic understanding of espionage law, grounded in general principles of law, which is more practically transferable to the cyber realm.

Keywords: *international law, information warfare, espionage, cyber law, Tallinn Manual 2.0, sovereignty, diplomatic law, consular law, general principles of law*

1. INTRODUCTION

Here is a story in two parts. In Part I, the Defense Minister for the Republic of Scamdinavia is honey-trapped by an attractive showgirl. During the course of their secret affair, the showgirl introduces the Minister to a senior naval attaché from the Embassy of Cyberia. The Minister, who quickly befriends the attaché, invites the latter to visit his home. Upon arrival, the attaché creates a diversion and seizes the opportunity to enter the Minister's private office, placing a pen-shaped recording device on his desk and photographing top-secret documents pertaining to the Department's security contracts and research spending. As a result, a number of top-secret Department of Defense projects are jeopardized, and the Minister is forced to resign.[1]

The second part begins with a series of phishing emails, sent to a number of major corporations across Scamdinavia, by a private hacking group with support and direction from Cyberia's central intelligence agency. The emails contain a trojan downloader. Within an eight-month period, roughly 50,000 computers are infected by the malicious code. Exploiting zero-day vulnerabilities in Microsoft XML Core Services, the malware begins modifying Windows registries, poisoning local DNS caches, disabling antivirus programs, and sequencing certain information harvesting and hard disk wiping processes. As a result of the attack, a number of financial institutions in Scamdinavia are unable to provide services and take weeks to fully restore functionality, causing significant economic losses. To make matters worse, the

1 This hypothetical is loosely based on one of the biggest spy scandals and political controversies of the Cold War era, the 1961 Profumo Affair. At the centre of the public blunder stood John Profumo, then Secretary of State for War, who was discovered to have had a sexual affair with model and showgirl Christine Keeler. Keeler was also romantically involved with Evgenii Ivanov, a senior naval attaché at the Soviet Embassy and an officer of the Soviets' Main Intelligence Directorate. At Keeler's invitation, Profumo and Ivanov met and soon became friends. Relying on his intimate access to Profumo's home and office, Ivanov was able to photograph highly classified documents pertaining to allied contingency plans for the Cold War defense of Berlin, as well top-secret specifications of US spy planes and nuclear weapons. Secretary Profumo initially denied the allegations of impropriety raised against him, but he eventually was forced to resign from his post, a fact that played a role in hastening the end of Harold Macmillan's term as Prime Minister. For further reading see JONATHAN HASLAM, NEAR AND DISTANT NEIGHBORS: A NEW HISTORY OF SOVIET INTELLIGENCE, 207-209 (2015); Leon Watson, *I Did Betray My Country: Fifty Years After Profumo's Resignation, Christine Keeler Confesses She Passed Secrets to Russians*, DAILY MAIL (9 June 2013), available at http://goo.gl/kPyXQT.

secret data of major government contractors is breached, and a number of top-secret Department of Defense projects are jeopardized.[2]

Existing legal literature would have us assume that these two hypothetical scenarios are doctrinally distinct. The first scenario is a textbook example of interstate spying, and insofar as it is regulated at all, it is only subject to the scant, amorphous, and under-developed legal framework of *espionage law*.[3] The second scenario, on the other hand, involves an example of what is colloquially termed a "cyber attack", which is subject to an emerging, ever-evolving body of legal rules, known cumulatively as *cyber law*.[4] This dichotomy, however, is erroneous and misleading. In practice, espionage and cyber law function as communicating vessels, and so are better conceived as two elements of a complex system, Information Warfare (IW). The paper draws attention to the similarities between the practices – the fact that the actors, technologies, and targets are interchangeable, as are the knee-jerk legal reactions of the international community. In light of the convergence between peacetime low-intensity cyber operations and peacetime espionage operations, the two should be subjected to a

2 This hypothetical is inspired by the events that transpired in South Korea on 20 March 2013 and are commonly known as the "Dark Seoul" incident. The attack, which occurred at approximately 2:15pm, hit television broadcasters YNT and MBC, as well as banks KBS, Shinhan, Nonghyup, and Jetu. South Korea's communicating regulator, Park Jae-Moon, released a statement suggesting that: "unidentified hackers used Chinese IP addresses to contact servers of the six affected organizations and plant malware which attacked their computers." Based on previous practice of North Korea to spoof Chinese IP address, a number of high-ranking officials from South Korea pointed their finger to Pyongyang. For further reading see Jonathan A.P. Marpaung & HoonJae Lee, *Dark Seoul Cyber Attack: Could it Be Worse*, 6th Conference of Indonesian Students Association in Korea (7 July 2013), available at http://goo.gl/MgCI9u; *China IP Address link to South Korea Cyber-Attack*, BBC News (21 March 2013), available at http://goo.gl/wm43kQ.

3 As Prof. Chesterman has argued, intelligence exists "in a legal penumbra, lying at the margins of diverse legal regimes and at the edge of international legitimacy." Elsewhere he noted that: "despite its relative importance in the conduct of international affairs, there are few treaties that deal with it directly. Academic literature typically omits the subject entirely or includes a paragraph or two defining espionage and describing the unhappy fate of captured spies. For the most part, only special regimes such as the laws of war address intelligence explicitly. Beyond this, it looms large but almost silently in the legal regimes dealing with diplomatic protection and arms control." See Simon Chesterman, *The Spy Who Came In From the Cold War: Intelligence and International Law*, 27 MICH. J. INT'L. L. 1071, at 1072, 1130 (2006); Richard Falk, foreword, in ESSAYS ON ESPIONAGE AND INTERNATIONAL LAW v, v (Roland J. Stranger ed., 1962) ("traditional international law is remarkably oblivious to the peacetime practice of espionage. Leading treatises overlook espionage altogether or contain a perfunctory paragraph that defines a spy and describes his hapless fate upon capture"); Christopher D. Baker, *Tolerance of International Espionage: A Functional Approach*, 19 AM. U. INT'L. L. REV.1091, 1091 (2004) ("Espionage is curiously ill-defined under international law, even though all developed nations, as well as many lesser-developed ones, conduct spying and eavesdropping operations against their neighbors"); Gary D. Brown & Andrew O. Metcalf, *Easier Said Than Done: Legal Reviews of Cyber Weapons*, 7 J. NAT'L SEC. L. & POL'Y 115, 116 (2014) ("there is a long-standing (and cynically named) 'gentleman's agreement' between nations to ignore espionage in international law").

4 *See e.g.*, MICHAEL N. SCHMITT (ED.), TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (2nd ed., 2017); UN General Assembly Resolution on an International Code of Conduct for Information Security, UN Doc. A/66/359 (14 September 2011); Elaine Korzak, *UN GGE on Cybersecurity: The End of an Era?*, THE DIPLOMAT (31 July 2017), available at http://goo.gl/BSWfnm; Louise Arimatsu, *A Treaty for Governing Cyber-Weapons: Potential Benefits and Practical Limitations, in* 4TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT (Czosseck & Ziolkowski eds., 2012); Joseph S. Nye Jr., *The World Needs New Norms on Cyberwarfare*, THE WASHINGTON POST (1 October 2015), available at http://goo.gl/NuC4z7; Brad Smith, *The Need for a Digital Geneva Convention*, MICROSOFT ON THE ISSUES (14 February 2017), available at goo.gl/4xPN7F.

single regulatory framework, one which recognizes the role that intelligence plays in our public world order and which adopts a contextual and consequential method of inquiry.

Part 2 of this paper provides a descriptive account of the unique symbiotic relationship between espionage and cyber law. It further explains the reasons for this dynamic and applies its findings to the two hypothetical scenarios introduced above. Part 3 then situates the discussion surrounding this relationship within the broader discourse on IW, making the claim that the convergence identified in Part 2 could further be explained by an even larger convergence across all the various elements of the informational environment. Parts 2 and 3 serve as the backdrop for Part 4, which details the attempt of the drafters of *Tallinn Manual 2.0* to compartmentalize espionage law and cyber law, and the deficits of their approach. The paper concludes by proposing in Part 5 an alternative holistic understanding of espionage law, grounded in general principles of law, which is more practically transferable to the cyber realm.

2. LAW OF COMMUNICATING VESSELS

> "If you had a bent tube, one arm of which was the size of a pipe-stem and the other big enough to hold the ocean, water would stand at the same height in one as in the other. Thus discussion equalizes fools and wise men in the same way, and the fools know it."
>
> -Oliver Wendell Holmes[5]

The experiment described in the quote, what Justice Holmes called the "hydrostatic paradox of controversy", is merely the Justice's cynical take on a classic principle of fluid mechanics, according to which the levels of homogenous liquid in a system of connected containers will always aspire to be equal, since the pressures on those levels are equal. Thus, if additional liquid is added to one vessel, the liquid will immediately find a new equal level in all connected vessels. This image of the "communicating vessels" experiment carries with it a powerful metaphor, which has been used across the humanities and social sciences, from construing surrealist thought,[6] to characterizing international policies on torture.[7] In this paper, I argue that the trite principle could also be helpful in describing the dialectical relationship between espionage law and cyber law.

What do I mean by "espionage" and "cyber"? It is worth recalling that: "no

5 2 JOHN T. MORSE, LIFE AND LETTERS OF OLIVER WENDELL HOMES 40 (1896). The statement was made by Holmes in response to an article in *The Nation* which harshly criticized his philosophy.

6 ANDRÉ BRETON, COMMUNICATING VESSELS (Translated by Mary Ann Caws & Geoffrey Harris, 1990).

7 STEVEN DEWULF, THE SIGNATURE OF EVIL: (RE)DEFINING TORTURE IN INTERNATIONAL LAW 535-551 (2011).

internationally recognized and workable definition of 'intelligence collection' exists."[8] Similarly "there are no common definitions for Cyber terms – they are understood to mean different things by different nations/organizations".[9] Given these innate ambiguities, it is important that I provide working definitions for both terms at the outset of this paper. To begin with, I am only interested in those cyber and espionage operations that occur in peacetime, given that wartime spying and cyber warfare are more constrained by the rules of international humanitarian law, and in any event occur at a far lesser rate than their peacetime equivalents. Limiting myself to peacetime cyber operations further narrows the scope of cyber activities to be examined, as it excludes from review those operations that by their scale and effect are likely to trigger an international armed conflict or to provoke responses in self-defense. Our attention thus automatically shifts to Low-Intensity Cyber Operations (LICOs). These are "below-the-threshold" operations which have not only proven to be significantly costly in recent years, but are in fact commonplace, as Michael Schmitt notes: "Few, if any, cyber operations have [ever] crossed the armed attack threshold".[10]

With Espionage Operations (EOs), I tend to cast the net quite wide, using the terms "espionage", "intelligence collection", "surveillance", and "reconnaissance" interchangeably, thus rejecting method-based definitional distinctions. Instead, I use the term EOs to mean a peacetime operation which encompasses the following four elements: (1) the operation involves the gathering, analysis, verification, and dissemination of information of relevance to the decision-making process of a State or States or otherwise serves some State interests; (2) the operation is launched by agents of a State or States, or those with a sufficient nexus to the State or States in question; (3) the operation targets a foreign State or States, their subjects, associations, corporations, or agents, without the knowledge or consent of that State or those States; and (4) the operation involves some degree of secrecy and confidentiality, as to the needs behind the operation and/or the methods of collection and analysis employed,

8 Sulmasy and Yoo, *Counterintuitive: Intelligence Operations and International Law*, 28 MICH. J. INT'L. L. 625, 637 (2007).

9 *Cyber Definitions*, NATO Cooperative Cyber Defence Centre of Excellence, available at http://goo.gl/wtAkWP.

10 Michael N. Schmitt, *"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law*, 54 VA. J. INT'L L. 697, 698 (2014). For further reading on the nature of LICOs see: Beatrice Waldon, Note, Duties Owed: Low-Intensity Cyber Attacks and Liability for Transboundary Torts in International Law, 126(5) YALE L. J. 1242 (2017). See also James R. Clapper, Statement of the Record, US Cybersecurity and Policy, Senate Armed Services Committee (29 September 2015), available at goo.gl/aWSgKH (where Clapper makes an alarming prediction: "we foresee an ongoing series of low-to-moderate level cyber-attacks from a variety of sources over time, which will impose cumulative costs on US economic competitiveness and national security").

so to ensure its effectiveness.[11] Notice that I exclude from review various forms of unconcealed open-source intelligence gathering, such as reading a newspaper, visiting a social media website, or gathering information in the course of routine diplomatic relations (element 4). I further exclude from my analysis domestic forms of surveillance focusing solely on interstate activities, launched by one State and its proxies (element 2) against another State and its proxies (element 3).

Already visible is the close proximity in nature between EOs and LICOs, for our definition of LICOs could also be limited only to interstate interactions (especially if we are to distinguish between LICOs and more local forms of domestic cyber crime). The only difference, therefore, between EOs and LICOs rests on the first element. Unlike EOs, LICOs can only be employed against electronic information (as opposed to non-electronic physical properties, *e.g.* a passport kept in a dresser or printed bank records stored in a cabinet). Moreover, LICOs are different as they may extend beyond the mere passive copying and storing of data to other more aggressive and coercive forms of electronic intrusion (e.g. altering, removing, disrupting, degrading, or destroying certain information, programs, systems, or networks).[12]

Therefore, if we put EOs and LICOs in a Venn diagram (see below in Figure 1), not only will the circle-circle overlap be significant (encompassing different types of cyber espionage and electronic surveillance operations), but the remaining sets will share profound similarities. I provide below a list of hypothetical examples of operations which are either exclusively EOs, exclusively LICOs, or in between, to exemplify those similarities.

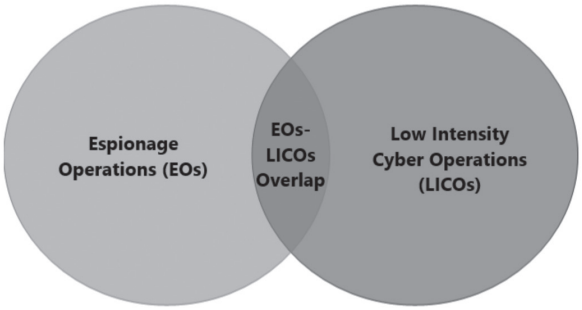
It is this affinity between EOs and LICOs that creates the "communicating vessels" phenomenon. Any attempt to modify or extend existing bodies of international law to better regulate LICOs will inevitably result in tidal waves that will engulf EOs.

11 This definition mirrors in some respects, and departs from in others, the definition put forward by Dermarest: "espionage can be defined as the consciously deceitful collection of information, ordered by a government or organization hostile to or suspicious of those the information concerns, accomplished by humans unauthorized by the target to do the collecting" (Geoffrey Dermarest, *Espionage in International Law*, 24 DENV. J. INT'L L. & POL'Y 321, 326). Note that as highlighted in Dermarest's definition, and as a general rule, intelligence operations involve some degree of secrecy and confidentiality to ensure their effectiveness (operations *de cape et d'épée*, coupled with some degree of deceitful intent). That said such is not always mandated (*e.g.* open source intelligence collection).

12 Note that my definition of EOs excludes "covert action" operations. These types of activities have a different primary purpose than the acquisition of intelligence. They seek the "purposive attenuation of the options of the target", influencing economic, ideological, political, diplomatic, and military conditions abroad. See W. MICHAEL REISMAN AND JAMES E. BAKER, REGULATING COVERT ACTION: PRACTICES, CONTEXTS, AND POLICIES OF COVERT COERCION ABROAD IN INTERNATIONAL AND AMERICAN LAW 10-12 (1992) (Reisman and Baker provide a useful list of examples of covert activities ranging from psychological operations and disinformation to political assassinations). If I were to include covert action into the definition of EOs, additional similarities between EOs and LICOs will surface (consider, for example, Russian interferences in elections as reflecting both covert action and a "below the threshold" cyber intrusion). In other words, expanding the definition of EOs to include covert action will entail extending its purpose beyond "mere passive copying and storing of information to other more aggressive types" of intrusions (namely the disruptive, degrading, and destructive kind).

Conversely, any attempt at normative compartmentalization or regulatory insulation could be equated to challenging a law of physics and would not pass the smile test.

FIGURE 1: VENN DIAGRAM OF EOS AND LICOS INTERSECTION



Exclusively EOs:

- Launching a spy satellite into space to engage in geo-spatial monitoring of a rogue country.
- Placing human agents in a major oil company, gathering information about its strategic plans.
- Gathering information about a government ministry relying on diplomatic engagements and open source materials.
- Placing cameras and microphones in the apartment of cyber criminals and monitoring their business dealings.
- Entering a training camp for a terrorist organization and seizing certain documents relating to an impending attack.

Exclusively LICOs:

- Jamming the communications links of a commercial satellite and sending it false GPS coordinates.
- Launching a ransomware attack against a major oil company, shutting down its operations for a short period.
- Launching a DDoS operation against a non-essential government service website.
- Hacking the devices of cyber criminals and blocking their access to a certain cryptocurrency.
- Installing malware on laptop computers at a terrorist training camp, circumventing a terrorist plot by altering certain data stored therein.

EOs-LICOs Overlap:

- Hacking a spy satellite for the purpose of gathering information about its technical specifications.
- Installing malware on the tablet of an oil company's CEO to gather information about the company's strategic plans.
- Hacking the DNS server of a government ministry and monitoring the internet activities of the ministry's staff.
- Hacking the devices of cyber-criminals and monitoring their business dealings by remotely activating certain sensors.
- Installing spyware on laptop computers at a terrorist training camp, and seizing certain documents relating to an impending attack.

To further my point, let us examine some areas of convergence between peacetime EOs and LICOs. First, both passive intelligence collection and mildly more aggressive cyber intrusions are launched by the same primary actors – State intelligence and security agencies and/or their proxies – and using the same advanced technological tools. Unit 8200 of Israel provides one good example,[13] and APT33 with its ties to Iran's Cyber Army offers another.[14] This reality is owed in part to the fact that traditional EOs now rely heavily on cyber techniques to increase their likelihood of success and broaden their scope of impact. For 16th century Sir Francis Walsingham, the father of modern intelligence agencies, "a global mass surveillance program involved paying off travellers in the ports of Lyon and merchant adventurers in the bazars of Hamburg".[15] Today, we cannot imagine an intelligence agency that would be satisfied with such low-tech techniques. SIGINT-based tools, such as the hacking of connected devices and the interception of electronic communications (either targeted or in bulk) have now significantly overshadowed the old historical techniques. The rise to predominance of Cyber-HUMINT, as its own distinct discipline, proves that even the most traditional of spying methodologies is not immune from this wave of digitalization.[16] Once an agency controls a band of cyberspies, calibrating between passive collection and moderately more offensive intrusions is left to its discretion and capacity limitations. So it is not surprising that the NSA is hoarding zero-day

13 John Reed, *Unit 8200: Israel's Cyber Spy Agency*, FINANCIAL TIMES (10 July 2015), available at goo.gl/951paE.

14 Eric Auchard, *Once 'Kittens' in cyber spy world, Iran gains prowess: security experts*, REUTERS (20 September 2017), available at https://goo.gl/DCmDkf; Jaqueline O'Leary *et al., Insight into Iranian Cyber Espionage*, FIREEYE (20 September 2017), available at https://goo.gl/vcS6Wc.

15 Asaf Lubin, *A Principled Defence of the International Human Right to Privacy: A Response to Frédéric Sourgens*, 42(2) YALE J. INT'L. L. 1, 2 (2017).

16 Andy Greenberg, *Cyberespionage is a Top Priority for CIA's New Directorate*, WIRED (9 March 2015), available at goo.gl/YWp5Zx (discussing the CIA's "digital overhaul" and quoting Jim Lewis from the Center for Strategic and International Studies, who notes: "Those 'humint' operations, as the intelligence community calls them, typically involve real spies on the ground, unlike the NSA's remote cyber espionage or the cyberwarfare activities of the Pentagon's Cyber Command. 'This kind of cyber activity has become increasingly important to them' … That combination of humint and digital operations could mean a spy infiltrating an organization to plant spyware by hand, for instance, or a digital investigation to check the bona fides of a source or agent. 'If you think of NSA operations as a vacuum cleaner and Cyber Command as a hammer, this is a little more precise, and it's about supporting human operations'").

vulnerabilities,[17] that the CIA controls a whole vault of cyber tools,[18] or that the FBI hacks thousands of foreign computers in the dark web with a trove of malware.[19]

Second, both EOs and LICOs thrive on "plausible deniability" and demand increased levels of deception and secrecy, intrinsically resisting mechanisms of accountability. Think of an undercover agent who is masquerading one day as a 30-year-old Danish female protester at a reproductive rights rally and the next day as a 55-year-old German wheelchair-bound male social worker. Now think of the Chinese hacker who is spoofing his way through the Tor network, one day hijacking the computer of a real Danish protester and the next adopting the online identity of an actual German social worker. Both operations, due to their unique nature, create similar and significant evidentiary hurdles for assigning individual and State responsibility under traditional international legal frameworks.[20]

Finally, both EOs and LICOs target information in ways that are non-kinetic and below-the-threshold, triggering the same knee-jerk international legal reactions. The victims of spying and cyber operations have a limited basket of potential claims that they might raise for a violation of international law, namely: violations of sovereignty, territorial integrity, the principle of non-intervention, the prohibition on extraterritorial enforcement, certain human rights abuses (such as the rights to privacy and freedom of expression), certain property rights abuses (including IP rights), and other potential State and individual immunities and privileges, depending on the subject matter of the operation.[21] What is more, common to both EOs and LICOs is the fact that the international norms enumerated in the above list are sufficiently under-defined to leave ambiguity as to whether an actual violation of a primary rule of international law had occurred. The *Tallinn Manual 2.0* was in this regard an attempt to clarify (if not codify) the "key aspects of the public international law governing 'cyber operations' during peacetime".[22] Put differently, the experts in *Tallinn 2.0* sought to elucidate the law of LICOs in isolation from the law on EOs. As I will show later, this unfortunate compartmentalized approach adopted by the *Manual*'s authors proves counterproductive at offering effective regulation. For now, let me conclude this section by showing in Table 1 how the two hypotheticals that opened this paper exemplify the convergence between EOs and LICOs.

17 *See e.g.*, Andy Greenberg, *The Shadow Brokers Mess is What Happens when the NSA Hoards Zero-Days*, WIRED (17 August 2016), available at goo.gl/zUdceh.

18 *See e.g.*, Lorenzo Franceschi-Bicchierai, *The secret-spilling organization launches a new series where it will release the source code of alleged CIA tools from the Vault 7 series*, MOTHERBOARD (9 November 2017), available at goo.gl/5C8eyN.

19 *See e.g.*, Joseph Cox, *The FBI Hacked over 8,000 Computers in 120 Countries Based on One Warrant*, MOTHERBOARD (22 November 2016), available at goo.gl/wWRtm2.

20 *See e.g.* John S. Davis *et al., Stateless Attribution: International Accountability in Cyberspace*, RAND CORPORATION (2017), available at https://www.rand.org/pubs/research_reports/RR2081.html; Dieter Fleck, *Individual and State Responsibility for Intelligence Gathering*, 28 MICH. J. INT'L. L. 687 (2007).

21 For potential violations from EOs, *see generally* Chesterman, n. 3. For potential violations from LICOs see Waldon, n. 10, at 1469-1477.

22 MICHAEL N. SCHMITT (ED.), TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (2nd ed., 2017) 3.

TABLE 1: AREAS OF CONVERGENCE BETWEEN EOS AND LICOS AS REFLECTED IN THE HYPOTHETICALS

	Part I: Classic EO	Part II: Classic LICO
Instigator	Cyberia's Intelligence	Private Hackers with Support from Cyberia's Intelligence
Tech Employed	Recording Device and Photography	Malware Capable of Both Copying Data and More Disruptive Functions
Accountability Thwarting Mechanism	Unidentified Showgirl, Clandestine Operation	Untraceable Phishing Emails and Hard-To-Detect Trojan Downloader
Goal of Operation	Information on Top Secret DOD R&D Projects	Information on DOD R&D Projects, Economic Disruption and Losses
Potential International Law Violations	Sovereignty, Non-Intervention, Diplomatic Law, Privileges and Immunities, Property Rights, Privacy Rights	Sovereignty, Non-Intervention, Privileges and Immunities, Property Rights, Privacy Rights

3. INFORMATION WARFARE: COMMUNICATING VESSELS WITHIN A UNIFIED SYSTEM

Dr Martin Libicki of the RAND Corporation gave one of the keynote addresses in the 8th International Conference on Cyber Conflict. In his remarks, he made the claim that the old 1990s DoD catch-phrase "Information Warfare" (IW) was making a comeback.[23] IW as a unified theory suggests that "competition over information would be the high ground of warfare,"[24] and that such competitions would involve "the protection, manipulation, degradation and denial of information".[25] It employs the following litmus test: "If information is used to perpetrate an act that was done to influence another to take or not take actions beneficial to the attacker then it can be considered IW".[26] Due to this broad test, different scholars at different times have introduced different elements that form part of IW. Libicki, for example, in his 1995 short monograph *What Is Information Warfare*, introduced it as a heptagon of methods of varying maturity, encompassing:

23 Martin C. Libicki, *The Convergence of Information Warfare*, STRATEGIC STUD. Q. 49, 50 (2017) ("given today's circumstances, in contrast to those that existed when information warfare was first mooted, the various elements of IW should now increasingly be considered elements of a larger whole rather than separate specialties that individually support kinetic military operations").

24 *Id*., at 49.

25 MARTIN C. LIBICKI, WHAT IS INFORMATION WARFARE? X (1995).

26 A. JONES AND G. KOVACICH, GLOBAL INFORMATION WARFARE: THE NEW DIGITAL BATTLEFIELD 5 (2nd ed., 2016).

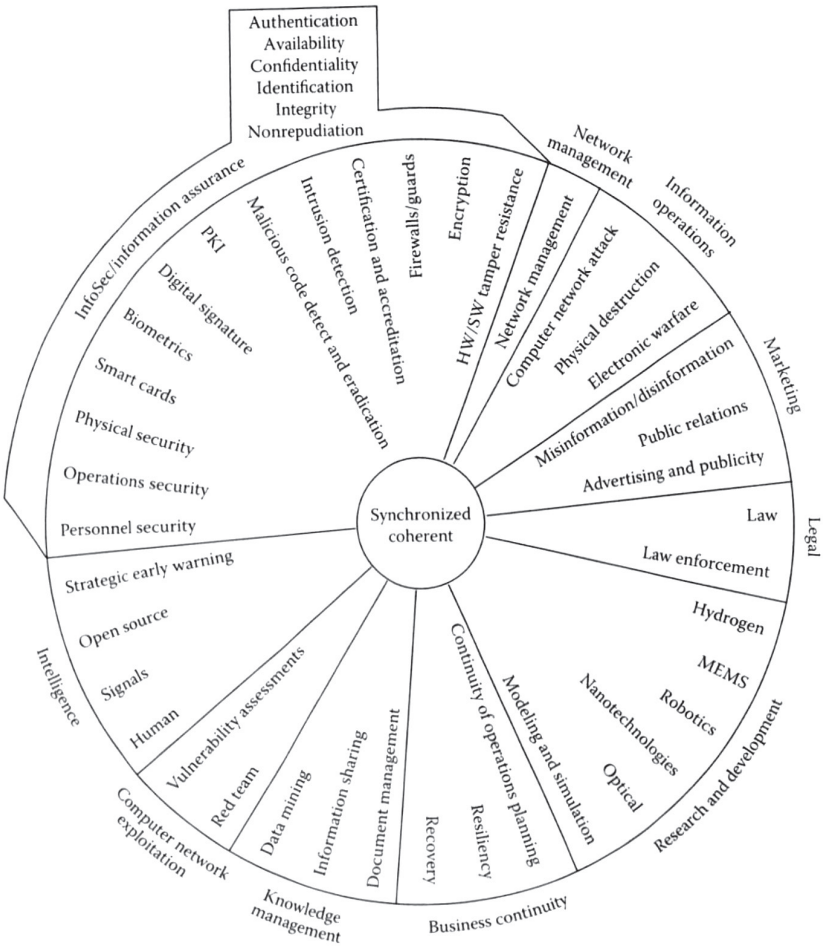
> "(i) command-and-control warfare (which strikes against the enemy's head and neck); (ii) intelligence based warfare (which consists of design, protection, and denial of systems that seek sufficient knowledge to dominate the battle space); (iii) electronic warfare (radio-electronic or cryptographic techniques); (iv) psychological warfare (in which information is used to change the minds of friends, neutrals, and foes); (v) 'hacker' warfare (in which computer systems are attacked); (vi) economic information warfare (blocking information or channelling it to pursue economic dominance); and (vii) cyberwarfare (a grab bag of futuristic scenarios)".[27]

Jones and Kovacich go even further, arguing that IW covers a whole spectrum of elements including, *inter alia*: lawfare, business continuity, knowledge management, information security, computer network exploitation, and intelligence.[28]

27 Libicki, n. 25, at X. Note that today Libicki seems to take a far more condensed approach to the elements encompassing IW, suggesting it covers ISR operations (intelligence, surveillance, and reconnaissance), electronic warfare (EW0, psychological operations (PSYOP), and Cyber Operations. *See* Libicki, n. 23, at 49. Directive 3600.1 of the US DoD similarly adopted this multi-dimensional approach in defining IW's core and supporting capabilities. The original directive was adopted in 1996 but has since been amended twice in 2006 and 2013. In its latest iteration it defines "Information Operations" as "the integrated employment, during military operations, of information-related capabilities (IRC) in concert with other lines of operations to influence, disrupt, corrupt, or usurp the decision making of adversaries and potential adversaries while protecting our own" (DODD O-3600.01, Information Operations (IO) 12 (2 May 2013), available at goo.gl/wJJX6T). The directive proceeds to note that IRCs constitute "tools, techniques, or activities" employed within a dimension of the information environment. These can include, but are not limited to, "a variety of technical and non-technical activities that intersect the traditional areas of electronic warfare, cyberspace operations, military information support operations (MISO), military deception (MILDEC), influence activities, operations security (OPSEC), and intelligence." *Id*., at 1.

28 *See* Jones and Kovacich, n. 26, at 6.

FIGURE 2: JONES AND KOVACICH'S ELEMENTS OF INFORMATION WARFARE



Regardless of which model of IW you adopt, all seem to include both certain EOs and LICOs as components of the broader theater of informational conflict. Libicki argues that the recent convergence of the IW's various elements, and the theory's broader resurgence as a unified doctrine, can be explained by three emerging circumstances:

> "First, the various elements can use many of the same techniques, starting with the subversion of computers, systems, and networks, to allow them to work. Second, as a partial result of the first circumstance, the strategic aspects of these elements are converging. This makes it more likely that in circumstances where

> one element of IW can be used, other elements can also be used. Hence, they can be used together. Third, as a partial result of the second circumstance, countries – notably Russia, but, to a lesser extent, North Korea, Iran and China – are starting to combine IW elements, with each element used as part of a broader whole."[29]

I highlight the discourse on IW because I feel it is important that we place the unique dialectical relationship between EOs and LICOs within a broader informational enviornment. These are two communicating vessels which form part of an even larger machine and the operating logic of that machine, as laid down in the above quote by Libicki, helps further explain the special relationship of EOs and LICOs. Assistant Secretary of Defense Eric Rosenbach once referred to cyber operations as filling the gap between diplomacy and economic sanctions on the one hand, and military action on the other. He called this gap, "the space between" and claimed that cyber operations within this space assist policy-makers in achieving their national interest.[30] The imagery of the space between is useful, but unlike Rosenbach's depiction, it encompasses much more than just cyber operations. A far larger spectrum of informational action, both cyber and non-cyber, occupies this "space between", with intelligence gathering and covert action constituting a significant and historical component. Any attempt at regulating some aspects of this space, in isolation from others, would be ill-fated.

4. THE COMPARTMENTALIZATION APPROACH AND THE TALLINN MANUAL 2.0

Against this backdrop, I want to begin portraying what was attempted in the *Tallinn Manual 2.0*. Rule 32 on "peacetime cyber espionage" is located in Section 5 of the *Manual*, which covers those cyber operations that the Group of Experts (GoE) deemed to be "not *per se* regulated by international law". According to the GoE, customary international law "does not prohibit espionage *per se*",[31] and therefore

29 *See* Libicki, n. 23, at 50.

30 For further reading see Thomas E. Ricks, *The Future of War: Cyber is Expanding the Clausewitzian Spectrum of Conflict*, FOREIGN POLICY (13 November 2014), available at goo.gl/1Nrsmi.

31 Note that the Experts rely on a single source to make this claim, basing themselves on the Office of General Counsel, Department of Defense Law of War Manual. However, paragraph 16.3.2, to which they cite, makes no reference to a lack of customary regulation of espionage under international law, quite the opposite is speaks clearly of "long-standing and well-established considerations" and "long-standing international norms" which govern this practice. *See* DEPARTMENT OF DEFENSE, LAW OF WAR MANUAL 990 (2016) ("international law and long-standing international norms are applicable to State behavior in cyberspace, and the question of the legality of peacetime intelligence and counterintelligence activities must be considered on a case-by-case basis. Generally, to the extent that cyber operations resemble traditional intelligence and counter-intelligence activities, such as unauthorized intrusions into computer networks solely to acquire information, then such cyber operations would likely be treated similarly under international law. The United States conducts such activities via cyberspace, and such operations are governed by long-standing and well-established considerations, including the possibility that those operations could be interpreted as a hostile act.")

determinations of lawfulness should be made on a case-by-case basis taking into account the particular methods employed in the conduct of the specific EO.[32] This allowed the GoE to avoid the need to address the hot potato of comprehensively explaining the law and practice of government espionage. What is more, it furthered the GoE's desire to compartmentalize spying, in its traditional sense, from the more specific cyber espionage and LICOs which regulation the *Manual* sought to elucidate. But as Chesterman has taught us, claiming that espionage is not *per se* regulated under international law is nothing more than a straw man: "Intelligence is less a lacuna in the legal order than it is the elephant in the room".[33] Well, the elephant was alive and well during the *Tallinn Manual* plenary sessions. It swayed its trunk and stomped its feet; but was nonetheless ignored.

Tossing to the side the question of the lawfulness of peacetime intelligence gathering, the GoE dodged the need to speak in higher granularity as to the conduct of interstate spying. Instead, the way was paved for the experts to engage in more general and casuistic reasoning. Throughout their commentary, the experts extract and extend legal rules from a series of tailored hypothetical scenarios, of their own design, which they then analyse in isolation from one another and in accordance with predominantly treaty norms. This "divide-and-conquer" approach is far from harmonious and results in a series of fragmented statements made throughout the *Manual*, each with varying degrees of consensus behind it. Every one of these statements can be compared to liquid being added to one of the vessels. Due to the communicative nature of cyber law and espionage law, as discussed above, any regulation of cyber espionage put forward by the experts – that is to say any regulation of the EOs-LICOs overlap area in our original Venn diagram – automatically sends equilibrium-adjusting tidal waves across the entire system. The experts did not acknowledge these tidal waves, nor did they address the impractical legal realities that they would inevitably create. Let us take up only two examples within the limits of this paper.

The GoE took a territorially protectionist approach to sovereignty violations. According to them:

> "[I]n the cyber context [...] it is a violation of territorial sovereignty for an organ of a State, or others whose conduct may be attributed to the State, to conduct cyber operations while physically present on another State's territory against that State or entities or persons located there."[34]

32 *Tallinn Manual 2.0*, n. 22, at 169-170 ("while the International Group of Experts agreed that there is no prohibition of espionage per se, they likewise concurred that cyber espionage may be conducted in a manner that violates international law due to the fact that certain methods employed to conduct cyber espionage are unlawful").

33 Chesterman, n. 3, at 1072.

34 *Tallinn Manual 2.0*, n. 22, at 19. This rule is then extended to the territorial sea (Rule 48) and the territorial airspace (Rule 55). The GoE is most explicit in the context of the physical tapping of submarine communication cables for the purpose of collecting data. The GoE agreed that "doing so in the territorial or archipelagic waters of another State constitutes a violation of that State's sovereignty". *Id*., at 257.

The GoE provide the example of an agent of one State who uses a USB flash drive to introduce malware into cyber infrastructures in another State and claim that this would result in a sovereignty violation.[35] The caveats provided ("in the cyber context", "cyber operations", etc.) are an attempt at compartmentalization, and have little meaning. If spies cannot clandestinely use a USB flash drive in the territory of a foreign country without it resulting in a sovereignty violation, it follows that they cannot also take photographs, handle HUMINT sources, or steal physical documents in that territory. Especially not in an age where all of these activities *de facto* require some form of cyber enabling. Going down this rabbit-hole, under basic rules of syllogistic logic, if every act of territorial spying results in a sovereignty violation, and every sovereignty violation is a violation of international law,[36] then territorial spying violates international law. Lo and behold, the same experts that concluded that espionage was not "prohibited *per se*", have just *per se* prohibited espionage in its most elementary form.[37] Their approach would seem to suggest that the only lawful way to conduct espionage in the 21st century is either by remote means,[38] or with consent (from the targeted State) or authorization (from the UN Security Council).

A second example comes in the form of the applicability of diplomatic and consular law to cyber espionage. The GoE argues that if a sending State launches spyware from within its diplomatic mission against the cyber infrastructures of another State that would constitute "an abuse of the diplomatic function and therefore an internationally wrongful act."[39] Similarly, if the receiving State or third States intercepted the electronic communications of diplomatic missions and consular posts, they would be violating "the confidentiality of diplomatic and consular communications",

35 *Ibid*. Note that the GoE later backtrack this definitive statement, arguing that they could not agree "on the lawfulness of close access cyber espionage operations, such as the insertion of USB flash drive into a computer located on one State's territory by an individual acting under the direction or control of another State". *Id*., at 171.

36 AJIL Unbound has recently held an online symposium titled "sovereignty, cyberspace, and Tallinn Manual 2.0" which focused on whether sovereignty constitutes a stand-alone binding international legal norm that may be violated. In this debate, I second the view put forward by Phil Spector that there is ample evidence to assert that sovereignty is in fact a binding rule. *See* Phil Spector, *In Defense of Sovereignty, In The Wake of Tallinn 2.0*, 111 AJIL UNBOUND 219 (2017).

37 Not only that, but the experts also claim that certain LICOs employed to enable spying operations, e.g. using cyber intrusions to 'herd' the target's communications to a platform more susceptible to surveillance, might itself trigger separate grounds for sovereignty violations. *Tallinn Manual 2.0*, n. 22, at 172.

38 *Id*., at 19 ("the mere interception of wireless signals from outside the target state's territory does not constitute a violation of that State's sovereignty"). Though even on the point of remote surveillance, there was those experts who argued that a severity test should be employed and that if the consequences suffered from the remote surveillance were so severe, they might too result in a sovereignty violation (*Id*., at 171). Put differently, for certain members of the GoE even spying from outer space, the high seas, or international airspace, might violate sovereignty if they reach a certain degree of severity. This echoes to me the Soviet concept of "danger theory" pushed, and rejected, in the 1960s following the U2 Spy Plane incident. The crux of the Soviet position was that sovereignty might be violated without incursions into national territory, so long as certain national rights were endangered due to a particular surveillance practice. For further reading *see* Joseph R. Soraghan, Reconnaissance Satellites: Legal Characterization and Possible Utilization for Peacekeeping, 13(3) McGill L. J. 458, 471-472 (1967) (quoting the work of Ronald Christensen, he notes that Soviet Russia regarded "her sovereignty rights as going beyond her territorial borders, ceasing, it seems, not even at the borders of another state, and, perhaps pervading the entire universe. No one anywhere, she says, has the right to endanger the Soviet Union").

39 *Id*., at 211-212, 229.

which is central to their function, and therefore will also result in an internationally wrongful act.[40] Once again, note that the repeated references to cyber technologies are inconsequential. The GoE, in essence, is banning espionage from within or against diplomatic missions, regardless of the method employed. If you cannot do it with a malware, there is nothing to justify doing it with your bare hands. The fact that "diplomacy and intelligence gathering have always gone hand in hand,"[41] and that the practice of spying from and on diplomatic missions is as historical as it is commonplace,[42] was not even mentioned in *Tallinn Manual 2.0*, let alone addressed or resolved. Consider the following three reported allegations from the past two decades: (1) In the lead-up to the UN Security Council vote authorizing the use of force against Iraq in 2003, the US and the UK spied on every single delegation to the Security Council;[43] (2) During the G20 talks in Toronto in 2010, the US and Canada spied on large numbers of heads of states and other diplomats in attendance;[44] (3) Between 2012-2017 Chinese agencies used backdoors into computer networks at the African Union Headquarters (networks which they had paid for and installed as a gift) in order to spy on the various delegations.[45] If one wanted to apply *Tallinn Manual 2.0* rules to these three operations, one would have to conclude that all of them violated international law. The same experts who sought to isolate intelligence gathering – to not *per se* address its lawfulness – ended up banning some of the most basic methods through which it is acquired and thereby the practice as a whole. Attempting to only regulate LICOs resulted in tidal waves that inadequately constrained EOs.

In attempting to cage the espionage elephant (by limiting their analysis to specific and self-selected cases of cyber espionage), the GoE found themselves engaging in textual treaty derivation which regurgitated the myth system while ignoring the operational code.[46] The experts did not appreciate fully what CIA analyst James Jesus Angleton

40 *Id*., at 221.

41 Chesterman, n. 3, at 1072.

42 Craig Forcese, *Spies Without Borders: International Law and Intelligence Collection*, 5 J. NAT'L SEC'Y L. & POL'Y 179, 197 (2011); Ashley Deeks, *An International Legal Framework for Surveillance*, 55(2) VIRG. J. INT'L L. 291, 313 (2015) (citing Antonin Scalia who at the time of working for the DOJ OLC drafted a memorandum which concluded that "the practice of spying on foreign missions was so widespread that the "inviolability" provision of the VCDR should not be read to prohibit such activities).

43 *See e.g*. Martin Bright and Peter Beaumont, *Britain spied on UN allies over war vote*, THE GUARDIAN (7 February 2004), available at http://goo.gl/fXhd8U.

44 *See e.g*. Paul Owen, *Canada 'allowed NSA to spy on G8 and G20 summits*', THE GUARDIAN (28 November 2013), available at http://goo.gl/HJB6mD.

45 *See e.g*. Reuters, *China rejects claim it bugged headquarters it built for African Union*, THE GUARDIAN (29 January 2018), available at http://goo.gl/i5yt2g.

46 As Professor W. Michael Reisman noted "in law things are not always what they seem," and one needs to be particularly mindful of the existence of "two 'relevant' normative systems: one which is supposed to apply and which continues to enjoy lip service among elites and one which is actually applied". Reisman describes the tension between the myth and the code as a "dynamic process" and a "symbiotic relationship". Acknowledging that the international law governing EOs and LICOs does not exist solely in the myth or solely in the code, but rather in the space between the two, would have benefited the quality of *Tallinn Manual 2.0*'s overall analysis. For further reading see W. Michael Reisman, *On the Causes of Uncertainty and Volatility in International Law, in* THE SHIFTING ALLOCATION OF AUTHORITY IN INTERNATIONAL LAW: CONSIDERING SOVEREIGNTY, SUPREMACY AND SUBSIDIARITY 44-45 (Tomer Broude & Yuval Shany eds., 2008).

described as the "wilderness of mirrors" that is part and parcel of spycraft. Explaining the legal intricacies of espionage requires one to embrace the notion that all law inevitably involves certain forms of *lex simulata* and *lex imperfecta*. Merely citing the law-in-the-books, while avoiding the-law-in-action, pays a disservice to the experts' overall courageous goal of legal elucidation and codification. The *Tallinn Manual 2.0* could have (and should have) engaged in a far more deliberate, nuanced, and comprehensive investigation into the international law of intelligence, which would have inspired the development of more harmonious and sensible cyber norms with practicability for both scholars and practitioners.

5. PROPOSING AN ALTERNATIVE HARMONIOUS ACCOUNT

The *Tallinn Manual 2.0* could have started by acknowledging that customary international law recognizes a sovereign nation's right to spy – because it does. Our international legal order, and within it more specifically our "contemporary global security system", is dependent upon a "reliable and unremitting flow of intelligence to the pinnacle elites".[47] A plethora of legal sources, enshrined in both treaty and custom, effectively recognize the existence of a derivative liberty right of States to peacetime intelligence gathering. These sources include:

1. The right of States to survival, recognized by the ICJ in the Nuclear Weapons advisory opinion[48] (and the related collective right of self-determination of peoples);
2. The laws on the use of force (and their recognition of both a customary and a Charter-based right for individual and collective self-defense);
3. Collective monitoring obligations under UN and Treaty Law (as encompassed for example in the fields of disarmament and counter-proliferation law, counter-terrorism law, sanctions regimes, environmental law, disaster relief, and the fight against illicit trafficking);
4. International human rights law (and the obligation of States to respect and ensure the right to life, liberty, and security of all persons subject to their jurisdiction, as well as the discretion of States to derogate from certain rights in times of emergency as well as balance them off in the name of protecting national security interests);

47 Myres S. McDougal, Harold D. Lasswell & W. Michael Reisman, *The Intelligence Function and World Public Order*, 46 TEMP. L.Q. 365, 434 (1973).

48 Use of Nuclear Weapons, Advisory Opinion, I.C.J. Reports 1996, 226, 263 (8 July 1996) ("The Court cannot lose sight of the fundamental right of every State to survival and thus its right to resort to self-defense in accordance with Article 51 of the Charter, when its survival is at stake").

5. International humanitarian law (and the obligation of States to develop "effective intelligence gathering systems", already in peacetime and in preparation for armed conflict, so to be able "to collect and evaluate information concerning potential targets" during the war);[49] and
6. International Accountability Regimes (certain obligations and requirements derived from both international criminal law and the frameworks governing State responsibility for internationally wrongful acts).

Within the scope of this paper, I cannot delve into a comprehensive analysis of each of these sources. Instead, let me focus on the right of self-defense, as a single example. Dating back to the Caroline incident of 1837, the right of a State to engage in pre-emptive self-defense in order to avert an attack that is "instant, overwhelming, leaving no choice of means, and moment of deliberation"[50] has been extensively analysed.[51] Even those who still maintain, based on the wording of UN Charter Article 51, that a right of self-defense applies only "if an armed attack occurs," cannot ignore diverse and robust subsequent practice by States.[52] The 2004 High-level Panel on Threats, Challenges, and Change established by the UN Secretary-General thus recognized that "a threatened State, according to long established international law, can take military action as long as the threatened attack is imminent, no other means would deflect it, and the action is proportionate."[53] Regardless of what interpretation of "imminence" one adopts, from a classically restrictive "Pearl Harbor"-type position to a highly permissive "Bush doctrine"-type position,[54] both ends of the spectrum, and everything in between, will embrace a State's derivative right to engage in peacetime intelligence gathering. For how else will a State know when a threat reaches whatever level of imminence is deemed sufficient to justify military action? If a State is entitled to retaliate against imminent threats to its survival, by definition it must be allowed to engage in peacetime espionage to gather the information necessary to reach that very conclusion.

Even were we to adopt the formalistic and anachronistic approach that only Article 51 holds (and therefore that a State may only react to an imminent threat by seeking

49 *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia*, ICTY, ¶29 (June 2, 2000), available at http://goo.gl/btGZ6y.

50 *See generally*, R.Y. Jennings, The Caroline and McLeod Cases, 32 AM. J. INT'L L. 82 (1938).

51 For a summary of the literature, see Christopher Greenwood, *Self-Defence*, MAX PLANCK ENCYCLOPEDIA PUB. INT'L. L. (Apr. 2011), available at http://goo.gl/zwaErV. For a more recent review of the literature, see Monica Hakimi, *North Korea and the Law on Anticipatory Self-Defense*, EJIL: TALK! (Mar. 28, 2017), available at http://goo.gl/4XPZeb.

52 W. Michael Reisman & Andrea Armstrong, The Past and Future of the Claim of Preemptive Self-Defense, 100 AM. J. INT'L. L. 525, 526 (2006) (noting that anticipatory self-defense was not, in their view, "in the contemplation of drafters of the Charter, though claimed by many to have been grafted thereon by subsequent practice," followed by a showing of such practice through case studies).

53 *Secretary General's High-Level Panel on Threats, Challenges, and Change, A More Secure World: Our Shared Responsibility*, UNITED NATIONS 63 (2004), available at http://goo.gl/JxTQKb.

54 For more moderate interpretations, see Daniel Bethlehem, *Principles Relevant to the Scope of a State's Right of Self-Defense Against an Imminent or Actual Armed Attack by Nonstate Actors*, 106 AM. J. INT'L L. 769 (2012); Jeremy Wright, The Modern Law of Self-Defense, EJIL: TALK! (Jan. 11, 2017), available at http://goo.gl/1QCaHH.

Security Council authorization) there would still be a derivative right for States to engage in peacetime intelligence gathering. For how else will a delegation be able to prove to the Security Council that a threat is mounting, so to convince its members to vote in favour of an authorization of the use of force? To the extent that the United Nations does not have its own intelligence capacities, the Security Council must rely on Member States in order to fulfil its mandate of maintaining peace and security. Note in this regard that the UN Security Council has in fact acknowledged the function that Member States' intelligence plays in its ability to exercise this mandate. Most recently it adopted this view in Resolution 2396, concerning threats to international peace and security caused by terrorist acts. Acting under Chapter VII the Council not only called on Member States to "intensify and accelerate" their peacetime intelligence collection efforts, it went on to suggest exactly what measures they should employ. The Council decided that Member States "shall develop and implement systems to collect biometric data, which could include fingerprints, photographs, facial recognition, and other relevant identifying biometric data". Other measures ordered by the Council were certain capabilities for the collection, processing, and analysis of passenger name record (PNR) and advance passenger information (API) data, the development and implementation of watch lists and databases on suspected terrorists, and increased cooperation with information communication technology companies in gathering a myriad of digital records and their later sharing through bilateral and multilateral arrangements.[55]

By recognizing that a right to spy exists as a matter of customary international law, the international community inexplicitly created a caveat to the myth system enshrined, *inter alia*, in Articles 2(1), 2(4), and 2(7) of the UN Charter, as well as in certain international legal regimes (such as the ones governing diplomatic and consular relations). Countries are willing to accept as tolerable certain assaults on their territorial sovereignty, political independence, their jurisdiction to determine their domestic affairs, and immunities and privileges, in the name of maintaining the important functions that intelligence plays in our public world order.[56] So long as the surveillance serves the *raison d'être* of our international system, the fundamental goals of all law – "the minimization of violence, the maintenance of minimum order, and as approximate an achievement of the politics of human dignity as each situation allows"[57] – the practice will be stomached even by those who have been discontentedly

55 UN Security Council Resolution 2396 concerning Threats to International Peace and Security Caused by Terrorist Acts, UN Doc. S/RES/2396 (21 December 2017).

56 For more on this function see Myres S. McDougal, Harold D. Lasswell & W. Michael Reisman, *The Intelligence Function and World Public Order*, 46 TEMP. L.Q. 365 (1973).

57 W. Michael Reisman, *Editorial Comment: Assessing Claims to Revise the Laws of War*, 97 AM. J. INT'L L. 82, 83 (2003).

subjected to it.[58] Note that this position was suggested, though ultimately not adopted, by a minority of the experts in *Tallinn Manual 2.0*:

> "A few of the experts were of the view that the extensive State practice of conducting espionage on the target State's territory has created an exception to the generally accepted premise that non-consensual activities attributable to a State while physically present on another's territory violate sovereignty. They emphasized, however, that this exception is narrow and limited solely to acts of espionage"[59]

Of course, acknowledging the right to spy would only be the first step in articulating the broader law on espionage. A fundamental source of international law mostly ignored by the GoE is that of general principles of law, which stand on the same footing as treaties and custom.[60] One of the typical uses of general principles is as "standard clarifiers", serving the purpose of defining "the depth and contours of broad or amorphous legal provisions" where international conventions and customs offer little organizational help.[61] One such general principle is the principle of "Abuse of Rights". Sir Hersch Lauterpacht recognized that "there is no legal right, however well established, which could not, in some circumstances, be refused recognition on the ground that it has been abused".[62] Applying the Abuse of Rights doctrine to our newly articulated Right to Spy creates the basis for the *Jus Ad Explorationem* (the law governing the launching of EOs). When spying is launched to achieve goals other than the ones for which it was originally intended, the particular operation will no longer be tolerable. Spying may only serve the national security interests of a State or the

58 Note that stomaching it from an international law point of view is different from domestically prohibiting spying and working extensively to prevent it. This is the essence of the "liberty right" to spy, as a weaker right, that does not create an obligation on third parties to condone or facilitate it. This GoE acknowledged the practice of State domestic criminalization of espionage, see *Tallinn Manual 2.0*, n. 22, at 174 ("States are entitled to, and have, enacted domestic legislation that criminalises cyber espionage carried out against them").

59 *Id*., at 19. See also at 171 ("A few of the experts took the view that [territorial cyber espionage] would not be unlawful, suggesting that acts of espionage represent an exception to the prohibitions of violations of sovereignty and intervention").

60 In the Introduction to *Tallinn Manual 2.0*, Professor Schmitt addresses which "rules and commentary" guided the GoE. It is quite visible from his description that the experts were solely interested in articulating treaty and customary international rules. The third source of international law, that of general principles, is not once mentioned by the project director in that section and is rarely brought up as such throughout the *Manual*. *Id.*, at 3-5.

61 Alain Pellet, *Article 38*, in THE STATUTE OF THE INTERNATIONAL COURT OF JUSTICE: A COMMENTARY 731, 850 (Andreas Zimmermann et al. eds., 2nd ed., 2012) (noting that the ICJ "will usually only resort to [General Principles of Law] in order to fill a gap in the treaty or customary rules available to settle a particular dispute"); CHARLES T. KOTUBY JR. AND LUKE A. SOBOTA, GENERAL PRINCIPLES OF LAW AND INTERNATIONAL DUE PROCESS: PRINCIPLES AND NORMS APPLICABLE IN TRANSNATIONAL DISPUTES 31-32 (2017) ()the authors cite the example of the ICSID tribunal using general principles to determine the precise content of the "fair and equitable treatment" standard, taking this interpretive approach due to the fact that "treaties and international conventions. … are not of great help to this end").

62 SIR HERSCH LAUTERPACHT, THE DEVELOPMENT OF INTERNATIONAL LAW BY THE INTERNATIONAL COURT 164 (1958).

broader interests of maintaining peace and security for the international community in general.[63] Thus, for example, if spying is done for the purpose of advancing the personal economic interests of a particular leader or those of specific corporations or industries,[64] or if it is conducted to facilitate a dictatorship or to commission an internationally wrongful act,[65] such spying operations are used "for an end which is different from which the right has been created",[66] and would therefore constitute an abuse of that very right.

Moreover, even in cases where the operation does serve a lawful purpose, but in its choice of means or targets (the *Jus In Exploratione*) the State adopts certain measures which are either customarily prohibited (*e.g.* torture and other cruel, inhuman or degrading treatment; or arbitrary interference with the customary human rights to privacy or freedom of expression), or which go beyond "unexpressed but generally accepted norms and expectations",[67] the operation might nonetheless be deemed unlawful. Again, general principles of law such as good faith, proportionality, rule of law, effectiveness, fairness, and comity,[68] might serve as useful tools in both interpreting existing treaty and customary norms (*e.g.* determining what constitutes as torture, or other cruel inhuman or degrading treatment; determining what violates the international human rights to privacy and freedom of expression) and clarify standards where the law has not yet caught up with the development of new surveillance and

63 *See* Asaf Lubin, *The Dragon-Kings Restraint: Proposing a Compromise for the EEZ Surveillance Conundrum*, 57 WASHBURN L. J. 1, 56 (2018).

64 Note that this idea was entertained to some degree by certain members of the GoE. *See Tallinn Manual 2.0*, n. 22, at 169, fn 386 (citing the 2015 US-Chinese commitment not to support cyber-enabled theft of intellectual property, and to a similar commitment taken by the G20 leaders that same year, the GoE cautioned that States may have committed themselves *inter se* to certain restrictions on industrial espionage. Nonetheless the GoE stopped short of determining that such practice was unlawful).

65 This resembles the position of the GoE that cyber espionage operations may be unlawful if they "constitute an integral and indispensable component of an operation that violates international law." *See Id*., at 171-172.

66 Alexandre Kiss, *Abuse of Rights*, MAX PLANCK ENCYCLOPEDIA OF INTERNATIONAL LAW, para. 5 (2006).

67 Roger D. Scott, *Territorially Intrusive Intelligence Collection and International Law*, 46 A.F. L. Rev. 217, 226 (1999) ("as long as unexpressed but generally accepted norms and expectations associated with espionage are observed, international law tolerates the collection of intelligence").

68 None of these general principles were sufficiently addressed in *Tallinn Manual 2.0*. Quite the opposite, the GoE even challenged the customary nature of proportionality as a binding legal requirement (*Tallinn Manual 2.0*, n. 22, at 204-205). For an analysis of proportionality as a general principle of international law *see* Kotuby and Sobota, n. 61, at 114-119. Similarly, an array of human rights standards, common to surveillance jurisprudence, and their applicability to both LICOs and EOs were hardly addressed by the authors. These include *inter alia* the principles of legality, necessity, proportionality, ex ante authorization, minimization procedures and safeguards from abuse, *ex post* review, independent oversight, non-discrimination, notification requirements, and access to remedy and justice.

cyber technologies.[69] Of course, making these determinations requires the use of contextual and consequential methods of inquiry.[70]

Determining the lawfulness of a particular LICO, including specifically cyber espionage operations, is not for the fainthearted. One should be willing to engage the *Jus Ad Explorationem* and the *Jus In Exploratione*, in light of the function that intelligence plays in our public world order, and in view of a contextual- and consequential-based analysis. It is therefore the reality that in some instances foreign agents introducing USB flash drives filled with spyware into national cyber infrastructures might indeed violate the international law of espionage, whereas in other instances they might not. We consider the intrusion on sovereignty or on diplomatic immunities only as factors in a far more layered legal analysis. This type of nuanced application will be relevant to all of the other hypotheticals introduced in the *Manual*: from certain cyber intrusions that 'herd' a target's communications to a platform more susceptible to surveillance, through tapping underwater submarine cables in the territorial sea, to spying on diplomats at the United Nations. Some of these might meet the above standards and criteria and would therefore be tolerated and stomached by the international community; others might not and would therefore be condemned, potentially even triggering State obligations for reparation. Far from rushing to provide rigid rules, *Tallinn Manual 2.0* should have recognized the symbiosis that exists across the informational domain, as manifested in the communicative nature of cyber and espionage law and should have thus been more modest in its approach. Instead of a rulebook, the GoE should have provided government lawyers with a map and a compass.

6. CONCLUSION

Dr Seuss taught us that "sometimes the questions are complicated and the answers are simple". In the area of cyber and espionage law, however, both the questions and the answers are complicated. This places a burden of humility on rule prescribers and rule appliers. In this paper, I have tried to highlight how, in our liberal rush to demonstrably regulate the cyber domain, a pursuit that we undertake for all the right reasons and with all the right intentions, we might end up leaving scorched earth.

69 M. Cherif Bassiouni, *A Functional Approach to "General Principles of International Law"*, 11 MICH. J. INT'L. L. 768, 777 (1989-1990) (where he suggested that general principles prevent "the static application of anarchic norms and procedures to what is admittedly an evolving legal process designed to frame or regulate the dynamic exigencies and needs of a community of nations with changing interests and mutable goals and objectives. To state that international law has faced and is likely to face increasing new challenges, if for no other reason than to meet the fast-growing and changing technological advances, is a truism. Thus the demands on international law must be accommodated through an expanded usage of 'General Principles'").

70 Reisman and Baker take this analysis a step further by applying a similar methodology (though at a higher level of abstraction) to the regulation of covert action. *See* W. MICHAEL REISMAN AND JANES E. BAKER, REGULATING COVERT ACTION: PRACTICES, CONTEXTS AND POLICIES OF COVERT COERCION ABROAD IN INTERNATIONAL AND AMERICAN LAW (1992).

When policy-makers are provided with sufficiently accurate information as to the levels and types of threats posed by their adversaries, their intentions, and capabilities, they are more likely to calibrate their responses properly, and are less likely to rely on force as a means for guarding against startling attacks or strategic surprises. Intelligence gathering, in this context, serves a stability-enhancing function in public world order, by increasing the potential for pacific settlement of disputes and reducing the chances for violence. As George Washington said: "To be prepared for war is one of the most effectual means of preserving peace".[71] The communicative nature of cyber law and espionage law entails that we need to take a degree of caution so that we do not regulate the former to a point where we can no longer benefit from the positive functions served by the latter.

A legal regime that tries to address LICOs without being mindful and cognizant of the tidal waves that such regulations will inevitably create for EOs is one that is doomed to be rejected by States. Far more troubling, however, is the fact that such a legal regime will not even serve our initial goals of enhancing the rule of law, stability, and the peaceful resolution of conflicts. The former President of the Republic of Estonia, Toomas Hendrik Ilves, opens *Tallinn Manual 2.0* by criticizing those who rely on realpolitik to dismiss international law as mere "window-dressing".[72] I share his criticism, but to adopt a set of rules that only echo the myth system while ignoring the operational code will only give fuel to those who scoff at the power of international law to effectively shape and bound government actions and expectations.

71 President George Washington, First Message to Congress on the State of the Union (Jan. 8, 1790).
72 *See Tallinn Manual 2.0*, n. 22, at xxiii.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Internet Intermediaries and Counter-Terrorism: Between Self-Regulation and Outsourcing Law Enforcement[1]

Krisztina Huszti-Orban
School of Law
University of Minnesota
Minneapolis, United States
khusztio@umn.edu

Abstract: Recent years have seen increasing pressure on Internet intermediaries that provide a platform for and curate third-party content to monitor and police, on behalf of the State, online content generated or disseminated by users. This trend is prominently motivated by the use of ICTs by terrorist groups as a tool for recruitment, financing, and planning operations. States and international organizations have long called for enhanced cooperation between the public and private sectors to aid efforts to counter terrorism and violent extremism. However, as the Special Rapporteur on Freedom of Expression noted in his latest report to the Human Rights Council, 'the intersection of State behaviour and corporate roles in the digital age remains somewhat new for many States'.

Detailed information on the means and modalities of content control exercised by online platforms is scarce. Terms of service and community standards are commonly drafted in terms that do not provide sufficiently clear guidance on the circumstances under which content may be blocked, removed or restricted, or access to a service may be restricted or terminated. Users have limited possibilities to challenge decisions to restrict material or access to a service. Moreover, as private bodies, such platforms are generally subject to limited democratic or independent oversight. At the same time, having private actors such as social media companies increasingly undertake traditionally public interest tasks in the context of Internet governance is likely unavoidable, as public authorities frequently lack the human or technical resources to satisfactorily perform these tasks.

1 This work was supported by the UK's Economic and Social Research Council [grant number ES/M010236/1].

Against this background, this paper aims to examine ways to define the contours of the division of responsibilities in countering terrorism and violent extremism between the public and private spheres. It addresses ways to ensure that Internet intermediaries carry out quasi-enforcement and quasi-adjudicative tasks in a manner compliant with international human rights norms and standards.

Keywords: *terrorism, violent extremism, human rights, Internet intermediaries, freedom of expression*

1. INTRODUCTION: ONLINE PLATFORMS AS GATEKEEPERS OF THIRD-PARTY CONTENT

It is difficult to overstate the role of the Internet intermediaries that provide a platform for and curate online content in facilitating the public's access to seek, receive, and impart information, including discourse on issues of public interest. Individuals' exercise of free speech is increasingly channelled through online platforms, which also enable governments to communicate with their constituencies and similarly facilitate the dissemination of messages by other actors. Many major online platforms (social media portals and search engines being prime examples) function on the basis of business models centred around hosting third-party content. The companies running these platforms regularly claim that the platforms function as mere distribution channels that exercise no or limited editorial intervention over the content published. Some of these sites have extremely high levels of user activity and interactivity,[2] allowing them to reach broad and diverse audiences in a manner that was not feasible before.[3] This, at the same time, makes meaningful real-time monitoring challenging or even impossible and editorial intervention time- and resource-intensive.

Online platforms regulate their use through terms of service and community standards. The private regulatory mechanisms used by these platforms generally represent an efficient alternative to public regulation in the online space. The terms and standards are pre-established and unilaterally imposed on all users who want access to the services offered, providing the platform with quasi-normative power when it comes to user behaviour. This power extends not only to the substantive aspects of use, such

2 It has been reported that every 60 seconds 510,000 comments are posted on Facebook, 293,000 statuses are updated, and 136,000 photos are uploaded. See Zephoria Digital Marketing, 'The Top 20 Valuable Facebook Statistics – Updated January 2018', 8 May 2017, https://zephoria.com/top-15-valuable-facebook-statistics/, accessed 15 January 2018. The daily video content watched on YouTube has reached 1billion hours this year. See YouTube Official Blog, 'You know what's cool? A billion hours' (27 February 2017) https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html, accessed 15 January 2018.

3 See Dave Chaffey, 'Global social media research summary 2017' (Smart Insights, 27 April 2017) http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, accessed 15 January 2018.

as the content that users are authorized to share or access, but also to enforcement-related ones, such as the criteria for decision-making and the technical tools used for the implementation of such decisions. In addition to these quasi-normative and quasi-executive functions, platforms frequently enjoy quasi-adjudicative power by requiring that disputes with users are settled via internal or other alternative dispute resolution or remedy mechanisms.

Such private 'sovereignty' should nonetheless be subject to public scrutiny to avoid arbitrary or abusive use of power. This is particularly important in light of some platforms undertaking functions traditionally catered for by the State. The argument that online platforms have become the digital age equivalent of public squares has been gaining traction in recent years.[4] Due to their reach, use, and level of interactivity, some of these platforms arguably play a public interest role. Studies show that people have increasingly been getting their news from social media.[5] Social media platforms have further been instrumental in disseminating information about political developments at home and abroad, humanitarian crises, and allegations of violations and abuses committed by States and Non-State actors.[6] In some countries or provinces, certain social media platforms are so dominant that to many inhabitants they represent the Internet itself.[7] As such, the information these inhabitants have access to online is restricted to whatever is available on these platforms. As offline information flows in these contexts are frequently restricted, social media platforms may constitute the main source of information, including of public interest information.

4 See Alissa Starzak, 'When the Internet (officially) became the public square' (Cloudfare, 21 June 2017) https://blog.cloudflare.com/internet-became-public-square/, accessed 15 January 2018; Ephrat Livni, 'The US Supreme Court just ruled that using social media is a constitutional right' (Quartz, 19 June 2017) https://qz.com/1009546/the-us-supreme-court-just-decided-access-to-facebook-twitter-or-snapchat-is-fundamental-to-free-speech/, accessed 15 January 2018.

5 See Jordan Crook, '62% of U.S. adults get their news from social media, says report' (TechCrunch, 26 May 2016) https://techcrunch.com/2016/05/26/most-people-get-their-news-from-social-media-says-report/, accessed 15 January 2018; Jane Wakefield, 'Social media "outstrips TV" as news source for young people', (BBC News, 15 June 2016) http://www.bbc.co.uk/news/uk-36528256, accessed 15 January 2018.

6 Christoph Koettl, 'Twitter to the rescue? How social media is transforming human rights monitoring', (Amnesty International USA, 20 February 2013) http://blog.amnestyusa.org/middle-east/twitter-to-the-rescue-how-social-media-is-transforming-human-rights-monitoring/, accessed 15 January 2018; Juliette Garside, 'Rioters' use of social media throws telecoms firms into spotlight' (The Guardian, 21 August 2011) https://www.theguardian.com/business/2011/aug/21/riots-throw-telecoms-firms-social-media-controls-into-spotlight, accessed 15 January 2018; Clay Shirky, 'The Political Power of Social Media: Technology, the public sphere and political change' Foreign Affairs (January/ February 2011) Vol. 90, No.1, 28-41.

7 See Megan Specia and Paul Mozur, 'A war of words puts Facebook at the center of Myanmar's Rohingya crisis' (The New York Times, 27 October 2017) https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html?mtrref=www.google.com, accessed 12 March 2018; Casey Hynes, 'Internet use is on the rise in Myanmar, but better options are needed' (Forbes, 22 September 2017) https://www.forbes.com/sites/chynes/2017/09/22/Internet-use-is-on-the-rise-in-myanmar-but-better-options-are-needed/#1ef96e44448e, accessed 12 March 2018; Corynne McSherry, Jeremy Malcolm, Kit Walsh, 'Zero Rating: What it is and why you should care' (Electronic Frontier Foundation, 18 February 2016) https://www.eff.org/deeplinks/2016/02/zero-rating-what-it-is-why-you-should-care, accessed 12 March 2018.

The full picture needs to be considered in light of technological developments that have provided new means and modalities for controlling the content available online. Online platforms and those who provide and facilitate access to them have considerable power in shaping the information that gets disseminated; that is, they have *de facto* authority when it comes to regulating online content. As offline news consumption continues to decrease, particularly with younger demographics, these actors can exert significant influence over individuals' access to information, freedom of opinion, expression, and association, and over interlinked political and public interest processes.[8] The issue has figured prominently in recent discussions centring around the role of social media in influencing democratic, including electoral, processes.[9]

In addition to these regulatory functions, platforms have increasingly been undertaking policing and law enforcement functions traditionally considered to be State tasks. At times, such roles are delegated through law, as is the case of the German Network Enforcement Act.[10] However, platforms increasingly undertake such functions without their being formally delegated by state authorities, in an attempt to avoid liability or pre-empt State regulation.

This paper aims to examine the division of responsibilities between the public and private sphere in countering terrorism and violent extremism in a context where the 'playground' is privately owned and operated infrastructure, with uneven levels of State regulation. It addresses means and modalities to ensure that Internet intermediaries, with particular focus on social media platforms, carry out quasi-enforcement and quasi-adjudicative tasks in a manner compliant with international human rights norms and standards. The analysis will pay particular attention to relevant developments in European Union (EU) laws and policies and Member State practices.[11]

2. STATE TRENDS TO OUTSOURCE ONLINE (CONTENT) POLICING

Recent years have seen increasing pressure on Internet intermediaries that provide a platform for and curate third-party content to monitor and police, on behalf of the State,

8 Bruce Schneier, Data and Goliath (New York: W.W. Norton & Company, 2015), 114-116.

9 See, for example, Ryan Goodman and Justin Hendrix, ‚Facebook users have the right to know how they were exposed to Russian Propaganda' (Just Security, 23 October 2017) https://www.justsecurity.org/46171/facebook-users-right-to-know-exposed-russian-propaganda/, accessed 12 March 2018; Hannes Grassegger and Mikael Krogerus, 'The data that turned the world upside down' (Motherboard VICE, 27 January 2017) https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win, accessed 12 March 2018.

10 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Netzwerksdurchsetzungsgesetz - NetzDG] (2017).

11 The reasons for choosing to demonstrate related issues by reference to the EU framework are the more detailed nature of EU regulation and its interpretation and also numerous current developments at the EU and Member State level. Many of the concerns raised are, however, valid beyond the EU.

online content that is generated or disseminated by users. This trend is prominently motivated by the use of ICTs and social media, in particular, by terrorist groups as a tool for recruitment, propaganda outreach, fundraising, and planning operations.[12] Discussions on the role and responsibilities of relevant online platforms in preventing and countering terrorism and violent extremism have intensified in the wake of recent attacks perpetrated by individuals linked to or inspired by ISIL.[13] Some policy-makers have expressed dissatisfaction with the efficiency of monitoring terrorist and violent extremist content and have warned platforms about the need to 'do more' if they want to avoid State intervention through binding regulation and sanctions.[14]

For its part, the tech industry has attempted to tackle the problems posed by terrorist or extremist third-party content through coordinated initiatives aimed at bolstering the efficiency of individually taken measures. Coordinated initiatives include the Global Internet Forum to Counter Terrorism,[15] the EU Internet Forum, bringing together EU entities, governments and technology companies,[16] the Code of Conduct on Countering Illegal Hate Speech Online,[17] and the Shared Industry Hash Database,[18] to name a few. Individually, companies have pledged to take further action to counter the use of their platforms for terrorist and other unlawful purposes by employing

12 See Brendan I. Koerner, 'Why ISIS is winning the social media war' (Wired, April 2016) https://www.wired.com/2016/03/isis-winning-social-media-war-heres-beat/, accessed 15 January 2018; David P. Fidler, 'Countering Islamic State exploitation of the Internet' (Council on Foreign Relations, 18 June 2015) https://www.cfr.org/report/countering-islamic-state-exploitation-internet, accessed 15 January 2018.

13 Andrew Sparrow, Alex Hern, 'Internet firms must do more to tackle online extremism, says No 10' (The Guardian, 24 March 2017) http://www.theguardian.com/media/2017/mar/24/internet-firms-must-do-more-to-tackle-online-extremism-no-10, accessed 15 January 2018; Jessica Elgot, 'May and Macron plan joint crackdown on online terror' (The Guardian, 12 June 2017) https://www.theguardian.com/politics/2017/jun/12/may-macron-online-terror-radicalisation, accessed 15 January 2018.

14 Amar Toor, 'France and the UK consider fining social media companies over terrorist content' (The Verge, 13 June 2017) https://www.theverge.com/2017/6/13/15790034/france-uk-social-media-fine-terrorism-may-macron, accessed 15 January 2018; Samuel Gibbs, 'Facebook and YouTube face tough new laws on extremist and explicit video' (The Guardian, 24 May 2017) https://www.theguardian.com/technology/2017/may/24/facebook-youtube-tough-new-laws-extremist-explicit-video-europe, accessed 15 January 2018; Kate McCann, 'Facebook "must pay to police internet" or face fines: UK Parliament' (The Canberra Times, 30 April 2017) http://www.canberratimes.com.au/technology/technology-news/facebook-must-pay-to-police-internet-20170430-gvvz2e.html, accessed 15 January 2018.

15 Microsoft Corporate Blogs, 'Facebook, Microsoft, Twitter and YouTube announce formation of the Global Internet Forum to Counter Terrorism' (26 June 2017) https://blogs.microsoft.com/on-the-issues/2017/06/26/facebook-microsoft-twitter-youtube-announce-formation-global-internet-forum-counter-terrorism/, accessed 15 January 2018.

16 European Commission, 'EU Internet Forum: a major step forward in curbing terrorist content on the internet. Press release' (Brussels, 8 December 2016) http://europa.eu/rapid/press-release_IP-16-4328_en.htm, accessed 15 January 2018.

17 The initiative is from the European Commission, together with Facebook, Microsoft, Twitter and YouTube. The Code of Conduct is available at http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf, accessed 15 January 2018.

18 Google, 'Partnering to help curb the spread of terrorist content online' (5 December 2016) https://www.blog.google/topics/google-europe/partnering-help-curb-spread-terrorist-content-online/, accessed 15 January 2018.

artificial intelligence and 'human expertise' to identify 'extremist and terrorism-related' content.[19]

3. ONLINE PLATFORMS AND COUNTER-TERRORISM

Relevant corporate obligations are included in a variety of laws adopted at the national level, among others those tackling hate speech, cybercrime, counter-terrorism, violent extremism, and intermediary liability. Many jurisdictions also encourage self- and co-regulation.

A. Terrorism and Violent Extremism: Dilemmas of Definition

Despite a plethora of multilateral treaties, Security Council resolutions, and other international and regional instruments addressing terrorism-related issues,[20] an internationally agreed definition of terrorism or an agreed list of terrorism-related offences is lacking. As a result, relevant definitions are to be found in laws and policies adopted at the level of States, causing considerable discrepancies between different domestic frameworks.

Particularly pertinent to our context are preparatory and ancillary offences and, newly, offences criminalizing the advocacy of terrorism, including 'glorification', 'apology', 'praise' or 'justification' of terrorism.[21] United Nations human rights mechanisms and other stakeholders have raised concerns over some definitions lacking precision, stressing the potential negative human rights implications of definitions of terrorism

19 See, for example, Google, 'Four steps we're taking today to fight terrorism online' 18 June (2017) https://www.blog.google/topics/google-europe/four-steps-were-taking-today-fight-online-terror/, accessed 15 January 2018; Monika Bickert, Brian Fishman, 'Hard Questions: How We Counter Terrorism', (15 June 2017) https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/, accessed 15 January 2018; Twitter Inc. 'An update on our efforts to combat violent extremism' (18 August 2016) https://blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html, accessed 15 January 2018.

20 See United Nations Counter-Terrorism Implementation Task Force, International Legal Instruments, https://www.un.org/counterterrorism/ctitf/en/international-legal-instruments, accessed 15 January 2018.

21 The UN Human Rights Committee has stressed that offences such as 'praising', 'glorifying', or 'justifying' terrorism must be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression. See United Nations Human Rights Committee, General Comment 34. Article 19: Freedoms of opinion and expression (CCPR/C/GC/34), para. 46. Similarly, the Secretary-General and the UN Special Rapporteur on Counter-Terrorism have expressed concerns about the 'troubling trend' of criminalizing the glorification of terrorism, stating that this amounts to an inappropriate restriction on expression. See Protecting Human Rights and Fundamental Freedoms While Countering Terrorism. Report of the Secretary-General (A/63/337) and United Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/31/65).

and related offences that are overly-broad[22] or attach criminal sanctions to conduct that falls short of incitement to terrorism or advocacy of national, racial or religious hatred constituting incitement to violence.[23]

Laws and policies addressing violent extremism similarly raise definitional concerns. While the term 'violent extremism' and related notions such as 'extremism' and 'radicalization' are prominently present in current political discourse at the international, regional, and national levels, none of these terms have internationally agreed definitions.[24] Many of the relevant definitions found in domestic laws and policies have been criticized for being vague and at times encompassing manifestations that are lawful under international human rights law.[25] In some jurisdictions, these concepts have become dissociated from violence,[26] thereby raising the potential for abusive implementation, as such definitions risk selectively blurring the distinction between belief and violent conduct. Such approaches, especially when not accompanied by robust safeguards, risk leading to the suppression of views that deviate from the social norms accepted by the majority, under the guise of preventing extremism; and measures may target thought, belief, and opinion, rather than actual conduct.

22 See, for example, Protecting Human Rights and Fundamental Freedoms While Countering Terrorism. Report of the Secretary-General, (A/68/298); Report of the United Nations High Commissioner for Human Rights on the Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/28/28); International Commission of Jurists, Report of the Eminent Jurists Panel on Terrorism, Counter-Terrorism and Human Rights (2009). See also, Cathal Sheerin, 'The threat of 'glorifying terrorism' laws' (IFEX, 2 February 2017) https://www.ifex.org/europe_central_asia/2017/02/02/glorifying_terrorism_charges/, accessed 12 March 2018; EDRi, 'European Union Directive on counter-terrorism is seriously flawed' (30 November 2016) https://edri.org/european-union-directive-counterterrorism-seriously-flawed/, accessed 12 March 2018; Amnesty International, 'EU: Orwellian counter-terrorism laws stripping rights under guise of defending them' (17 January 2017) https://www.amnesty.org/en/latest/news/2017/01/eu-orwellian-counter-terrorism-laws-stripping-rights-under-guise-of-defending-them/, accessed 12 March 2018; Amar Toor, 'France extends draconian anti-terrorism laws' (The Verge, 17 February 2016) https://www.theverge.com/2016/2/17/11031006/france-extends-state-of-emergency-paris-attacks , accessed 12 March 2018; Amnesty International, 'Tweet… if you dare. How counter-terrorism laws restrict freedom of expression in Spain' (March 2018), Index no. EUR 41/7924/2018.

23 See Article 20, International Covenant on Civil and Political Rights. See also, Amnesty International, 'Tweet… if you dare. How counter-terrorism laws restrict freedom of expression in Spain'. In France, the Constitutional Court has recently struck down an amendment to the Penal Code criminalizing 'regular consultation' of content deemed to be inciting or glorifying terrorism. See Nadim Houry, 'French legislators rebuked for seeking to criminalize online browsing' (Human Rights Watch, 15 December 2017) https://www.hrw.org/news/2017/12/15/french-legislators-rebuked-seeking-criminalize-online-browsing, accessed 12 March 2018; Conseil constitutionnel, Décision n° 2017-682 QPC du 15 décembre 2017.

24 Acknowledging this shortcoming, the Secretary-General in his Plan of Action to Prevent Violent Extremism stated that violent extremism is to be defined at the national level, while emphasizing that such definitions must be consistent with obligations under international human rights law.

25 See Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism (A/HRC/31/65) and Report on Best Practices and Lessons Learned on How Protecting and Promoting Human Rights Contribute to Preventing and Countering Violent Extremism. Report of the United Nations High Commissioner for Human Rights (A/HRC/33/29).

26 A number of countries also target 'extremism' that is non-violent. For example, extremism is defined in the United Kingdom as 'vocal or active opposition to fundamental values, including democracy, the rule of law, individual liberty and the mutual respect and tolerance of different faiths and beliefs'. See HM Government, Prevent Strategy (2011), Annex A; HM Government, Counter-Extremism Strategy (2015, October), para. 1.

The potential and actual uses of the counter-terrorism and preventing violent extremism framework to stifle dissent, to persecute human rights defenders, journalists, and the political opposition, and to criminalize the work of humanitarian organizations has been addressed at length elsewhere.[27] Online platforms having to operationalize such laws and policies may find themselves contributing to the negative human rights impact of these frameworks. Even in cases where related domestic legal and policy frameworks do not present these shortcomings, the discrepancies between different domestic frameworks inevitably raise difficulties for online platforms, in particular those that operate worldwide (or at least in numerous jurisdictions), making it difficult to comply with all relevant domestic laws.

B. Online Platforms as De Facto Content Regulators

1) Means and Modalities of Content Review

Many platforms rely on a combination of artificial intelligence (AI) and human expertise to review and moderate content. The use of AI to spot terrorist or violent extremist content is a relatively new development,[28] and platforms such as Facebook acknowledge that it is a tool that must be complemented by human review.[29] Using algorithms to assess compliance with the law and terms of service or community standards provides for a time-efficient way for dealing with large volumes of material. It is one advocated by bodies such as the European Commission, which encourages online platforms to 'step up investment in, and use of, automatic detection technologies'.[30]

Algorithms, however, are not fool-proof, as they are not necessarily well-equipped to understand context, different forms of humour and satire,[31] and may not pick up on certain subtleties.[32] For example, hash-matching or even fingerprinting algorithms are not capable of analysing meaning or context, such as whether certain content contains

27 See Interagency Standing Committee, Sanctions Assessment Handbook: Assessing the Humanitarian Implications of Sanctions (United Nations, 2004); Kate Mackintosh and Patrick Duplat, 'Study of the Impact of Donor Counter-Terrorism Measures on Principled Humanitarian Action' (United Nations Office for the Coordination of Humanitarian Affairs and the Norwegian Refugee Council, July 2013).

28 See Monika Bickert and Brian Fishman, note 19.

29 Monika Bickert and Brian Fishman, 'Hard Questions: Are We Winning the War on Terrorism Online?' (Facebook, 28 November 2017) https://newsroom.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/, accessed 15 January 2018; Lynsey Barber, 'Facebook's now using artificial intelligence to remove terror content' (CityA.M., 29 November 2017) http://www.cityam.com/276626/facebooks-now-using-artificial-intelligence-remove-terror, accessed 15 January 2018.

30 European Commission, 'Communication on tackling illegal content online, towards enhanced responsibility of online platforms' (28 September 2017) https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms, accessed 15 January 2018.

31 Julia Krüger, 'Kommentar: Das Recht auf den Tweet' (Netzpolitik.org, 6 January 2018) https://netzpolitik.org/2018/kommentar-das-recht-auf-den-tweet/, accessed 15 January 2018.

32 See, for example Julia Reda, 'When filters fail: These cases show we can't trust algorithms to clean up the Internet' (28 September 2017) https://juliareda.eu/2017/09/when-filters-fail/, accessed 15 January 2018.

terrorist propaganda or hate speech, or reveals criminal intent.[33] As a result, they may end up removing not only videos produced by terrorist groups for recruitment purposes, but also media analysis of these videos, or even footage uploaded by human rights groups reporting on abuses.[34] Some machine-learning algorithms, such as natural language processing tools, are better suited for the kind of analysis required in this context. However, even their use comes with limitations. Experts argue that these tools cannot be applied with the same reliability across different contexts, as language use differs across different cultural, demographic, and linguistic groups.[35] An algorithm trained to parse out anti-Muslim hate speech may achieve lower levels of accuracy when attempting to identify anti-Semitic hate speech, for example. As with any machine learning algorithm, these tools can also amplify existing biases (including social and other bias existing in a language). This may result in algorithms over-censoring groups that are already marginalized.[36] Dialects that are underrepresented in mainstream text are also more likely to be misinterpreted, leading to algorithms performing less accurately,[37] and many of the existing natural language processing tools only work for English or other high-resource languages.[38]

These limitations suggest that unchecked use of algorithms for content management may lead to screening that is over- or under-inclusive. The margin of error would prove particularly problematic in the case of large online platforms. For example, Facebook has at some point reported that it receives one million user violation reports a day.[39] If all these reports were processed through AI tools, it would mean hundreds of thousands of mistaken decisions per day.[40] For meaningful oversight of decisions made by AI tools, integrating the human-in-the-loop principle needs to be ensured. Unfortunately, most social media platforms do not provide meaningful information

33 *Ibid*. See also Evan Engstrom and Nick Feamster, 'The limits of filtering: A look at the functionality & shortcomings of content detection tools' (Engine, March 2017) 13-15 and 17-21.

34 See, for example, Daphne Keller, 'Problems with filters in the European Commission's platforms proposal' (Stanford Center for Internet and Society, 5 October 2017) http://cyberlaw.stanford.edu/blog/2017/10/problems-filters-european-commissions-platforms-proposal, accessed 12 March 2018.

35 See Bermingham et al., 'Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation', Proceedings of the International Conference on Advances in Social Network Analysis and Mining (2009), 3; Su Lin Blodgett and Brendan O'Connor, 'Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English', Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Conference (2017) 1-2, https://arxiv.org/pdf/1707.00061.pdf, accessed 15 January 2018.

36 Jieyo Zhao et al., 'Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints', Proceedings of the Conference on Empirical Methods in Natural Language Processing (2017), https://arxiv.org/pdf/1707.09457, accessed 15 January 2018.

37 Su Lin Blodgett and Brendan O'Connor, note 35, 1-2; Rachael Tatman, 'Gender and Dialect Bias in YouTube's Automatic Captions', Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing, 53–59 (2017), http://rachaeltatman.com/sites/default/files/papers/EthNLP06.pdf, accessed 15 January 2018. See also Natasha Duarte, Emma Llanso, Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' (Centre for Democracy and Technology, November 2017) 15.

38 *Id*., 14.

39 Sara Ashley O'Brien, 'Facebook gets 1 million user violation reports a day' (CNN Tech, 12 March 2016) http://money.cnn.com/2016/03/12/technology/sxsw-2016-facebook-online-harassment/index.html, accessed 15 January 2018.

40 Natural language processing tools reportedly do not possess an accuracy rate higher than 80%. See Natasha Duarte, note 37, 5.

on content review procedures and the criteria that determine whether certain content will be reviewed by AI, human moderators, or both.[41]

Having content reviewed by human moderators does not necessarily assuage all concerns. Assessing what may amount to hate speech, incitement to terrorism, 'glorification' of terrorism or violent extremist content frequently requires a rather complex analysis to be conducted by a highly trained, specialized, and adequately resourced workforce. The reality, however, does not seem to fit this picture. Reports indicate that low-paid and insufficiently trained moderators frequently end up being the *de facto* gatekeepers of freedom of expression online.[42] Moreover, bearing in mind the overwhelming pace at which content is posted, relying primarily on human monitoring, particularly in near real-time, would be next to impossible.

Many large social media platforms operate worldwide, or at least in numerous jurisdictions. This makes it difficult or even impossible to produce a universally suitable set of rules for their algorithms and moderators. As described above, such rules need to take into account the differences between domestic legal systems and the scope of prohibited content in different jurisdictions and linguistic, cultural, social, and other contexts.

2) Safeguards, Transparency, and Accountability

Detailed information on the means and modalities of content control exercised by online platforms is scarce. Terms of service and community standards are commonly drafted in vague terms and do not provide sufficiently clear guidance on the circumstances under which content may be blocked, removed or restricted, or access to a service restricted or terminated, including the criteria used for such assessments. Facebook's Director of Global Policy Management, Monika Bickert, explained that the company does not share details of its policies to avoid encouraging people 'to find workarounds'.[43] This also means reduced transparency, including when it comes to the internal consistency of the application of these policies, and may as a result lead to reduced accountability.

41 See Monika Bickert, note 19. While the so-called 'Facebook files' provide some insight into the moderation process, many questions remain. Moreover, moderation policies of other major social network platforms remain obscure.

42 See Olivia Solon, 'Counter-terrorism was never meant to be Silicon Valley's job. Is that why it's failing?' (The Guardian, 29 June 2017) https://www.theguardian.com/technology/2017/jun/29/silicon-valley-counter-terrorism-facebook-twitter-youtube-google; accessed 15 January 2018; Olivia Solon, 'Underpaid and overburdened: The life of a Facebook moderator' (The Guardian, 25 May 2017) https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content, accessed 15 January 2018; Till Krause and Hannes Grassegger, 'Inside Facebook' (Süddeutsche Zeitung, 15 December 2016) http://www.sueddeutsche.de/digital/exklusive-sz-magazin-recherche-inside-facebook-1.3297138, accessed 15 January 2018; Nick Hopkins, 'Facebook struggles with 'mission impossible' to stop online extremism' (The Guardian, 24 May 2017) https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremism, accessed 15 January 2018.

43 Monika Bickert, 'At Facebook we get things wrong – but we take our safety role seriously' (The Guardian, 22 May 2017) https://www.theguardian.com/commentisfree/2017/may/22/facebook-get-things-wrong-but-safety-role-seriously, accessed 15 January 2018.

Information provided *ex post facto* is similarly lacking. Users are frequently not informed of the origin of removal requests, the procedure that led to removal or rejection of removal and the criteria used.[44] They also have limited possibilities to challenge decisions to restrict content or access to a service. To tackle this shortcoming, the recently adopted German Network Enforcement Act requires companies to report on a biannual basis describing their means and modalities for handling complaints and disclosing the criteria for removing or blocking content. It similarly calls on companies to inform both the complainant and the users affected by particular measures, including the reasoning for the decision. The law, however, does not explicitly require companies to provide users with the option to challenge these decisions.

As relevant measures by private companies are generally taken in enforcement of terms of service and not on the basis of specific legislation, it is frequently not possible to challenge them in court. Platforms may also impose internal or other alternative dispute resolution mechanisms, should disputes arise. Moreover, as private bodies, such platforms are generally not subject to democratic or independent oversight in the way that public authorities are, despite their effectively carrying out regulatory, executive, and adjudicative functions.[45] Removing the possibility of independent, including judicial, review of measures that interfere with human rights is problematic in general and particularly so having in mind recent legal and policy developments. Businesses are potentially facing fines and sanctions imposed by States if they do not restrict unlawful content.[46] On the other hand, should they remove lawful content in the process, affected individuals have limited avenues of redress. In case of doubt, businesses will more likely err on the side of over-censoring.

C. The Scope of Responsibility of Online Intermediaries

Online platforms that host or store user-generated content and enable access to and retrieval of this content by the author and other users[47] qualify as Internet intermediaries. Such intermediaries, as opposed to authors and publishers of content, are generally protected against liability for third-party content, with certain caveats. The scope of this exemption differs depending on jurisdiction.[48] For example, under the EU e-Commerce Directive, hosting intermediaries do not incur liability as long

44 See Annemarie Bridy and Daphne Keller, 'U.S. Copyright Office Section 512 Study: Comments in Response to Notice of Inquiry' (31 March 2016) 29.

45 'Zachary Loeb – Who moderates the moderators? The Facebook files' (Boundary 2, 7 June 2017) http://www.boundary2.org/2017/06/zachary-loeb-who-moderates-the-moderators-on-the-facebook-files/, accessed 15 January 2018.

46 See Section C *infra* : The Scope of Responsibility of Online Intermediaries.

47 Monica Horten, 'Content 'responsibility': The looming cloud of uncertainty for Internet intermediaries' Center for Democracy and Technology (September 2016) 5. See also Jaani Riordan, The Liability of Internet Intermediaries (Oxford University Press, 2016) Chapter 2.

48 See Article 19, 'Internet Intermediaries: Dilemma of Liability' (2013); Eric Goldman, 'Facebook isn't liable for fake user account containing non-consensual pornography' (Forbes, 8 March 2016) https://www.forbes.com/sites/ericgoldman/2016/03/08/facebook-isnt-liable-for-fake-user-account-containing-non-consensual-pornography/#40ba670379b2, accessed 15 January 2018.

as they 'expeditiously' remove or disable access to illegal content once they have 'actual knowledge' of its existence.[49] Under EU law, it is not permitted to impose a general obligation to monitor content or to 'actively seek facts or circumstances indicating illegal activity'.[50] Similarly, so-called 'notice and stay-down' injunctions, involving an obligation to ensure that content, once removed, will not reappear on the platform, are also problematic to the extent that their implementation requires general monitoring.

The idea of introducing such a burden on intermediaries has emerged in current debates, with policy-makers calling for stricter regulation of the liability of Internet intermediaries when it comes to countering terrorism, violent extremism, and hate speech. Proposals include imposing fines and other sanctions on social media platforms 'that fail to take action against terrorist propaganda and violent content',[51] and even having social media companies bear the costs of authorities policing content online.[52] The introduction of criminal liability for platforms was discussed and ultimately rejected by the European Parliament in the context of the Directive on Combating Terrorism. However, the European Commission, in its *Communication on Tackling Illegal Content Online: Towards enhanced responsibility of online platforms*, recommended that tech companies proactively look to identify illegal content on their platforms with the help of artificial intelligence, stressing that 'online platforms should also be able to take swift decisions […] without being required to do so on the basis of a court order or administrative decision'.[53] The Commission considers that online platforms can take the recommended proactive measures without fear of losing their liability exemption under the e-Commerce Directive.[54]

Other developments similarly come close to recommending or requiring proactive monitoring by intermediaries, potentially also affecting the internal consistency of the EU legal framework. Article 28a of the review proposal to the Audio-Visual Media Services (AVMS) Directive[55] provides that video-sharing platforms[56] must

49 Article 14, Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (e-Commerce Directive).

50 Article 15, e-Commerce Directive. See also *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Netlog NV*. Case C-360/10 (2012) (European Court of Justice).

51 Toor, note 14; Gibbs, note 14.

52 See House of Commons Home Affairs Committee, 'Hate crime: abuse, hate and extremism online' (25 April 2017); McCann, note 14.

53 European Commission, note 30.

54 While the Communication addresses the compatibility of such proactive measures with Article 14 of the e-Commerce Directive, it does not pay similar attention to Article 15.

55 European Commission, Proposal for a Directive of the European Parliament and of the Council Amending Directive 2010/13/EU: On the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audio-visual media services in view of changing market realities.

56 It must be noted that some civil society organizations and some Member States caution against the inclusion of video-sharing platforms, in particular social media ones, within the scope of the Directive. See EDRi, 'EDRi Position on AVMSD Trilogue Negotiations' (14 September 2017) https://edri.org/files/AVMSD/edriposition_trilogues_20170914.pdf, accessed 15 January 2018.

take measures to 'protect all citizens' from content containing incitement to violence, discrimination or hate.[57] In addition to providing for a rather vague definition of such content,[58] the proposed provision may be interpreted as requiring proactive monitoring.[59]

As a result of such developments, the EU will have to assess the compatibility of the e-Commerce Directive with other instruments addressing the role of Internet intermediaries in combating hate speech and other illegal content, especially in light of the decision not to reopen the e-Commerce Directive. It is in this vein that the European Commission has adopted the above-mentioned *Communication on Tackling Illegal Content Online*[60] and is developing measures that set common requirements across the Union for companies when it comes to removing illegal content, as a means to avoid 'overzealous rules that differ between EU countries'.[61]

What seems to be missing is the human rights-based analysis of such new obligations. This shortcoming comes even though human rights concerns posed by far-reaching intermediary liability and, in particular, its negative impact on freedom of speech and interlinked rights, have repeatedly been flagged by international human rights mechanisms[62] and civil society actors.[63] It is questionable whether the course of action proposed in the Commission's *Communication* can be construed in line with human rights standards,[64] including as spelled out in the EU Council's *Human Rights*

57 See EDRi, 'EDRi's analysis on the CULT compromise on Article 28a of the draft Audiovisual Media Services Directive (AVMSD) proposal' (13 April 2017) https://edri.org/files/AVMSD/compromise_article28a_analysis_20170413.pdf, accessed 15 January 2018.

58 For example, a compromise amendment under discussion provides for the following: 'protect all citizens from content containing incitement undermining human dignity, incitement to terrorism or content containing incitement to violence or hatred directed against a person or a group of persons defined by reference to nationality, sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, gender, gender expression, gender identity, sexual orientation, residence status or health.' (emphasis added) See European Parliament. Committee on Civil Liberties, Justice and Home Affairs. (2016) Amendments 47-171. (2016/0151(COD)).

59 While the draft explicitly mentions that it is without prejudice to articles 14 and 15 of the e-Commerce Directive, the intended scope of the duty of care is still unclear. See also Horten, note 47, 14.

60 European Commission, 'Liability of online intermediaries', (15 June 2017) https://ec.europa.eu/digital-single-market/en/liability-online-intermediaries, accessed 15 January 2018.

61 Catherine Stupp, 'Gabriel to start EU expert group on fake news' (Euractiv, 30 August 2017) https://www.euractiv.com/section/digital/news/gabriel-to-start-eu-expert-group-on-fake-news/, accessed 15 January 2018.

62 See Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (A/HRC/35/22), para. 49. See also, Joint Declaration by the United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, presented at the UNESCO World Press Freedom Day event (3 May 2016).

63 See note 48.

64 For criticism of the Communication, see for example Daphne Keller, note 34; Graham Smith, 'Towards a filtered internet: the European Commission's automated prior restraint machine' (Cyberleagle, 25 October 2017) http://www.cyberleagle.com/2017/10/towards-filtered-internet-european.html, accessed 12 March 2018.

Guidelines on Freedom of Expression Online and Offline,[65] bearing in mind its emphasis on protecting intermediaries from an obligation of blocking Internet content 'without prior due process'. Indeed, the *Communication* seems to stress *ex post facto* modalities of redress at the expense of 'prior due process'. In this respect, it states that platforms should be able to take 'swift decisions' to take action with respect to illegal content 'without being required to do so on the basis of a court order or administrative decision'. This is the case in particular when such content has been flagged by a law enforcement authority. Law enforcement authorities may be so-called 'trusted flaggers', together with other 'specialized entities with specific expertise in identifying illegal content'. In some cases, platforms 'may remove content upon notification from the trusted flagger without further verifying the legality of the content themselves'.

One entity identified as a trusted flagger in the context of assessing terrorist and violent extremist content is the Internet Referral Unit (IRU) of Europol. The IRU flags content that contravenes the EU legal framework related to terrorism and also content that goes against the terms of service set by platforms.[66] However, terms of service instituted by platforms commonly impose restrictions that go beyond what could lawfully be imposed in compliance with freedom of expression standards.[67] This approach creates the risk that content will be blocked, filtered or removed beyond what would be permissible under international human rights law. It may also result in undermining regular safeguards that protect against excessive interference, including the right to an effective remedy, as the end decision is ultimately delegated to private entities.[68]

Relevant developments have to be noted at Member State level as well. Germany has recently adopted the controversial[69] Network Enforcement Act,[70] imposing onerous obligations on social media platforms with more than two million registered users. Platforms falling within the ambit of the law face fines of up to €50 million if they

65 Council of the European Union, EU Human Rights Guidelines on Freedom of Expression Online and Offline (Foreign Affairs Council meeting, Brussels, 12 May 2014) http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/EN/foraff/142549.pdf, accessed 12 March 2018.

66 See European Parliament, 'Question for written answer to the Commission' (16 March 2017) http://www.europarl.europa.eu/sides/getDoc.do?type=WQ&reference=E-2017-001772&language=FR, accessed 12 March 2018; Answer given by Mr Avramopoulos on behalf of the Commission (12 June 2017) http://www.europarl.europa.eu/sides/getAllAnswers.do?reference=E-2017-001772&language=EN, accessed 12 March 2018. See also, Graham Smith, note 64.

67 See e.g. Elizabeth Nolan Brown, 'YouTube says no to sexual humor, profanity, partial nudity, political conflict, and "sensitive subjects" in partner content' (Reason, 1 September 2016) http://reason.com/blog/2016/09/01/youtube-bans-sex-drugs-and-politics , accessed 12 March 2018. As privately-run outlets, social media platforms can of course decide to shape the content hosted by them in order to facilitate the creation of a space that fits their business model, by enabling a more family-friendly or minor-friendly environment, for example.

68 See European Digital Rights (EDRi). (2011, January). The Slide from 'Self-Regulation' to 'Corporate Censorship'. Retrieved from https://edri.org/files/EDRI_selfreg_final_20110124.pdf.

69 'Wirtschaft und Aktivisten verbünden sich gegen Maas-Gesetz' (Der Spiegel, 11 April 2017) http://www.spiegel.de/netzwelt/netzpolitik/heiko-maas-wirtschaft-und-netzszene-protestieren-gegen-hassrede-gesetz-a-1142861.html, accessed 15 January 2018.

70 Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken.

fail to remove or block access to 'clearly illegal' content within 24 hours[71] and other illegal content within 7 days[72] after having been put on notice through a complaint. The law includes no guidance on how to distinguish 'clearly illegal' entries from merely 'illegal' ones. Such lack of clear guidance, when paired with a threat of hefty fines, becomes a definite incentive to over-censor in case of doubt.

Implementation of the Act started on the 1st of January 2018 and related incidents have already drawn attention to the limits of algorithmic moderation[73] as well as the discrepancies in the approach to moderating content demonstrated by different social media platforms.[74] In addition to cases of lawful content being removed by overeager platforms, some argue that it also results in obstructing prosecution of related crimes, as deletion of online content frequently results in deletion or improper retention of evidence needed to plead the case in court.[75] The Act will inevitably influence how major social media sites approach users' freedom of expression, with its impact in all probability extending beyond Germany's borders due to the cross-border nature of information flows and also the likelihood of it influencing similar legal and policy initiatives in other jurisdictions.

Changes in laws and policies aimed at more effectively tackling terrorist and extremist content and hate speech have also been contemplated in other jurisdictions. In this respect, the UK House of Commons Home Affairs Committee has recommended that Internet intermediaries proactively identify illegal content and expressed dissatisfaction with such platforms for only reviewing content after it has been flagged by users or other stakeholders and for not ensuring that blocked and removed content does not resurface.[76] Similarly, the *French-British Action Plan on the Use of the Internet for Terrorism Purposes*[77] (also known as the Macron-May Plan) calls on platforms to proactively identify terrorist content and prevent it from being made available by automating the detection and suspension or removal of content, based on both the posting person or entity and the actual content of the post. This measure

71 Unless the social media network agrees a different timeline with the competent law enforcement authority. Netzwerksdurchsetzungsgesetz, Article 1 §3 (2) No. 2.

72 Unless the unlawful character of the content in question depends on factual circumstances to be determined or unless the social media network transmits the case to an authorized self-regulatory mechanism (Einrichtung der regulierten Selbstregulierung). Netzwerkdurchsetzungsgesetz, Article 1 §3 (2) No. 3.

73 See note 31.

74 *Ibid*.

75 Bernhard Rohleder, 'Germany set out to delete hate speech online. Instead, it made things worse.' (The Washington Post, 20 February 2018) https://www.washingtonpost.com/news/theworldpost/wp/2018/02/20/netzdg/?utm_term=.331d14c7fb0a, accessed 12 March 2018.

76 House of Commons Home Affairs Committee, note 52. See also, Elliot Harmin, '"Notice-and-stay-down" is really "filter-everything"' (Electronic Frontier Foundation, 21 January 2016) https://www.eff.org/deeplinks/2016/01/notice-and-stay-down-really-filter-everything, accessed 15 January 2018.

77 French-British Action Plan on the Use of the Internet for Terrorism Purposes (Paris, 13 June 2017) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/619333/french_british_action_plan_paris_13_june_2017.pdf, accessed 15 January 2018.

drew criticism for advocating both far-reaching monitoring and prior restraint.[78] The Plan also recommends measures that go beyond the existing 'notice and take-down' process, which has also been reinforced through the establishment of Europol's Internet Referral Unit and the UK's own domestic referral unit, raising the possibility of a 'notice and stay-down' obligation.

4. CONCLUSION

The Internet has frequently been described as a neutral tool that can be instrumentalised in various ways.[79] It is fundamental in facilitating the public's ability to seek, receive, and impart information and may provide a platform for persons and groups that are less included in debates of public interest, such as women and individuals belonging to minority groups, but it also enables terrorist groups and other criminal actors to convey their messages and use it as a recruitment and operational planning tool.

As online content continues to be generated at a staggering rate, attempts to control its flow encounter significant challenges and, due to the particularities of the digital space, tech companies running these online platforms are better positioned to regulate their functioning, while State powers in this respect may be more limited. There are clear expectations on the part of States that online platforms take more responsibility when it comes to illegal third-party content. Many governments view the use of automated decision-making tools as an essential component of handling content. The choice is understandable, having in mind the volume of the material that is being produced, the pace of such production and the need to take swift action. However, the limitations of existing technology are significant. If algorithms are used for regulating content, they become the rule, the rule-maker in the case of machine learning algorithms, and the tool for enforcement. The rules behind the algorithms become the *de facto* standards for the platform and beyond.

The duty of States to protect the human rights of those within their jurisdiction, including from undue interference by third parties such as businesses, is well-established. Outsourcing such tasks – whether formally or informally, through actively encouraging corporate governance or through omission or acquiescence – without establishing adequate safeguards and oversight systems, fails to comply with that duty.[80] The rise of automated processes without a corresponding strengthening of users' rights is likely to lead to undermined protection, and while ensuring *ex post*

78 See, for example, Monica Horten, 'Macron-May Internet deal: Necessary measures or prior restraint?' (Iptegrity.com, 28 July 2017) http://www.iptegrity.com/index.php/internet-freedoms/1068-macron-may-internet-deal-necessary-measures-or-prior-restraint, accessed 15 January 2018.

79 Anja Mihr, Cyber Justice: Human Rights and Good Governance for the Internet (Springer, 2017), 47.

80 See, for example, Emily B. Laidlaw, Regulating Speech in Cyberspace (Cambridge University Press, 2015) Chapter 6.

facto safeguards and modalities for redress is important, it is not sufficient, particularly as existing studies indicate that these tools go underused.[81]

There are, of course, legitimate and practical justifications for stressing the role and responsibility of social media companies in the context of countering terrorism and violent extremism. Due to the control and influence they exercise over content on their platforms, meaningful action could not be taken without their cooperation. Having private actors such as social media companies increasingly undertake traditionally public tasks in the context of Internet governance is probably unavoidable, especially as public authorities (including the judiciary) in most States do not have the human or technical resources to satisfactorily perform these tasks.

While it is inevitable for relevant private actors to play an increasingly significant role, including the taking up of quasi-executive and quasi-adjudicative tasks, this should not be done without proper guidance and safeguards. At this point, however, the outsourcing results in lowering or removing existing human rights safeguards and protections. Social media companies are stuck with tasks that they are not particularly well-equipped to carry out. For example, it is questionable whether private actors are well-placed to assess whether a particular measure is necessary and proportionate in the interest of national security or public order.

Social media platforms should be given clear and detailed instructions and guidance if they are to carry out such assessments. If control over elements of the right to freedom of expression are outsourced to these outlets, independent oversight of their conduct in this respect needs to be ensured, to guarantee transparency, accountability and respect for the right to remedy of individuals whose rights are unjustly interfered with in the process. The necessity for safeguards is not simply due to intermediaries lacking the relevant legal expertise, but a basic matter of legal principle requiring that measures impacting human rights be subjected to independent oversight by public, preferably judicial, authorities rather than left up to private bodies.

The challenges that arise in this domain call for ways to bridge public and private dimensions involved in promoting and protecting human rights. This in turn would require ensuring complementarity and synergy between various systems of regulation.[82]

81 See note 44, Appendix B.
82 See note 80, 233-234.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

From Grey Zone to Customary International Law: How Adopting the Precautionary Principle May Help Crystallize the Due Diligence Principle in Cyberspace

Peter Z. Stockburger
Dentons US LLP
San Diego, California, USA
peter.stockburger@dentons.com

Abstract: The international principle of "due diligence" is well recognized under international law, and is an outgrowth of the general obligation of States to "do no harm". The due diligence principle imposes an obligation on States to take affirmative action to ensure their territory or objects over which they maintain sovereign control are not used for internationally wrongful purposes. The due diligence principle has been recognized by international scholars and jurists since the early 20th century, and has been adopted as a principle of customary international law in the international environmental law context by States and courts, including the International Court of Justice. The International Court of Justice has specifically endorsed a procedural aspect of due diligence – that States must conduct environmental impact assessments, where appropriate, as a precautionary measure to ensure their territory is not used for internationally wrongful purposes. In 2013 and 2017, the Tallinn Manual and Tallinn Manual 2.0 confirmed the due diligence principle applies in cyberspace. However, in both manuals, the experts could not agree on the scope of its application. And, in 2017, the Tallinn Manual 2.0 experts agreed that the due diligence obligation does not include a preventive feature, as is reflected in international environmental law. This paper examines this grey area of international law, and whether and to what extent the

precautionary principle, as adopted in the international environmental law context, could be applied in cyberspace. After an examination of the precautionary principle as applied, this paper argues its application in cyberspace would help crystallize the due diligence principle from a grey zone in international law into customary international law of cyberspace by introducing a procedural due diligence requirement for States to conduct a cyber impact assessment where appropriate.

Keywords: *due diligence, cyber due diligence*

1. INTRODUCTION

The principle of State sovereignty is considered "the most fundamental" principle of all international law,[1] and has been defined as the "supreme authority of every [S]tate within its territory"[2] to exert "independence" over the "functions of a State" to the "exclusion of any other State".[3] This principle, however, is not without limit. A number of "principles and rules of conventional and customary international law derive from the general principle of sovereignty",[4] including the "corollary"[5] principle of non-intervention, which is codified at Article 2 of the United Nations (UN) Charter and restricts States from unlawfully interfering against the territorial integrity or political independence of another State.[6] The principle of non-intervention therefore restricts States in their exercise of sovereignty from using their territory or objects over which they maintain sovereign control for purposes "detrimental to the rights of other States."[7] This specific obligation is often referred to as the duty to not commit transboundary harm,[8] and is well reflected in the writings of Oppenheim as early as 1912,[9] the 1928 *Island of Palmas* award,[10] and in the International Court of Justice's (ICJ or Court) 1949 *Corfu Channel* judgment.[11]

1 Michael Schmitt, Grey Zones in the International Law of Cyberspace, 42 Yale J. of Int'l L. Online 1, 4 (2017).
2 Lassa Oppenheim, Oppenheim's International Law, at 564 (Robert Jennings & Arthur Watts eds., 9th edn, 1992).
3 *Island of Palmas (Neth. v. U.S.)*, 2 RIAA 829, 838 (Perm. Ct. Arb. 1928) (hereinafter, *"Island of Palmas"*).
4 Int'l Group of Experts, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations 11 (Rule 1) (Michael N. Schmitt ed., 2017) (hereinafter, "Tallinn Manual 2.0").
5 *Military and Paramilitary Activities in and against Nicaragua (Nicar. V. U.S.)*, 198 I.C.J. 14, 106 (June 27) (hereinafter, *"Nicaragua"*).
6 *Ibid*.; U.N. Charter Art. 2(4).
7 Schmitt, note 1, at 11.
8 Stephen Fietta et al., The South China Sea Award: A Milestone for International Environmental Law, The Duty of Due Diligence and The Litigation of Maritime Environmental Disputes? 29 Geo. Envtl. L. Rev. 711, 723 (2017).
9 Lassa Oppenheim, International Law: A Treatise, 243-44 (2nd edn, 1912).
10 *Island of Palmas*, note 3, at 829-90.
11 *Corfu Channel (U.K. v. Alb.)*, 1949 I.C.J. 4, 22 (Apr. 9) (hereinafter, *"Corfu Channel"*).

To carry out this prohibition against transboundary harm, and by extension the principle of non-intervention, States have agreed to carry out their activities with "due diligence." The due diligence obligation imposes an independent duty on States to take affirmative action to stop or prevent their territory, or the items or persons within their jurisdictional control, from knowingly being used to cause internationally wrongful acts.[12] This principle is well established "in the rules, and interpretation thereof, of numerous specialised regimes of international law[,]"[13] most notably in international environmental law. In 2010, the ICJ affirmed the principle of due diligence as reflective of customary international law in its *Case Concerning Pulp Mills on the River Uruguay* between Argentina and Uruguay (*Pulp Mills*) judgment[14] wherein the Court endorsed a preventive interpretation of the principle as "a customary rule"[15] and made clear that a State is "obliged to use all the means at its disposal in order to avoid activities which take place in its territory, or in any area under its jurisdiction, causing significant damage to the environment of another State".[16] The ICJ specifically recognized States have a procedural due diligence obligation to conduct an environmental impact assessment (EIA) "before embarking on an activity having the potential adversely to affect the environment of another State[.]"[17] This principle, generally known as the precautionary principle in international environmental law, requires States to take preventive measures even in the absence of scientific certainty. The principle was further endorsed by the ICJ in its 2015 judgment in the case concerning the *Construction of a Road in Costa Rica Along the San Juan River* between Nicaragua and Costa Rica (Costa Rica).[18]

Whether and to what extent the due diligence principle, and the precautionary principle, apply in cyberspace has been the subject of extensive debate over the past five years.[19] In 2009, the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE) commissioned an independent group of experts (IGE) to examine whether and to what extent general principles of international law apply in cyberspace.[20] The IGE produced two manuals in response - the 2013 *Tallinn Manual on the International Law Applicable to Cyber Warfare ("Tallinn Manual 1.0")* and the 2017 *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations ("Tallinn Manual 2.0")*.[21] In both, the IGE endorsed the application of the due diligence principle in

12 *Ibid*.
13 Tallinn Manual 2.0, note 1, at 30, Rule 6, ¶1.
14 *Pulp Mills on the River Uruguay (Argentina v. Uruguay)*, 2010 I.C.J. 14, 55-56 (Apr. 20, 2010) (hereinafter, *"Pulp Mills"*); *Corfu Channel*, note 11, at 22.
15 *Pulp Mills*, note 14, at 55.
16 *Id*. at 55-56.
17 *Id*. at 83.
18 *Certain Activities Carried Out by Nicaragua in the Border Area (Costa Rica v. Nicaragua) and Construction of a Road in Costa Rica along the San Juan River (Nicaragua v. Costa Rica)*, 2015 I.C.J. 665, 706-707 (Dec. 16, 2015) .
19 Schmitt, note 1, at 11; Tallinn Manual 2.0, note 4, at 30 (Rule 6).
20 Tallinn Manual 2.0, note 4, at 1.
21 Int'l Group of Experts, Tallinn Manual on the International Law Applicable to Cyber Warfare (Michael N. Schmitt ed., 2013) (hereinafter, "Tallinn Manual 1.0"); Tallinn Manual 2.0, note 4.

cyberspace,[22] but could not agree on its scope.[23] In the 2017 *Tallinn Manual 2.0*, for example, the IGE agreed the due diligence principle applies in cyberspace,[24] but was "divided as to the interpretation of the due diligence obligation".[25] Specifically, the IGE agreed the principle generally applies when cyber operations "having serious adverse consequences vis-à-vis a legal right of a State are mounted from another State's territory",[26] but could not agree that there was a preventive or precautionary element tied to this obligation.[27] The IGE also noted that because "not every State involved in pre-publication consultations readily accepted the application of due diligence to cyberspace as a matter of customary law", there was a view, not shared by the IGE, "by which the premise of applicability is *lex ferenda* (what the law should be), rather than *lex lata* (current law)".[28] This view, according to the IGE, appears to be based in part on the 2013 and 2015 reports of the United Nations Groups of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of Informational Security (GGE),[29] which only agreed that States "should," rather than must, take actions necessary to put an end to cyber operations emanating from their territory which are harmful to other States.[30]

This paper examines this grey area of international law, and whether a preventive or precautionary principle should, as the *lex ferenda*, apply in cyberspace. This paper specifically explores whether applying a procedural due diligence requirement in cyberspace, similar to the procedural due diligence obligation in environmental law, would help crystallize the due diligence obligation in cyberspace and close the gap recognized by the *Tallinn Manual 2.0.* In so doing, this paper argues that States should agree to conduct a cyber impact assessment as a procedural due diligence requirement that each would undertake before embarking on an activity having the potential adversely to affect the cyber infrastructure or interests of another State. This principle, of course, is not the *lex lata*. States have not agreed to this approach in cyberspace. But because there are analogies to be drawn between significant and irreparable environmental harm and the harm that a serious and adverse cyber operation could impose on States, this paper argues the *lex ferenda* should properly consider the application of a precautionary approach in cyberspace to further ensure States have clear rules concerning due diligence in their cyber operations vis-à-vis one another.

22 Tallinn Manual 1.0, note 22, at 26.
23 *Id*. at 28.
24 Schmitt, note 1, at 11; Tallinn Manual 2.0, note 4, at 30 (Rule 6).
25 Schmitt, note 1, at 11.
26 *Ibid*.
27 *Id*. at 13; Tallinn Manual 2.0, note 4, at 41-42 (Rule 6) cmt. 42; *id*. at 44-45 (Rule 7) cmts. 7-10.
28 Schmitt, note 1, at 11; Tallinn Manual 2.0, note 4, at 31 (Rule 6) cmt. 3.
29 Schmitt, note 1, at 11.
30 Rep. of the Grp. of Governmental Experts on Devs. In the Field of Info. & Telecomm. In the Context of Int'l Sec., U.N. Doc. A/68/98, ¶ 23 (June 24, 2013) (hereinafter, "2013 GGE Report"); Rep. of the Grp. of Governmental Experts on Devs. In the Field of Info. & Telecomm. In the Context of Int'l Sec., U.N. Doc. A/70/174, ¶¶ 13(c), 28(e) (July 22, 2015) (hereinafter, "2015 GGE Report").

This paper is divided into four parts. **Part I** examines the history of the due diligence principle as it has developed under international law. **Part II** examines the development of the precautionary approach in international environmental law. **Part III** examines the application of the due diligence principle in cyberspace, as reflected in the *Tallinn Manual 1.0, Tallinn Manual 2.0*, and the 2013 and 2015 GGE Reports. And **Part IV** explores how, if adopted, a precautionary approach may help further crystallize due diligence in cyberspace by imposing a procedural due diligence obligation on States.

2. PART I - DEVELOPMENT OF DUE DILIGENCE UNDER INTERNATIONAL LAW

The obligation of "due diligence" is well recognized in international law, and dates back to the writings of Grotius and Vattel.[31] The principle has been applied in various specialized regimes of international law, including international human rights, humanitarian, trade, and environmental law.[32] The ICJ expressly endorsed the due diligence principle in its 1949 *Corfu Channel* judgment, stating there are "certain general and well-recognized principles" of international law, including "every State's obligation not to allow knowingly its territory to be used for acts contrary to the rights of other States".[33] The ICJ further endorsed the principle in the case concerning the *Prevention and Punishment of the Crime of Genocide*.

In addition to these general developments, the principle of due diligence has received considerable attention in the international environmental context. It was first endorsed in the 1938 *Trail Smelter* Arbitral Award, which determined that Canada was required to take protective measures to reduce the air pollution in the Columbia River Valley caused by sulphur dioxide emitted by zinc and lead smelter plants in Canada, only seven miles from the Canadian-US border:[34]

> Under the principles of international law, no State has the right to use or permit the use of its territory in such a manner as to cause injury by fumes in or to the territory of another or the properties or persons therein, when the case is of serious consequence and the injury is established by clear and convincing evidence.[35]

The ICJ further endorsed this principle in 2010 and 2015, and introduced the preventive principle within the due diligence obligation in the *Pulp Mills* and *Costa*

31 Stephen Fietta, et al., The South China sea Award: A Milestone for International Environmental Law, The Duty of Due Diligence and The Litigation of Maritime Environmental Disputes? 29 Geo. Envtl. L. Rev. 711, 723 (2017).

32 Fietta, note 32, at 723 (citing Friendly Relations Declaration, multiple Security Council resolutions, the four Geneva Conventions of 1949, and multiple arbitral decisions).

33 *Corfu Channel*, note 11, at 22.

34 *Trail Smelter (U.S. v. Can.)*, 3 R.I.A.A. 1905, 1965 (1938).

35 *Ibid*.

Rica judgments. In its 2010 *Pulp Mills* judgment, the ICJ affirmed the principle of due diligence as reflective of customary international law, and relied on its articulation of the principle in its 1949 *Corfu Channel* judgment.[36] From this general principle, the ICJ additionally recognized that within the due diligence principle there exists a principle of prevention which is also "a customary rule",[37] and obliges States to "use all the means at [their] disposal in order to avoid activities which take place in its territory, or in any area under its jurisdiction, causing significant damage to the environment of another State".[38] The ICJ made clear in its judgment that it may now be considered a requirement under general international law to undertake an environmental impact assessment where there is a risk that the proposed industrial activity may have a significant adverse impact in a transboundary context, in particular, on a shared resource.[39]

Although the Court's judgment in *Pulp Mills* referred only to industrial activities, the Court further expanded on the principle in its 2015 *Costa Rica* judgment and affirmed the principle of due diligence and that the requirement of an EIA "applies generally to proposed activities which may have a significant adverse impact in a transboundary context".[40] The Court stated that in order to "exercise due diligence in preventing significant transboundary environmental harm, a State must, before embarking on an activity having the potential adversely to affect the environment of another State, ascertain if there is a risk of significant transboundary harm, which would trigger the requirement to carry out an environmental impact assessment.[41] This principle, the preventive principle, is also known as the precautionary principle.

3. PART II - DEVELOPMENT OF PRECAUTIONARY PRINCIPLE

A. 1971 - 1991

Most commentators agree that the "precautionary" principle traces back to 1971 and the concept of *Vorsorgeprinzip* (foresight) under German environmental law.[42] This principle was asserted by Germany ten years later during international conferences held to discuss the protection of the North Sea,[43] and was adopted in 1987 as part of the Ministerial Declaration Calling for Reduction of Pollution, which stated in relevant part:

36 *Pulp Mills*, note 14, at 55-56; *Corfu Channel*, note 11, at 22.
37 *Pulp Mills*, note 14, at 55.
38 *Id*. at 55-56.
39 *Id*. at 83.
40 *Costa Rica*, note 18, at 706.
41 *Id*. at 706-707.
42 Ling Chen, Realizing the Precautionary Principle in Due Diligence, 25 Dal. J. Leg. Stud. 1, 4 (2016); Mary Stevens, The Precautionary Principle in the International Arena, 2 Sus. Dev. Law & Pol. 13, 13 (2002).
43 Stevens, note 42, at 13.

> [in] order to protect the North Sea from possibly damaging effects of the most dangerous substances, a precautionary approach is necessary which may require action to control inputs of such substances even before a causal link has been established by absolute clear scientific evidence.[44]

The principle was also referenced in the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer, which provides that States must "protect the ozone layer by taking precautionary measures to control equitably total global emissions that deplete it".[45]

By 1990, the principle had received widespread adherence. It was applied at the third conference on the protection of the North Sea[46] and was also included in Great Britain's 1990 White Paper on Britain's Environmental strategy, which provided:

> We must analyze the possible benefits and costs both of action and of inaction. Where there are significant risks of damage to the environment, the Government will be prepared to take precautionary action to limit the use of potentially dangerous pollutants, even where scientific knowledge is not conclusive, if the balance of the likely costs and benefits justifies it. This precautionary principle applies particularly where there are good grounds for judging either that action taken promptly at comparatively low cost may avoid more costly damage later, or that irreversible effects may follow if action is delayed.[47]

Europe further endorsed the principle in 1991 in a meeting between parties to the 1927 London Dumping Convention,[48] and in the Bamako Convention of 1991 which requires States party to prevent the "release into the environment of substances which may cause harm to humans or the environment without waiting for scientific proof regarding such harm".[49]

B. 1992 To The Present

The precautionary principle gained momentum in 1992, and was endorsed in multiple international instruments, including Article 2 of the 1992 Convention for the

44 Chen, note 42, at 5.

45 *Montreal Protocol on Substances that Deplete the Ozone Layer*, 16 September 1987, 1522 UNTS 3 (entered into force 1 January 1989).

46 Final Declaration of the Third International Conference on Protection of the North Sea, Mar. 7-8, 1990. 1 YB Int'l Envtl Law 658, 662-73 (1990).

47 This Common Inheritance: Britain's Environmental Strategy, Sept. 1990 at § 1.18.

48 London Dumping Convention Amendments (1991).

49 *Bamako Convention on the Ban of the Import into Africa and the Control of Transboundary Movement and the Management of Hazardous Wastes Within Africa*, Jan. 30, 1992, OAU/CONF/COOR/ENV/MIN/AFRI/CONV.1(1) Rev. 1, reprinted in 30 L.L.M. 773.

Protection of the Marine Environment of the Northeast Atlantic and Article 15 of the landmark Rio Declaration, which was signed at the UN Conference on Environment and Development and provides that:

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.[50]

The 1992 UN Framework Convention on Climate Change also endorsed the precautionary principle:

> The parties should take precautionary measures to anticipate, prevent, or minimize the causes of climate change and mitigate its adverse effects. Where there are threats of serious or irreversible damage, lack of full scientific certainty should not be used as a reason for postponing such measure, taking into account that policies and measure to deal with climate change should be cost-effective so as to ensure global benefits at the lowest possible cost.[51]

Article 6 of the 1995 Agreement for the Implementation of the Provisions of the 1982 UN Convention on the Law of the Sea relating to the Conservation and Management of Straddling Fish Stocks and Highly Migratory Fish Stocks further endorsed the application of the precautionary approach,[52] and provided that States party are required to use the precautionary approach to conserve, manage, and exploit the stocks of straddling fish and highly migratory fish and "shall be more cautious when information is uncertain, unreliable or inadequate".[53] Under this principle, States cannot delay or refuse to take conservation and management measures because of inadequate scientific information.[54] States are also required to implement the precautionary principle when developing scientific information and technology to mitigate uncertainties relating to the size of fish stocks, and collect data to assess the impact of certain fishing activities.[55]

50 Rio Declaration at art. 15.
51 United Nations Framework Convention on Climate Change, May 9, 1992, art. 3, para. 3, U.N. Doc. A/CONF.151/26.
52 UNGA, Conference on Straddling Fish Stocks and Highly Migratory Fish Stocks, 6th Sess, *Agreement for the Implementation of the Provisions of the United Nations Convention on the Law of the Sea of 10 December 1982 relating to the Conservation and Management of Straddling Fish Stocks and Highly Migratory Fish Stocks*, UN Doc A/CONF.164/37, September 1995.
53 *Id*. at Art. 6.1, 6.2.
54 *Ibid*.
55 *Id*. at art. 6.3.

The 2000 Cartagena Protocol on Biosafety to the Convention on Biological Diversity also applies the precautionary principle to the control of transboundary movements of genetically modified organisms,[56] wherein the principle is reflected in paragraph 4 of its preamble[57] and Articles 1, 10(6) and 11(8).[58] Articles 10(6) and 11(8), both of which track precautionary language, include language such as "lack of scientific certainty", "insufficient relevant scientific information and knowledge", and the minimization of "potential adverse effects".[59]

As noted above, the ICJ embraced the precautionary principle in the 2010 *Pulp Mills* judgment, and made clear that the due diligence principle carries with it a procedural element – the undertaking of an EIA in appropriate circumstances to determine if there is a risk of significant transboundary harm, which would trigger the requirement to carry out an environmental impact assessment.[60] The Court further articulated that the content of the EIA is to be made in "light of the specific circumstances of each case":[61]

> it is for each State to determine in its domestic legislation or in the authorization process for the project, the specific content of the environmental impact assessment required in each case, having regard to the nature and magnitude of the proposed development and its likely adverse impact on the environment as well as to the need to exercise due diligence in conducting such an assessment.[62]

The Court further elaborated on this procedural due diligence obligation in the *Costa Rica* judgment, noting that if the:

> environmental impact assessment confirms that there is a risk of significant transboundary harm, the State planning to undertake the activity is required, in conformity with its due diligence obligation, to notify and consult in good faith with the potentially affected State, where that is necessary to determine the appropriate measures to prevent or mitigate that risk.[63]

56 *Cartagena Protocol on Biosafety to the Convention on Biological Diversity*, 29 January 2000, 2226 UNTS 208 (entered into force 11 September 2003).
57 *Id*. at preamble, para. 4.
58 *Id*. at arts 1, 10(6), 11(8).
59 *Id*. at art 10(6), 11(8).
60 *Costa Ric*a, note 18, at 706-07.
61 *Id*. at 707.
62 *Pulp Mills*, note 11, at 83.
63 *Costa Rica*, note 18, at 707.

4. PART III - DUE DILIGENCE IN CYBERSPACE

Whether and how the principle of due diligence and its precautionary approach apply in cyberspace has been examined closely by scholars and jurists over the past five years. Although there are myriad opinions on the application of due diligence in cyberspace, this paper focuses solely on those opinions set out in the *Tallinn Manual 1.0, Tallinn Manual 2.0*, and the 2013 and 2015 GGE Reports.

A. Tallinn Manual 1.0

The *Tallinn Manual 1.0* endorses the principle of due diligence in cyberspace by reaffirming the principle that a State may not "allow knowingly its territory to be used for acts contrary to the rights of other States".[64] The IGE concluded that States, in their cyber operations, are to "take appropriate steps to protect those rights".[65] The scope of that obligation, however, was the subject of extensive debate and disagreement. Indeed, due diligence was only dealt with in a single rule accompanied by a brief commentary. And the IGE could not achieve consensus on the parameters of the obligation. The IGE noted that the implementation of the due diligence principle in cyberspace is complicated by the nature of harmful cyber acts, "especially time and space compression, and their often-unprecedented character."[66]

The IGE therefore adopted a knowledge standard when applying the due diligence principle in cyberspace, noting that the principle of due diligence applies only if the territorial State has "actual knowledge" of the cyber operation and/or the threat in question.[67] The IGE could not "achieve consensus" as to whether the principle of due diligence applies if "the respective State has only constructive ('should have known') knowledge".[68] In other words, the IGE agreed it was:

> unclear whether a State violates [the principle of due diligence] if it fails to use due care in policing cyber activities on its territory and is therefore unaware of the acts in question. Even if constructive knowledge suffices, the threshold of due care is uncertain in the cyber context because of such factors as the difficulty of attribution, the challenges of correlating separate sets of events as part of a coordinated and distributed attack on one or more targets, and the ease with which deception can be mounted through cyber infrastructure.[69]

64 Tallinn Manual 1.0, note 22, at 26.
65 *Ibid*.
66 *Id*. at 27.
67 *Id*. at 28.
68 *Ibid*.
69 *Ibid*.

The IGE also could not agree on whether a State must take preventive measures to ensure the cyber hygiene of the infrastructure on its territory or whether "States should be required to monitor for malicious activity that might be directed at other States".[70]

B. Tallinn Manual 2.0

The Tallinn Manual 2.0 further confirmed that the due diligence principle applies to cyber operations originating from a State's territory,[71] making clear that the principle of due diligence was reflected in international law and applied in cyberspace as the *lex lata*.[72] Notwithstanding, the IGE rejected the notion that due diligence in cyberspace involves an "obligation of prevention", stating that the group of experts was in agreement that the "due diligence principle does not encompass an obligation to take material preventive steps to ensure that the State's territory is not used in violation [of the law]".[73] In reaching this decision, the IGE stated it "carefully considered whether the due diligence principle imposes a requirement to take preventive measures, such as hardening one's cyber infrastructure, to reduce general, as distinct from particularised, risks of future cyber operations falling within the purview of the [due diligence principle].[74]

Ultimately, the IGE "rejected the premise of a requirement to take purely preventive measures of a general nature"[75] based on the difficulty in mounting comprehensive and effective defences against all possible cyber threats.[76] Such a requirement, according to the IGE, would "impose an undue burden on States, one for which there is no current basis in either the extant law or current State practice."[77] The IGE further noted that "States have not indicated that they believe such a legal obligation exists with respect to cyber operations, either by taking preventive measures on this basis or by condemning the failure of other States to adopt such measures".[78]

The IGE also noted that because knowledge is a requirement under the principle of due diligence, it would be "contradictory to expand" the principle of due diligence to "hypothetical future cyber operations"[79] because a State cannot know of a "cyber operation that has yet to be decided upon by the actor".[80] Thus, having rejected the duty of prevention, the IGE concurred that a State is "not required to monitor cyber activities on its territory".[81]

70 Michael N. Schmitt, *In Defense of Due Diligence in Cyberspace*, The Yale Law Journal Forum at 71 (June 22, 2015).
71 Tallinn Manual 2.0, note 4, at 30 (Rule 6).
72 *Id*. at 31 (Rule 6). The IGE also acknowledge a view, which no member held, that the due diligence principle is not reflective of custom based on the non-mandatory language found in the 2013 and 2015 GGE Reports.
73 *Id*. at 32 (Rule 7).
74 *Id*. at 44.
75 *Ibid*.
76 *Id*. at 45.
77 *Ibid*.
78 *Ibid*.
79 *Id*. at 45.
80 *Ibid*.
81 *Ibid*.

The IGE, did, however, acknowledge the precautionary approach in international law. It "acknowledged the contrary view, which none of them held, that the due diligence obligation extends to situations in which the relevant harmful acts are merely possible".[82] "By it, States must take reasonable measures to prevent them from emanating from their territory."[83] The IGE notes that this view is based on the existence of an obligation to "take preventive measures in the context of transboundary environmental harm".[84] According to this position, a "State must take feasible preventive measures that are proportionate to the risk of potential harm. They have to take account of technological and scientific developments, as well as the unique circumstances of each case".[85]

The IGE rejected this principle, practically, because "if such an approach were to be adopted, it would be unclear when the obligation would be breached":

> One possibility is that a breach takes place when a target State is placed at the risk of harm by virtue of the territorial State not having taken appropriate measures to prevent harmful cyber operations being mounted from or through its territory. Another is that although the due diligence principle requires States to take appropriate preventive measures, they cannot be held responsible for having failed to do so unless and until the target State actually suffers the requisite harm.[86]

The IGE concluded that the "precise threshold of harm at which the due diligence principle applies is unsettled in international law".[87]

C. The 2013 and 2015 GGE Reports

In 2013, the UN GGE issued a report on the application of "norms derived from existing international law relevant to State behavior in cyberspace".[88] Concerning the due diligence principle, the GGE concluded that States must "meet their international obligations regarding internationally wrongful acts attributable to them", and "should seek to ensure that their territories are not used by non-State actors for unlawful use" of their cyber infrastructure.[89] In 2015, the GGE reaffirmed this principle, and stated that "States should not knowingly allow their territory to be used for internationally wrongful acts" using their cyber infrastructure.[90] The use of the word "should" instead of "shall" or "must" has raised questions as to whether States truly understand that

82 *Ibid*.
83 *Ibid*.
84 *Id*. at 45-46.
85 *Id*. at 46.
86 *Ibid*.
87 Tallinn Manual 2.0, note 4, at 36.
88 2013 GGE Report, note 31, at 2.
89 *Id*. at 8, ¶23.
90 2015 GGE Report, note 31, at 8, ¶13(c)

the due diligence principle is reflective of customary international law. "[As] due diligence is purportedly a primary rule of international law, a State's violation of which constitutes an internationally wrongful act, such hesitancy to accord the rule lex lata status produces a grey zone of international law."[91]

5. ADOPTING THE PRECAUTIONARY APPROACH IN CYBER

Whether the due diligence obligation reflects the *lex lata* in cyberspace is not the focus of this paper. This paper instead questions the 2017 *Tallinn Manual 2.0* IGE's conclusion that a preventive feature of due diligence cannot apply in cyberspace. The 2017 IGE rejected the application of the precautionary approach in cyberspace because States cannot harden their cyber defenses against all possible cyber threats.[92] The IGE also rejected its application because knowledge is a requirement to trigger the due diligence principle, and it would be "contradictory to expand" the principle of due diligence to "hypothetical future cyber operations" because a State cannot know of a "cyber operation that has yet to be decided upon by the actor".[93]

These are legitimate concerns. However, they would be mitigated if States adopted a procedural due diligence obligation, similar to the standard articulated by the ICJ in the 2010 *Pulp Mills* and 2015 *Costa Rica* judgments. In particular, a procedural due diligence approach in cyberspace would not require States to harden their systems against any possible cyber threat. Nor would it require States to guard against any "hypothetical future cyber operations". Instead, as the Court stated in *Pulp Mills* and *Costa Rica*, States would have a procedural due diligence obligation that would be triggered once the State embarks on any activity "having the potential adversely to affect the [rights and interests] of another State" to "ascertain if there is a risk of significant transboundary harm".[94] Specifically, in such circumstances, States would be required to conduct an "impact assessment" to determine if the State's actions in cyberspace would have the potential to adversely affect the rights and interests of another State.

This "impact assessment" could come in a variety of forms and would be circumscribed in "light of the specific circumstances of each case".[95] For example, as the Court noted in *Pulp Mills*, it would be for "each State to determine in its domestic legislation" the specific content of the impact assessment required in each case, having regard to "the nature and magnitude" of the proposed activity and its likely adverse impact on the

91 Schmitt, note 1, at 11.
92 *Id*. at 45.
93 *Ibid*.
94 *Costa Rica*, note 18, at 706-07.
95 *Id*. at 707.

rights and interests of other States, "as well as to the need to exercise due diligence in conducting such an assessment".[96] Further, as the Court stated in *Costa Rica*, if the impact assessment "confirms that there is a risk of significant transboundary harm, the State planning to undertake the activity is required, in conformity with its due diligence obligation, to notify and consult in good faith with the potentially affected State, where that is necessary to determine the appropriate measures to prevent or mitigate that risk".[97]

Adopting the preventive / precautionary approach in cyberspace would therefore introduce a procedural due diligence obligation on States, and would impose two distinct obligations on States. First, if the State plans to engage in activity having the potential adversely to affect the rights and interests of another State, the State would undertake a cyber impact assessment to ascertain if there is a risk of significant transboundary harm resultant from that action. Second, if the impact assessment confirms there is a risk of significant transboundary harm, the State planning to undertake the activity is required, in conformity with its due diligence obligation, to notify and consult in good faith with the potentially affected State, where that is necessary to determine the appropriate measures to prevent or mitigate that risk.

Adopting this obligation is not impossible for States, as many already implement the due diligence principle in many of their cyber strategies and domestic plans. In its 2011 International Strategy for Cyberspace, for example, the United States stated that "States should recognize and act on their responsibility to protect information infrastructures and secure national systems from damage or misuse".[98] Germany likewise adopted a due diligence approach in many of its national programs and strategies.[99] Similar jurisdictions have due diligence principles built into their programmes, including new data protection regulations in the European Union.

Adopting a procedural due diligence approach in cyberspace would also be consistent with international law. States are already bound to conduct their international relations with other States in "good faith,"[100] which has been defined as a sustained upkeep of negotiations over a period appropriate to the circumstances and with an awareness of the interests of the other party.[101] States could apply this principle when determining whether to enter into negotiations with other States regarding the results of their impact

96 *Pulp Mills*, note 11, at 83.
97 *Costa Rica*, note 18, at 707.
98 International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World, White House 10 (2011).
99 Annegret Bendiek, *Due Diligence in Cyberspace: Guidelines for International and European Cyber Policy and Cybersecurity Policy*, SWP Research Paper at 22 (2016).
100 Rogoff, *The Obligation to Negotiate in International Law: Rules and Realities*, 16 Mich, J. Int'l L. 141, 153 (1994).
101 *Application of the International Convention on the Elimination of All Forms of Racial Discrimination (Georgia v. Russian Federation)*, 2011 I.C.J. 157 (2011); *Arbitration between Kuwait and the American Independent Oil Co., (AMINOIL)* 21 ILM 1982, 1014; *Lac Lanoux Arbitration (France v. Spain)* (1957) 24 I. L. R. 101, 23 November 16, 1957.

assessment, and whether certain systems should be hardened, or further information should be exchanged.

Adopting a procedural due diligence approach in cyberspace would also address the underlying concern addressed in international environmental law – the prevention of significant and non-reversible transboundary harm. Over the past ten years alone, from the 2007 attack in Estonia to the 2016 attack in the United States, the scope and impact of detrimental cyber operations has been manifest. The precautionary principle would require an impact assessment be conducted, even if technical certainty is not conclusive to prevent transboundary harm. In this context, applying the precautionary approach in cyberspace would not, as the IGE supposes in the *Tallinn Manual 2.0*, place an unreasonable burden on States, because the obligation would not require the State to harden systems *per se* but only to conduct a procedural review to determine if there is a threat of significant harm to another State. Thus, under the formulation set out by the ICJ in the *Pulp Mills* and *Costa Rica* cases, the precautionary approach in cyberspace could blend procedural and substantive elements.

From a substantive perspective, it could be agreed between States that, as a general rule, States must take steps to mitigate any potential transboundary harm resultant from potential cyber operations using that territorial State's cyber infrastructure, even if there is no conclusive evidence of attribution, technical identification, or operational certainty. To effectuate this substantive obligation, as in the environmental context, a procedural obligation would be required by States that would place a lesser burden on them. This requirement would not, as the 2017 IGE suggests, require a State to anticipate every hypothetical attack. It would instead allow the territorial State to understand the current state of its national cyber infrastructure, to measure that against known threats within and outside its jurisdiction, and to make a determination as to whether it should consult with other States based on a threat analysis commensurate with the experience and resources of the territorial State. As the ICJ noted in *Costa Rica*, the scope and substance of such an assessment would be subject to the circumstances of each State.

Of course, there are certain guideposts that could be established by treaty that would outline the scope of any such impact assessment. States could agree, for example, that when triggered a general framework for review should be used similar to that provided in the National Institute of Standards and Technology's (NIST) *Framework for Improving Critical Infrastructure Cybersecurity*.[102] This uses a common language to address and manage cybersecurity risk for private business, focusing on a risk management framework. Many private-sector entities understand that the standard for private sector "due diligence" is compliance with the NIST Framework[103] and several

102 National Institute of Standards and Technology, *Framework for Improving Critical Infrastructure Cybersecurity* (2014) (hereinafter, "NIST Framework").

103 Why the NIST Cybersecurity Framework Isn't Really Voluntary, Info. Sec. Blog (Feb. 25, 2014), http://www.pivotpointsecurity.com/risky-business/nist-cybersecurity-framework.

States are engaged in NIST collaborations, including the United Kingdom, Japan, Korea, Estonia, Israel, and Germany.[104] In any event, this paper does not endorse any particular method of impact assessment, only that once triggered, States should agree that conducting an impact assessment is a procedural due diligence requirement.

By segregating procedural and substantive due process, the concern raised by the IGE in *Tallinn Manual 2.0* that the due diligence principle is difficult to effectuate in cyber space because of the "difficulty of mounting comprehensive and effective defences against all possible cyber threats"[105] would be mitigated. States would not have to mount comprehensive and effective defenses against all possible cyber threats. Territorial States would instead only need to conduct a procedural due diligence impact assessment, if triggered. In *Pulp Mills*, the ICJ noted that the scope and substance of EIAs would be dependent on the specific "nature and magnitude of the proposed development and its likely adverse impact on the environment".[106] Likewise in cyber, the scope and nature of an impact assessment would be dependent on the nature and magnitude of the particular cyber infrastructure in question. For example, an impact assessment conducted by the United States or China would be significantly more complex than that conducted of lesser cyber capable States. The standards could be flexible. But the underlying principle should be clear.

By adopting the precautionary approach, as reflected in the ICJ's jurisprudence, States would have a clear obligation that would help better crystallize the substantive due diligence obligation that has evaded State interest to date.

6. CONCLUSION

The IGE recognized in the *Tallinn Manual 2.0* that "in light of the nature of cyber activities, preventive measures are arguably prudent".[107] Applying the precautionary approach to the due diligence principle in cyberspace would help to crystallize the principle of due diligence, and encourage increased adherence, by implementing a prudent and understandable procedural obligation. The precautionary principle in cyberspace is, of course, not reflective of customary international law. This paper argues that instead the approach is the *lex ferenda*, or where the law should go. The benefits of the precautionary approach, especially delineating between procedural and substantive due diligence, would have clear benefits in cyberspace by providing more clear guideposts for States on what is required when carrying out due diligence. By requiring States to undergo critical assessments of their cyber infrastructure to

104 See Brian Fung, *A Court Just Made It Easier for the Government to Sue Companies for Getting Hacked*, Wash. Post (Aug. 24, 2015), https://www.washingtonpost.com/news/the-switch/wp/2015/08/24/a-court-just-made-it-easier-for-the-government-to-sue-companies-for-getting-hacked/?wpmm=1&wpisrc=nl_headlines.

105 Tallinn Manual 2.0 at p. 45, Rule 6, ¶ 8.

106 *Pulp Mills*, note 11, at ¶205.

107 Tallinn Manual 2.0, note 4, at 46 (Rule 7).

determine potential vulnerabilities, the precautionary approach would create a baseline obligation for States that could help to crystallize the due diligence principle in cyberspace, and help move this grey zone of international law to a principle of customary international law.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Pressing Pause: A New Approach for International Cybersecurity Norm Development

Cedric Sabbah
Office of the Deputy Attorney General (International Law)
Ministry of Justice
Israel[1]

Abstract: Over the last few years, the international community has devoted much attention to the topic of "international cyber norms". However, there appears to be a fundamental tension between these norm-development efforts and their real-world application as effective tools to reduce cyber risk and deter or prevent malicious state and non-state actors. Furthermore, in the current geopolitical climate, a broad agreement on global cyber norms seems improbable, as suggested by the lack of consensus in the course of the UN GGE 2017 process.

In the meantime, government officials tasked with developing and deploying cybersecurity policy and law face day-to-day challenges and are operating on a different track. Questions continuously arise with respect to the role of the state in formulating cybersecurity standards, information sharing, active defense and privacy protection. These questions are dealt with mostly in the "civilian" cybersecurity sphere and are occurring largely under the radar of the global "international cyber norms" community.

Against this backdrop, the paper suggests a shift in the approach to cyber norms. Its central thesis is that, at this juncture, rather than attempting to create a set of pre-defined aspirational norms aimed at achieving global stability, the international community should pay greater attention to discussions that are already occurring between cybersecurity regulators/authorities and should proactively support such discussions. Incremental and "bottom-up" processes, covering technical, policy and legal challenges at the domestic level, create fertile grounds for discussions that

1 The views and opinions stated herein belong to the author only, and are not reflective of Israel's Ministry of Justice or the Israeli government.

can be scaled up. This civilian, bottom-up approach is admittedly more mundane than the "aspirational cyber norms" track. Both tracks can and should continue to coexist in parallel, though the "civilian" track is more likely to result in a common taxonomy, legal/policy interoperability or common understandings that states can readily endorse, all of which could potentially ultimately lead to norms that enhance cybersecurity more pragmatically.

Keywords: *cyber norms, international law, cybersecurity law*

1. INTRODUCTION

The subject of "cyber norms" has been discussed at length in recent years, especially following the report on the subject issued in 2015 by a United Nations Governmental Group of Experts (GGE), regarding the use of information and communications technologies (ICT) by states.[2] Building upon the 2013 GGE Report,[3] the 2015 GGE Report acknowledged the application of basic concepts of international law, such as self-defense and state responsibility for internationally wrongful acts, to the cyber domain. It also recommended a series of "voluntary, non-binding norms" applicable in peacetime, which according to the Report were intended to reflect the international community's expectations as to "responsible behavior by states" in order to "increase stability and security in the global ICT environment."[4] The suggested norms covered a range of topics, from information sharing between states, to providing assistance to other states in dealing with cyber incidents, to protection of critical infrastructure.[5] The report was considered a significant development because representatives of 20 countries holding widely divergent views had produced a consensus text on certain topics that had previously been considered highly contentious. Another GGE was convened in 2016, with a mandate to expand on the 2015 GGE Report.[6] However, amid reports of profound rifts among the participating countries,[7] this GGE ended its work in 2017 without a consensus text being issued. Despite this setback, the subject

2 Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UN Doc. A/70/174 (22 July 2015) ("2015 GGE Report").

3 Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UN Doc. A/68/98 (24 June 2013).

4 2015 GGE Report, para. 9 and 10.

5 Id., par. (c), (f) and (h).

6 UNGA Resolution A/RES/70/237 (23 December 2015).

7 Michele Markoff, US Expert to the GGE, Explanation of Position at the Conclusion of the 2016-2017 UN Group of Governmental Experts (GGE) on Developments in the Field of Information and Telecommunications in the Context of International Security, June 23, 2017, <https://usun.state.gov/remarks/7880>. See also Arun M. Sukumar, Lawfare Blog, Tuesday, July 4, 2017 <https://lawfareblog.com/un-gge-failed-international-law-cyberspace-doomed-well>.

of "cyber norms" continues to draw attention, with some arguing that states should expand this exercise.[8]

The working assumption in this discussion, it seems, is that norms are inherently a good thing: broadly defined as "shared expectations about appropriate (or inappropriate) behavior within a given community",[9] they can lay down the "rules of the road" between states, and thus contribute to international stability.[10] This has generated a wide range of proposals and ideas in an effort to identify the "right" forum in which a discussion can be held[11] or the "right" norm that states can settle on,[12] and to devise ways in which to implement the 2015 GGE norms.[13]

To be sure, the general notion that norms might eventually play a positive role in stabilizing cyberspace remains relevant, and the work of the GGE processes has arguably advanced the global conversation.[14] However, these approaches have not yielded concrete results beyond the 2015 GGE Report. Finnemore and Hollis refer to "fatigue" from the multiplicity of projects in this field.[15]

Against this backdrop, this paper argues that a moderate shift in approach is called for, beginning with a reassessment of current norm-development efforts and their underlying premises. The first part presents a critique of cyber norms and the global community's expectations of them. It argues that given the present political context and divergences between the main players, the focus on "global stability" – arguably, the underlying theme of the 2015 GGE Report – is, at this point in time, overly ambitious, and that norm-development efforts should be untethered from this goal. The second part proposes to shift the emphasis, from "global stability" to domestic cybersecurity. Its central thesis is that, rather than the current top-down approach that

8 See for example Kubo Mačák. (2017). From Cyber Norms to Cyber Rules: Re-engaging States as Law-makers. Leiden Journal of International Law, 30(4), 877-899. doi:10.1017/S0922156517000358.

9 This paper adopts on the definition used by Duncan B. Hollis in his article, "China and the US Strategic Construction of Cybernorms: The Process Is the Product". Hoover Institute, Aegis Paper Series No. 1704, July 6, 2017, <https://www.hoover.org/research/china-and-us-strategic-construction-cybernorms-process-product>, at p. 1.

10 See for example UK National Cyber Security Strategy 2016-2021, para.6.3.3; Australia Cyber Security Strategy, p. 42, which emphasize this point.

11 See James A. Lewis, "Revitalizing Progress in International Negotiations on Cyber Security", in Centre for International Governance Innovation (CIGI), *Getting beyond Norms: New Approaches to International Cyber Security Challenges*, edited by Fen Osler Hampsen and Michael Sulmeyer, Sept. 5, 2017, pp. 13-18; Joseph S. Nye Jr., "Normative Constraints on Cyber Weapons", in Getting Beyond Norms, ibid., pp. 19-22.

12 For example, Tim Maurer, Ariel (Eli) Levite, George Perkovich, "Toward a Global Norm Against Manipulating the Integrity of Financial Data", White Paper, Carnegie Endowment for International Peace, March 27, 2017.

13 E.g. East-West Institute, "Promoting International Cyber Norms: A New Advocacy Forum", Dec. 2015; ICT4Peace open consultations on the United Nations Cybersecurity Norms Proposals, <https://ict4peace.org/call-for-global-open-consultations-on-the-united-nations-cybersecurity-norms-proposal/> (accessed on March 11, 2018); Mariarosaria Taddeo, "Deterrence by Norms to Stop Interstate Cyber Attacks", Minds & Machines (2017) 27:387–392, 390.

14 Eneken Tikk and Mika Kerttunen, "The Alleged Demise of the UN GGE: An Autopsy and Eulogy", Cyber Policy Institute, 2017.

15 Martha Finnemore and Duncan B. Hollis, Constructing Norms for Global, Cybersecurity, 110 AM. J. INT'L L. 425, 469 (2016).

has characterized norm-development efforts to date, the cybersecurity community would be better served by focusing more on bottom-up processes emanating from cybersecurity policies as they are developed and deployed domestically. It contains a non-exhaustive overview of topics and issues that pose concrete challenges in this sphere. It argues that, while some of these topics are already the subject of bilateral and multilateral conversations to a certain extent, they could benefit from more expanded regional and multilateral conversations. A broad roadmap for taking the discussion forward is then submitted.

Most critically, the approach suggested herein is not focused on a specific set of norms around which to center a global process, but on issues-based discussions between government officials tasked with developing and implementing cybersecurity policy and law at the domestic level. There is no predictable outcome for such an exercise – it may or may not produce guidelines, common understandings or norms, and the outcomes might be global or between like-minded countries only. Neither does this approach negate the importance of maintaining existing multilateral cyber norm diplomatic efforts. However, the paper argues that, short of achieving "global stability", as current norms processes set out to do, such a bottom-up, needs-driven approach can help enhance cybersecurity for the parties involved in a concrete way.

2. A CRITIQUE OF CYBER NORMS

A. Advantage of Cyber Norms

Cyber norms have undeniable political and policy advantages for states. As defined in the 2015 GGE Report, norms differ from international law rules in that they are not binding on states. As such, they provide a certain flexibility, allowing states to coalesce around a particular principle or value without compromising their official legal positions. In the case of the 2015 GGE, this may have enabled the United States, the United Kingdom, Germany, China and Russia – countries with profoundly different approaches to the application of international law to cyberspace and what "information security" means – to agree on a set of broad principles.[16]

Another argument in favor of cyber norms, for states, is signaling or deterrence. By expressing support for or adherence to a certain norm, states are putatively indicating to each other that they would treat the violation of such a norm as non-trivial. The 2015 GGE Report makes this goal explicit: "norms reflect the international community's expectations, set standards for responsible State behaviour, and allow the international community to assess the activities and intentions of States".[17] Cyber norms can

16 Finnemore & Hollis, n. 15, p. 470.
17 GGE Report 2015, para. 10.

indicate red lines, providing states with a justification to respond, for example through diplomacy or trade sanctions, when the line is crossed.[18]

The process by which norms are developed can also be seen a positive element. The very fact that governments are speaking with one another, voicing their disagreements and attempting to hash out a consensus, allows the discussion to move forward. The process provides an outlet for states that hold opposing positions to interact with each other and seek common ground. Even if the process does not necessarily generate concrete results, it does foster dialogue between countries, which ultimately is a stepping stone towards global stability. To paraphrase Finnemore and Hollis, the process is the product.[19]

These arguments are valid and sound. However, they should be weighed against the challenges, disadvantages and costs of current cyber norm development efforts.

B. Critical Perspective on Current Cyber Norm Development Efforts

1) Political Challenges

The question of how to achieve global stability in the use of ICTs is an intrinsically political one. The lack of consensus at the 2017 GGE regarding the applicability of international law to the use of ICTs, including specifically the availability of self-defense - despite statements to that effect in previous GGE reports[20] – underscores the ideological and political gaps that remain between the positions of the US and European states on the one hand, and Russia and China on the other.[21] These gaps have been further highlighted in recent months, as China and Russia have each enacted laws tightening controls on Internet access.[22] In parallel, Russia has been actively promoting a new "cybercrime" treaty[23] which adopts an approach to ICTs that is fundamentally different from that found in the Council of Europe's Cybercrime Convention.[24] It is unlikely that these gaps will be resolved in the short term via another iteration of the GGE process or some variant thereof.

18 See for example EU Draft Council Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities ("Cyber Diplomacy Toolbox"), June 7, 2017.

19 Finnemore and Hollis, n. 15, p. 453.

20 2015 GGE Report, par. 28(d) and (e); Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UN Doc. A/68/98, 24 June 2013), para. 19.

21 United Nations, General Assembly, Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General, A/69/723 (13 January 2015).

22 Samm Sacks, "China's Cybersecurity Law Takes Effect: What to Expect", Lawfare Blog, June 1, 2017, <https://www.lawfareblog.com/chinas-cybersecurity-law-takes-effect-what-expect>; Janet Burns, "Russian Laws Will Ban VPNs And Force Chat Users To Register, Giving Censors An Edge", Forbes, July 30.2017, <https://www.forbes.com/sites/janetwburns/2017/07/30/new-russian-laws-ban-vpns-and-force-chat-users-to-register-giving-censors-an-edge/#637dd7d02d7e>.

23 David Ignatius, "Russia is pushing to control cyberspace. We should all be worried", Washington Post, Oct. 24, 2017 <https://www.washingtonpost.com/opinions/global-opinions/russia-is-pushing-to-control-cyberspace-we-should-all-be-worried/2017/10/24/7014bcc6-b8f1-11e7-be94-fabb0f1e9ffb_story.html?utm_term=.30f8621ccc5c>.

24 Council of Europe, Convention on Cybercrime, ETS 185 (2001).

Furthermore, one cannot dissociate the cyber norms debate from the broader geopolitics at play. For example, the United States' qualification of the Sony attacks and of Russia's alleged interference in the 2017 US elections was couched in terms of core principles and values such as free speech and civil liberties.[25] US interests in those cases extended beyond questions of how ICTs are used, and touched on broader questions of interference in another state's internal affairs. Similarly, in a briefing regarding the United States' attribution of WannaCry to North Korea, Tom Bossert, then-current Assistant to the President for Homeland Security and Counterterrorism, made a connection between North Korea's behavior in its use of the ransomware and its nuclear missile program.[26] The difficult topics that successive GGEs wrestled with cannot be analyzed solely from a perspective of information and communication technologies – they are intrinsically tied to a complex web of national interests and alliances, national and international security, international trade and diplomacy.

Finally, the norms discussion is occurring against the backdrop of a broader debate on the future of Internet governance. As is often recalled, the International Telecommunications Union (ITU) has been an unfortunate battleground for this debate, and it remains so to date.[27] The question of whether the Internet can or should be "regulated" in any way at the ITU – a dicey question in itself – has become intertwined with questions of sovereignty "over" the Internet,[28] further complicating the norms debate.

There are good arguments to be made that, notwithstanding the above, agreement on core "global stability" issues is desirable and could conceivably be achieved. Some of the proposals advanced recently include protecting the integrity of financial data,[29] dealing with "states' responsibility arising from the actions of their citizens," a commitment to ensure that actions in cyberspace do not contravene their international commitments, treatment of election processes as protected infrastructure and norms for cybercrime.[30] While it may be possible to achieve a consensus around these types of issues in the medium or long term, the doubts raised in this paper relate to whether

25 White House Office of the Press Secretary, Statement by the Press Secretary on the Executive Order Entitled "Imposing Additional Sanctions with Respect to North Korea", January 2, 2015, <https://obamawhitehouse.archives.gov/the-press-office/2015/01/02/statement-press-secretary-executive-order-entitled-imposing-additional-s>; White House Office of the Press Secretary, Statement by the President on Actions in Response to Russian Malicious Cyber Activity and Harassment, December 29, 2016 <https://obamawhitehouse.archives.gov/the-press-office/2016/12/29/statement-president-actions-response-russian-malicious-cyber-activity>.

26 White House Press Briefing transcript, Dec. 19, 2017 <https://www.whitehouse.gov/briefings-statements/press-briefing-on-the-attribution-of-the-wannacry-malware-attack-to-north-korea-121917/>.

27 Samantha Dickinson, "How 'Cyber' Sidelined 'Development' at the ITU's World Telecommunication Development Conference", CFR Blog, Nov. 17, 2017 <https://www.cfr.org/blog/new-cyber-brief-countering-russian-information-operations-age-social-media>.

28 Paul Rosenzweig, "The Continuing Struggle for Control of Cyberspace--and the Deterioration of Western Influence", Lawfare, Jan. 13, 2014 <https://www.lawfareblog.com/continuing-struggle-control-cyberspace-and-deterioration-western-influence>.

29 Maurer, Levite, and Perkovich, n. 12.

30 *Getting beyond Norms*, n. 9, pp. 16 and 21.

such agreements could emerge as a result of a self-styled norms process, and whether this approach is appropriate for the near future.[31]

2) Practical Limitations

Several factors limit the practical utility of norms. For one, the purported effect of a particular cyber norm cannot be gaged with certainty, since cyber operations are not usually made public. Second, since the GGE norms of 2015 and subsequent reiterations of those norms by the G7 in 2016 and 2017,[32] the world has seen several cyber incidents attributed to nation-states. Public testimony given by the US Director of National Intelligence to a Senate committee in May 2017 attests to the magnitude of cyber threats by states.[33] Indeed, major incidents at least partially attributed to states, like WannaCry, NotPetya, the DNC hack, and election hacks in France,[34] occurred after the adoption of the 2015 GGE norms. Of course, since this list only represents attacks that have been reported, definitive conclusions cannot be drawn from these and similar data. And certainly, the occurrence of these incidents should not be attributed to a "failure" of the norms process. What is evident, however, is that these kinds of incidents illustrate the challenge of applying broad aspirational cyber norms to actual scenarios.

States are also developing their doctrines and strategies at their own cautious pace, based on actual operational needs and existing legal frameworks. The merits of making their conclusions more transparent can be debated, but the national defense and security community is currently on a somewhat slower and more prudent track than the one reflected in current efforts to promote cyber norms.[35] To the extent that a given norm might impact national defense/security interests, the more conservative approach of governmental departments and agencies entrusted with these interests must be acknowledged.

The broader issue here is not whether a particular cyber norm is in fact being implemented. It is that declaring the existence of a norm at a UN forum or similar forum does not guarantee its effectiveness. Norms may provide guidance and declare red lines, but when a country's core interests are at stake, norms arguably play a lesser role. As Tikk and Kerttunen noted,

31 White House, Fact Sheet: President Xi Jinping's State Visit to the United States, 25 September 2015,<obamawhitehouse.archives.gov/the-press-office/2015/09/25/fact-sheet-president-xi-jinpings-state-visit-united-states>.

32 G7 Principles on Actions in Cyber, May 27, 2016, <http://www.mofa.go.jp/files/000160279.pdf>; G7 Declaration on Responsible States Behavior on Cyberspace Lucca, 11 April, 2017, available at <www.mofa.go.jp/files/000246367.pdf>.

33 Daniel R. Coats, Director of National Intelligence, Statement for the Record, "Worldwide Threat Assessment of the U.S. Intelligence Community," Senate Armed Services Committee, 23 May 2017.

34 See full list at CSIS website, <https://www.csis.org/programs/cybersecurity-and-warfare/technology-policy-program/other-projects-cybersecurity>.

35 Max Smeets, "Europe Slowly Starts to Talk Openly About Offensive Cyber Operations", CFR Blog, Nov. 6, 2017 <https://www.cfr.org/blog/europe-slowly-starts-talk-openly-about-offensive-cyber-operations>; Robert Hackett, "Gasp! China admits to having cyber warriors", Forbes, Mar.26, 2015 <http://fortune.com/2015/03/26/china-admits-cyber-warriors/>.

> "[given] the premature understanding what cyber security is about and how it can or may affect international peace and security, it is hard to see how the necessary level of peer pressure can manifest between 193 actors with (justifiably) sovereign interests and authority."[36]

One notable case study in the norm-development process is the norm prohibiting cyber industrial theft, which was excluded from the 2015 GGE Report. It was embodied in a bilateral commitment between China and the United States in 2015,[37] after which it was replicated in other international texts.[38] There have been conflicting reports as to the extent to which China has actually adhered to that commitment.[39] If reports of a partial reduction in cyber industrial theft are accurate, they reinforce the point made above, that at present bilateral commitments based on reciprocal interests are more likely to be effective than multilateral ones. The replication of this particular norm, specifically in bilateral commitments between China and other countries, also suggests that it emerged from a concrete need of states to address a specific concern (theft of intellectual property by companies), as opposed to a broad attempt to promote international stability. Other bilateral agreements based on a pragmatic need to resolve specific issues might also work in similar fashion.[40]

3) Taxonomy and the Ambiguous Value of Constructive Ambiguity

Joseph Nye has shown that the international cyber domain is a "regime complex", composed of a multiplicity of sub-regimes (incident response, law enforcement, international standards, international law, etc.), each with its own set of frameworks and actors.[41] The discussion on cyber norms can be confusing because different states frame the issue differently. Among Western states, cybersecurity, cybercrime, and the applicability of the laws of armed conflict to the cyber domain are distinct (though related) concepts, each governed by its own legal or political regime. By contrast, the concept of "information security" as understood by Russia and China is significantly different.[42]

36 Tikk and Kerttunen, n. 12, p. 26.

37 White House, Fact Sheet: President Xi Jinping's State Visit to the United States, 25 September 2015,<obamawhitehouse.archives.gov/the-press-office/2015/09/25/fact-sheet-president-xi-jinpings-state-visit-united-states>.

38 G7 Declaration on Responsible States Behavior on Cyberspace Lucca, n.32 para.12.; G20 Leaders Communiqué, Antalya Summit, 15-16 November 2015, para. 26 <https://www.g20.org/profiles/g20/modules/custom/g20_beverly/img/timeline/Turquia/2015-g20-final-declaration-eng.pdf>; Reuters, "China, Canada vow not to conduct cyber attacks on private sector", June 26, 2017, <https://www.reuters.com/article/us-canada-china-cyber/china-canada-vow-not-to-conduct-cyber-attacks-on-private-sector-idUSKBN19H06A>.

39 Andy Greenberg, "China Tests the Limits of its Us Hacking Truce", in Washington Post, Oct. 31, 2017, <https://www.wired.com/story/china-tests-limits-of-us-hacking-truce/>; David Sanger, "Chinese Curb Cyberattacks on U.S. Interests, Report Finds", New York Times, June 20, 2016, <https://www.nytimes.com/2016/06/21/us/politics/china-us-cyber-spying.html>.

40 See, for example, Jack Goldsmith, "Contrarian Thoughts on Russia and the Presidential Election", Lawfare Blog, Jan. 10, 2017, <https://www.lawfareblog.com/contrarian-thoughts-russia-and-presidential-election>.

41 Nye, Joseph S. 2014. The Regime Complex for Managing Global, Cyber Activities. Global Commission on Internet Governance, Paper Series, 1.

42 UNGA, n. 21.

The 2015 GGE Report attempted to bridge this divergence of views through vaguely-drafted norms. For example, the 2015 Report includes a norm against attacking a country's "critical infrastructure" contrary to international law but provides no workable definitions or guidelines.[43] This is also the case with the norms regarding "due diligence", supply chain oversight and reporting of vulnerabilities.[44] One may argue that this type of constructive ambiguity is helpful in that it conveys an intelligible concept that states are free to define going forward.[45] One may also point to the current norm-development forums as positive efforts to infuse content to these norms. These arguments are certainly persuasive. However, the fundamental difficulty with this type of top-down push for achieving consensus is that it places the carriage before the horse: it glosses over the constructs around which the norms are built, declares a particular norm into existence, and only then seeks a way to operationalize it. This approach is not conducive to widespread implementation by states.

Indeed, events are unfolding at a rapid pace, challenging a short or mid-term conception of what a "stable ICT environment" might look like. The domestic policy landscape is continuously evolving: for example, it has recently been reported that Germany is actively exploring the possibility of enacting legal authority for state "hackbacks",[46] while China has adopted a sweeping cybersecurity law.[47] Moreover, the use of cyber tools by diverse actors – state, non-state, hacktivist groups and individuals – continues to rise, presenting new practical and legal challenges to states.[48] In short, it is difficult to deal with long-term stability through cyber norms, when the short and medium-term reality are filled with moving targets.

In summary, it is not argued that there is no room for a discussion on cyber norms involving core "global stability" issues. However, there is another, potentially more fertile ground for discourse in the field of domestic, "civilian" cybersecurity (defined below). Given the above factors, a more promising approach to cyber norms would be to promote and expand existing discussions in the domestic civilian sphere and allow norms within that sphere to emerge and evolve in a more organic fashion. The next part proposes a multi-stage analysis for how such a process might take place.

43 2015 GGE Report, para.13(f).
44 Ibid., para. 13(b), (h), (i), (j).
45 See discussion on "incompletely theorized" norms in Finnemore & Hollis, n. 15, p. 21.
46 Andrea Shalal, "German spy agencies want right to destroy stolen data and 'hack back' ", Reuters, Oct.5 2017, <https://www.reuters.com/article/us-germany-cyber/german-spy-agencies-want-right-to-destroy-stolen-data-and-hack-back-idUSKBN1CA1IN>.
47 Sachs, n. 22.
48 Paul Rosenzweig, "The Reality of Cyber Conflict: Warfare in the Modern Age", Heritage Foundation, 2017, <http://index.heritage.org/military/2017/essays/reality-cyber-conflict/>.

3. REFRAMING THE GLOBAL DISCUSSION ON CYBER NORMS: A POSSIBLE PATH FORWARD

The stated purpose of the cyber norms in the 2015 GGE Report was to "help to prevent conflict in the ICT environment and contribute to its peaceful use." Those objectives were ambitious, to say the least, and the current state of play suggests that the goal of global stability may be too much to pin on cyber norms.

Rather than attempting to tackle large, controversial issues that are fraught with political baggage, it may be more useful to enhance and broaden existing discussions around more mundane – yet no less important – issues of cybersecurity policy and regulation in the domestic, civilian sphere. Put otherwise, rather than asking "which cyber norms can enhance global stability in the cyber domain?", it is worth asking "what issues do cybersecurity officials have in the domestic arena, that could benefit from a broader conversation with their counterparts around the world?" As one commentator noted:

> "Given these near-dead ends, real issues might best be taken up bilaterally or multilaterally between countries and entities that have mutually agreed priorities and issues. Given political sensitivities, technical-level cooperation – be it between computer emergency response teams, law enforcement entities or judicial authorities – is likely more efficient than politicized formats."[49]

This admittedly unassuming starting point will not in and of itself produce world peace. However, if cybersecurity professionals engage in greater discussions of the type described below, this could help the international community or coalitions of like-minded countries to achieve a few discrete objectives in the field of domestic policy and law. This might contribute to greater security in the cyber domain, which could in turn enhance global stability over time. The approach proposed below is not intended to replace or subsume current large-scale "global stability" norm development efforts. Rather, it is a parallel track, which at this juncture should be afforded greater attention.

A. Framing the Discussion: Cybersecurity in the Civilian Sphere

Since the 1980s and 1990s, the body of policies and laws for protecting critical networks has matured into a full-fledged discipline. States are beginning to develop and update comprehensive cybersecurity strategies,[50] and are being increasingly active in the legislative sphere, as exemplified by the US Cybersecurity Information Sharing Act of 2015 and the EU NIS Directive. Furthermore, cybersecurity has percolated into the

49 Eneken Tikk, "Norms à la Carte", in Getting Beyond Norms, n. 9, p. 25.
50 See n. 10.

spectrum of regulatory issues, with regulators in the financial sector,[51] energy,[52] and transportation,[53] for example, developing sector-specific cybersecurity policies and rules. In the private sector as well, insurance companies, accounting firms, law firms and consulting firms have begun offering services in cybersecurity to their clients.[54]

For the most part, the topics covered by these areas do not involve complex questions of international law or international relations. They are mainly focused on building up robustness (sharing information about threat indicators, regulatory incentives for the private sector to improve defense, cyber awareness campaigns, supply chain oversight, etc.), and resilience (breach incident notification requirements, intervention of the national CERT, etc.), at the domestic level.[55] By way of illustration, on the domestic "civilian" end are topics such as how to protect personally identifiable information as part of an organization's information sharing with the government, application of the NIST framework to private entities, regulation of cybersecurity professionals, breach incident disclosure requirements in consumer protection law and securities law, cybersecurity regulation on the cloud, active defense in the private sector, and labeling requirements for software. The processes for policy development in these areas are usually unclassified and involve open consultations with the private sector. Similarly, these measures operate mainly in the civilian sphere, and they aim to promote domestic cybersecurity in the narrow sense of the term – reducing the risk of cyber incidents and the damages caused when such incidents occur.

At the other end of the spectrum are measures regarding the interface with the attacker or associated actors in the international sphere, for example deterrence tools, permitted actions above or below the "use of force" threshold under Article 2(4) of the UN Charter, the proposed norm about refraining from manipulating financial data, and broad questions of sovereignty and jurisdiction. Such topics are inherently more

51 Financial Services Board, Stocktake of Publicly Released Cybersecurity Regulations, Guidance and Supervisory Practices, Oct. 13, 2017, <http://www.fsb.org/wp-content/uploads/P131017-2.pdf>, Tom Gilheany, "The State of Cybersecurity Laws in the Financial Services Industry", in Talking Tech With Cisco Blog, May 18, 2017 <https://learningnetwork.cisco.com/blogs/talking-tech-with-cisco/2017/05/18/the-state-of-cybersecurity-laws-in-the-financial-services-industry>.

52 Energy Expert Cyber Security Platform, "Cyber Security in the Energy Sector Recommendations for the European Commission on a European Strategic Framework and Potential Future - Legislative Acts for the Energy Sector", February 2017, available at <https://ec.europa.eu/energy/en/news/new-report-cyber-security-energy-sector-published>.

53 For example: UK Government, Department of Transport, "Principles of cyber security for connected and automated vehicles", Aug. 6, 2017 <https://www.gov.uk/government/publications/principles-of-cyber-security-for-connected-and-automated-vehicles/the-key-principles-of-vehicle-cyber-security-for-connected-and-automated-vehicles>.

54 OECD (2017), Enhancing the Role of Insurance in Cyber Risk Management, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264282148-en; Lexis Nexis Business of Law Blog, "Beautiful Minds: 41 Legal Industry Predictions for 2016", Dec.16, 2015 <http://businessoflawblog.com/2015/12/legal-industry-predictions-2016/>.

55 Regarding the distinction between "robustness" and "resilience", see Matania, E. & Yoffe, L. & Mashkautsan, M. "A Three-Layer Framework for a Comprehensive National Cyber-security Strategy." Georgetown Journal of International Affairs, vol. 17 no. 3, 2016, pp. 77-84. Project MUSE, doi:10.1353/gia.2016.0038.

sensitive, approaching the core of a country's national security interests and raising complex international relations and international law questions.

This distinction between "domestic civilian" and "international" realms does not profess to create entrenched categories of cybersecurity policy and law, nor to suggest that any particular area in the cybersecurity discussion belongs exclusively to either realm. It merely highlights that certain areas of policy and law will tend to be easier for states to discuss in an open and transparent manner than others.

It should be stressed that the proposal to focus on the domestic civilian sphere is not meant to exclude the evolution of other norms in the field of defense and security, such as how to apply the law of state responsibility to attacks attributable to non-state actors, what "sovereignty" means,[56] and what "responsible state behavior" could look like in practice. Processes in both these areas can coexist and complement one another. The thrust of the argument here is that the domestic civilian cybersecurity sphere should garner more attention from the international community than it has to date, and may reveal itself to be a promising path forward.

B. Bottom-up Process Led by Domestic Cybersecurity Professionals

Having broadly defined the types of issues that could be discussed, it is equally important to describe the contours of possible discussions around these issues. Civilian cybersecurity discussions are driven by those government officials tasked with creating and deploying domestic policy and law. This includes officials involved with cyber education and awareness, defense of critical and non-critical infrastructure networks, handling of cyber events in real time within a CERT, policy development, engagement with the private sector, regulation and oversight.

Through this dialogue, cybersecurity professionals with diverse backgrounds develop a common language, share issues and questions of concern, learn from best practices, and achieve informal capacity building. The dialogue is technical, legal or policy-oriented or multidisciplinary. This is fundamentally a bottom-up process, which draws from the experience and expertise of cybersecurity professionals.

To be sure, there are already formal and informal discussions under way between different actors around these topics (within FIRST, the network of CERTs including national CERTs, as well as between sector-specific industry regulators). Our suggestion here is to expand upon, and refocus the international community's efforts around, these types of discussions.

56 Gary Corn, Tallinn Manual 2.0, Advancing the Conversation, Just Security (Feb. 15, 2017) <https://www.justsecurity.org/37812/tallinn-manual-2-0-advancing-conversation/>.

The process suggested above can be distinguished from the OSCE's confidence-building measures of 2013 and 2016.[57] Finnemore and Hollis have shown how, among other factors, the choice of a particular type of forum to promote a particular norm can be just as important as the content of the norm.[58] For example, when a proposed norm is developed within an existing organization (in this case, the OSCE), this has an impact on the way the norm is understood and its reach to a particular target audience. The OSCE's confidence-building measures were developed primarily in a top-down fashion, mostly through diplomatic action, and thus far, it does not appear that they have been "adopted" by the national CERT community. By contrast, a bottom-up process focused on "civilian cybersecurity" on the topic of confidence building, would likely result in a more technical set of standards based on the perceived needs of national CERT officials, which could then percolate upwards with the assistance of cyber diplomats.

The COE Cybercrime Convention can be taken as illustrative of the ways in which top-down and bottom-up efforts can converge. On the one hand, the Convention constitutes a relatively successful exercise in international law development in a different though related field. Adopted in 2001, it has been ratified by 56 countries and remains the benchmark text in the field of cybercrime. Thus, one might view the Convention as an example of the success of the "top-down" approach. At the same time, the Convention is an example of how the law developed bottom-up from a concrete specific need, namely, law enforcement cooperation to deal with cross-border cybercrime. The conference of state parties of the Convention constitutes a useful forum which is currently tackling several important issues, such as access to data on the cloud, and is attended by a mix of diplomats and practitioners.

An additional clarification is in order. The suggested focus on "domestic cybersecurity" should not be seen as a negating the need for discussions on "global stability". Similarly, diplomatic efforts should not compete with, or come at the expense of, bottom-up civilian-based technical efforts, or vice versa. On the contrary, these two processes can and should complement each other. However, the point made here is that up until now, bottom-up processes have been largely ignored in the cyber norms discussion.[59] A few concrete examples of how such processes can be amplified and harnessed will be suggested below.

57 Organization for Security and Co-operation in Europe, Decision No. 1106: Initial Set of OSCE Confidence-Building Measures to Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies, PC.DEC/1106, OSCE Permanent Council, 975th Plenary Meeting, 3 December 2013), <http://www.osce.org/pc/109168?download=true>; Decision No. 1202: OSCE Confidence-Building Measures To Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies, PC.DEC/1202, OSCE Permanent Council, 1092nd Plenary Meeting, 10 March 2016, <https://www.osce.org/pc/227281?download=true>.

58 Finnemore and Hollis, n. 15, p. 468.

59 A notable exception is the "CERT diplomacy" initiative raised at the 2017 Internet Governance Forum, which is addressed below.

C. Potential Areas of Discussion

As noted previously, there is no definitive list of cybersecurity topics that can neatly fit into a "civilian" category. Similarly, not every issue is necessarily conducive to broad multilateral discussions. Still, there are areas where common ground, or at least shared understandings, are more realistic. We provide below a few examples of such areas.

1) The Role of the State

The hybrid private-public nature of Internet infrastructure, coupled with the pervasiveness of connected devices, presents new challenges for domestic cybersecurity regulators. One of these is identifying the instances in which a national cybersecurity agency can and should intervene in the market in order to prescribe minimum standards. The need for government cybersecurity officials to manage risk, prioritize and classify types of organizations and networks, balance between rules-based and principles-based regulation making and optimize the use of deterrents and incentives, while maintaining the core authority to intervene when national security or public order or safety are at stake, requires difficult choices, constant engagement with the private sector, and an adaptive modus operandi. While domestic cybersecurity agencies might be developing this approach on their own, there could be much benefit to an expanded discussion on regulatory choices, pitfalls and best practices. The NIST Framework,[60] the OECD Recommendations on Digital Risk Management[61] and the OECD workshop on protecting critical infrastructure[62] provide useful starting points for such discussions.

2) Information Sharing Between the Public and Private Sectors

An underlying issue of concern for cybersecurity regulators is how to generate trust between the public and private sectors within a particular jurisdiction.[63] Relevant questions to be asked include: are current domestic policies and practices in this field optimal? Do they lead to actionable results? How can data collection practices be streamlined? Can and should a common information sharing standard be adopted? What type of approach vis-à-vis the private sector is desirable? In what cases are incentives more appropriate? How can individuals' personal information be protected in the course of information sharing? An expanded dialogue on how to improve

60 NIST, "Cybersecurity Framework," <www.nist.gov/cyberframework>.

61 OECD (2015), Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264245471-en.

62 See workshop website at <http://www.oecd.org/going-digital/digital-security-in-critical-infrastructure/>, accessed on March 11, 2013.

63 See, for example, the discussions held at the 2017 Internet Governance Forum regarding this topic: International cooperation between CERTS: technical diplomacy for cybersecurity (https://igf2017.sched.com/event/CTrn/international-cooperation-between-certs-technical-diplomacy-for-cybersecurity-ws38?iframe=no&w=100%&sidebar=yes&bg=no>); Cybersecurity 2.0 - Leveraging the Multistakeholder Model to Develop and Deploy Cybersecurity Policy (<https://igf2017.sched.com/event/CTri/cybersecurity-20-leveraging-the-multistakeholder-model-to-develop-and-deploy-cybersecurity-policy-of70?iframe=no&w=100%&sidebar=yes&bg=no>).

information sharing between the private and public sectors could lead to real solutions to such dilemmas.

3) Active Defense in the Private Sector

For the purposes of this paper, we define "active defense" as actions and measures taken to:

> "detect, analyse, identify and mitigate threats to and from communications systems and networks in real-time, combined with the capability and resources to take proactive or offensive action against threats and threat entities including action in those entities' home networks".[64]

The issue has been analyzed at length, leading to growing calls for a more sophisticated discussion on active defense in the private sector.[65] Possible policy discussions to be held include whether some of the risks attendant to active defense could be mitigated by adding elements of *ex ante* and *ex post* government oversight and entrusting the task to reputable cybersecurity companies under an accreditation system. Another policy issue is whether the perceived need to allow active defense could be diminished if "internet infrastructure" entities such as ISPs were better incentivized to take a more active role in detecting and mitigating attacks transiting through their networks.

4) Cybersecurity on the Cloud

The UN Commission on International Trade Law (UNCITRAL) has begun grappling with the contractual aspects of cloud services in the private sector,[66] and this topic seems ripe for further study from a cybersecurity perspective, particularly with respect to government procurement of cloud services from third party vendors.[67]

64 Robert Dewar, "The 'Triptych of Cyber Security': A Classification of Active Cyber Defence" (6th Annual Conference on Cyber Conflict, 2014), NATO Cooperative Cyber Defence Centre of Excellence, https://ccdcoe.org/cycon/2014/proceedings/d1r1s9_dewar.pdf.

65 Joe Uchill, "New bill would allow hacking victims to 'hack back' ", *The Hill*, Oct.13, 2017 <http://thehill.com/policy/cybersecurity/355305-hack-back-bill-hits-house>. See also Paul Rosensweig, Steven P. Bucci and David Inserra, "Next Steps for U.S. Cybersecurity in the Trump Administration: Active Cyber Defense", Heritage Foundation Backgrounder 3188, May 5, 2017 <www.heritage.org/sites/default/files/2017-05/BG3188.pdf >.

66 UNCITRAL Working Group IV on e-commerce.

67 American Technology Council, "Report to the President on Federal It Modernization", Dec. 13, 2017 <https://itmodernization.cio.gov/>UK Government Digital Service, "Government Cloud First Policy", Feb. 3, 2017 <https://www.gov.uk/guidance/government-cloud-first-policy>.

Other relevant topics include:

- cyber insurance (whether and how the market should be regulated, guidance on how to quantify cybersecurity risks);
- cybersecurity for the Internet of Things;[68]
- labeling and rating of software;[69]
- developing a common ontology and technical standards for cybersecurity.[70]

At the same time, it should be borne in mind that not all civilian efforts are worth pursuing at a global scale.[71] The challenge is to identify topics that could both benefit from and lend themselves to an international conversation.

D. The Formats of Potential Discussions

The format of an international discussion about a particular area can be as important as the topic itself, as it sets the stage for the types of discussions that are held and the expectations of participants.[72] Accordingly, we offer the following basic principles regarding the format for potential discussions around topics such as the ones discussed above.

1. A non-prescriptive process is more likely to enable participants to engage in an exploratory dialogue in which they consider a range of options. A discussion on norms should be allowed to emerge naturally as a result of the discussions, rather than established as a goal from the outset.
2. As mentioned earlier, the agenda should be set by cybersecurity officials involved with policy development and deployment. They are arguably best placed to define and discuss the challenges they face on a day-to-day basis.
3. The level of engagement (multilateral, regional or like-minded) plays an important role in expectations and outcomes. To state the obvious, the more global the forum, the more challenging it is to achieve consensus.
4. One cannot ignore the place of bilateralism. Several countries have opened lines of dialogue and entered into bilateral agreements and memorandums of understanding in the field of cybersecurity[73] and this trend will likely

68 Laura DeNardis & Mark Raymond, "The Internet of Things as a Global Policy Frontier", UC Davis Law Review, Issue 51:2 (December 2017), 475.

69 E.g. DHS designation of Kaspersky products as presenting security risks - DHS Statement on the Issuance of Binding Operational Directive 17-01, Sept. 13, 2017 <https://www.dhs.gov/news/2017/09/13/dhs-statement-issuance-binding-operational-directive-17-01>; see also Cyber Independent Testing Lab, founded by Sara and Peter Zatko (a.k.a Mudge).

70 Claire Vishik, Mihoko Matsubara, Audrey Plonk, "Key Concepts in Cyber Security: Towards a Common Policy and Technology Context for Cyber Security Norms", in International Cyber Norms Legal, Policy & Industry Perspectives, Anna-Maria Osula and Henry Rõigas (Eds.), NATO CCD COE Publications, Tallinn 2016.

71 Columbia School of International Public Affairs New York Cyber Task Force, "Building a Defensible Cyberspace", Sept. 2017, <https://sipa.columbia.edu/sites/default/files/3668_SIPA%20Defensible%20Cyberspace-WEB.PDF>, p. 14.

72 Finnemore and Hollis, n. 15, p. 468.

73 See, for example, Mapping of India's Cyber Security-Related Bilateral Agreements, <https://cis-india.org/internet-governance/blog/india-cyber-security-bilateral-agreements-map-dec-2016> (accessed on March 11, 2018), Australia Cyber Security Strategy, n. 10, p. 43.

continue in the near future. While the resulting texts may be phrased in broad language that encourages general cooperation rather than requiring compliance with concrete obligations, they create the framework for engagement between states within which future cybersecurity discussions can be held.

5. The creation of yet another global forum dealing with cybersecurity should be avoided. The focus should not be on adding to the high-level discussions that already exist, but on expanding the bottom-up, professional discussions that are currently under-exploited.

One practical way forward was recently explored at the Internet Governance Forum of 2017 in Geneva. There, national and private CERTs were identified as technical and largely apolitical actors at the frontline of incident response. These attributes position CERTs advantageously, as potentially significant actors on the global sphere. To tap into this potential, governments could further empower CERTs to engage with one another, broaden the scope of their discussions and cooperation, and take the lead in "cyber diplomatic efforts".[74] That being said, any expanded role for CERTs should be carefully crafted to avoid unduly politicizing their activities and tainting their technical mission. Another interesting outcome of the 2017 IGF was the proposal, in one of the panels, to leverage the multi-stakeholder model to enhance cybersecurity policy development and deployment.[75] While this panel was primarily focused on domestic cybersecurity, examples were given of how bottom-up domestic policy development processes can have international ripple effects. The NIST Framework was frequently cited as a useful standard for countries and entities outside the United States.

Another example could be to expand the work of technical, policy and legal working groups in bodies such as UNCITRAL and the OECD. These bodies enjoy broad membership with established structures and work methods, and their work is typically produced by subject-matter experts. As noted above, they have each undertaken work that touches on cybersecurity issues in the past, and they could be tasked with more such issues going forward. This requires a "bottom-up" push from cybersecurity officials to suggest clear mandates for working groups within these organizations, followed by a "top-down" push from capitals to promote these mandates when the relevant organization decides on its future work program.

Finally, a more adventurous endeavor could consist of creating one or more *ad hoc* topical and specialized forums, not necessarily tied to existing organizations. For example, one might imagine a forum similar to the Financial Action Task Force

74 A transcript of the session can be accessed at: <https://www.intgovforum.org/multilingual/content/igf-2017-day-3-room-xi-ws38-international-cooperation-between-certs-ws38-technical-diplomacy>. See summary here: <https://www.intgovforum.org/multilingual/index.php?q=filedepot_download/5902/858>.

75 A transcript of the session can be accessed at <https://www.intgovforum.org/multilingual/content/igf-2017-day-3-room-ix-of70-cybersecurity-20-leveraging-the-multistakeholder-model-to>. See summary here: <https://www.intgovforum.org/multilingual/index.php?q=filedepot_download/5921/1042>.

(FATF), which could work on developing global cybersecurity standards in specific areas (information sharing, professional qualifications, etc.). The FATF is a product of high-level ministerial cooperation and it has been highly influential in setting standards to combat money-laundering and the financing of terrorism. Arguably, a similar model could be adopted by cybersecurity agencies wishing to promote concrete steps towards enhancing global cybersecurity through domestic measures.

It goes without saying that the diplomatic community has a role to play in each of the examples provided above. Diplomatic efforts are needed to initiate, support and sustain the contacts between technical and policy professionals between different states, especially if some of the states will not be "like-minded". Such efforts will also be needed to lend visibility to the discussions taking place, so as to increase their reach and effectiveness.

4. CONCLUSION

The analysis above conveys a few recurring themes. The first is a shift in expectations: while acknowledging that some discussion of cyber norms might contribute to global stability, it would be unrealistic to expect that such stability can be achieved by declaring the existence of a norm or by attempting to operationalize a particular norm. The second theme is the need for a bottom-up approach, driven by actual needs of, and challenges faced by, government cybersecurity organizations. The third and most fundamental theme is the shift in emphasis, from the current discussions focused on global "stability", towards the more mundane goal of domestic cybersecurity.

In their comprehensive paper on cyber norms, Tikk and Kertunen have stated:

> "[...] cyber incident and risk assessments indicate more than state-on-state hostilities. Data breaches, website defacements, increasing cybercrime and botnet topologies, more than they speak of the potential of cyber warfare, testify of a cyber crisis surface where the risk of unwanted or unforeseen developments cannot be effectively prevented due to the still low awareness or obvious capacity gaps. Therefore, the GGE has, without necessarily meaning to, developed at least two separate agendas of international cybersecurity: one that can be understood and explained by way of traditional geopolitics and where the likelihood of conflict or no conflict does not depend significantly on ICT as such. Absent ICTs, the relationships between the US, China, Russia, Iran and North Korea remain largely the same. What geopolitics cannot

> exhaustively explain, is the surface of potential cyber crisis that has emerged by way of extensive adoption of ICTs across the world, without due acknowledgment of the accompanying risks and ways of their mitigation. Jumping on the international information highway has been too fast, too soon, for countries that are not able to run sustainable information systems and services: States that have to run on Windows XP, cannot be helped by any of the UN GGE recommendations."[76]

In very broad terms, the two agendas described above summarize the distinction made in this paper between "global stability", which current cyber norm efforts have been promoting, and domestic cybersecurity, which deserves greater attention from the international community. The effect of the suggested bottom-up, domestic cybersecurity approach is a series of open-ended processes, the milestones of which will likely be more incremental. Its successes will hopefully be enduring and substantive, though they will not grab national headlines. Under this approach, the role of civil society is crucial. Think-tanks, multinational corporations and academics can generate valuable ideas outside conventional thinking, conduct large-scale empirical research and provide a diversity of perspectives that can all feed in to these bottom-up processes. Diplomacy, too, plays a critical role in taking the domestic civilian cybersecurity discussion to the global arena. The challenge for the multi-stakeholder cybersecurity community, then, is to reassess current cyber norm development efforts, adjust expectations, refocus and leap forward with a new sense of purpose.

5. ACKNOWLEDGMENTS

The author thanks Paul Rosenzweig and Duncan Hollis for their insight and comments.

76 Tikk and Kerttuen, n. 14, p. 31.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Developing Collaborative and Cohesive Cybersecurity Legal Principles

Jeff Kosseff[1]
Assistant Professor of Cybersecurity Law
United States Naval Academy
Annapolis, MD, United States

Abstract: Legal discussions about combatting global cyber threats often focus on international cybercrime arrangements or the application of the law of war to cyberspace. While these discussions are vital, policy-makers and scholars have not devoted adequate attention to creating a global legal framework to bolster the defenses of public and private infrastructure. Due to the interconnected nature of cyberspace and the cross-border impacts of attacks, inadequate security in one country could harm another.

To build cyber strategies that rely in part on defense and deterrence by denial, governments should also focus both on the security of their systems and those of the private sector. Industry has been the target of some of the most destructive cyberattacks worldwide. Guiding international principles for a cyber security legal framework would help nations to build effective laws that reduce the likelihood of successful attacks, and increase resilience after attacks occur. Moreover, international collaboration on cybersecurity laws provides multinational companies with a more coherent legal framework. A patchwork of hundreds of different international security requirements is not only burdensome for companies, but it increases the potential for vulnerabilities, particularly if the company operates in countries with less stringent cybersecurity requirements.

This paper sets out the need for nations to discuss common legal principles for promoting and regulating cybersecurity, similar to the privacy principles articulated

1 Assistant Professor of Cybersecurity Law, United States Naval Academy, Annapolis, MD. J.D., Georgetown University Law Center. M.P.P., B.A., University of Michigan. Thanks to LCDR Joseph Hatfield and Professor Martin Libicki for helpful feedback. The views expressed in this paper are only those of the author, and do not represent the U.S. Naval Academy, Department of the Navy, Department of Defense, or any other party.

in the Organization for Economic Cooperation and Development's Fair Information Practices in 1980. As a starting point for discussion, this paper suggests four goals of common international principles for cybersecurity law: (1) modernization of cybersecurity laws; (2) uniformity of legal requirements; (3) coordination of cooperative incentives and coercive regulations; and (4) supply chain security. Although cybersecurity laws will always vary, international coordination could improve their efficacy by providing some degree of consistency. A dialogue also could help policy-makers learn from other nations' cybersecurity successes and failures.

Keywords: *cybersecurity; cooperation; principles; cybercrime; data security*

1. INTRODUCTION

Over the past decade, there has been great progress on international cooperation to combat cybercrime and build on norms to deter and deny states that leverage the asymmetric nature of cyber operations. All of these discussions are vital and must continue on the international stage. However, international legal discussions also must address cybersecurity law.

At the outset, this Paper defines "cybersecurity law," as the term is often used interchangeably to describe regulation of the private sector's computer systems and networks, federal programs that assist the private sector, cybercrime statutes and the legal norms of cyberwar. For the purposes of this paper, I broadly define cybersecurity law as domestic laws that seek to promote the confidentiality, integrity and availability of public and private computer systems, networks and information.[2] This expansive definition applies equally to governmental regulations and public-private partnerships and to incentives that have the ultimate goal of improving cybersecurity.

Improving the cybersecurity of public and private systems has two primary national security benefits. First, hardened defenses help to reduce or eliminate harm caused by an aggressor. Second, cybersecurity is an important part of a framework to deter attacks, provided that the aggressor is aware of the strong defenses. While deterrence by punishment is an important component of a cyber strategy, so too is deterrence by denial. Cyber deterrence requires nations to ensure that their laws provide adequate assistance and incentives for cybersecurity of both government and private infrastructure. Too often, the security of the private sector is missing from the greater discussion of national cybersecurity.[3] Governments worldwide have recognized the

2 *See* Jeff Kosseff, *Defining Cybersecurity Law*, 103 IOWA L. REV. 985 (2018).

3 *See* Kristen E. Eichensehr, *Public-Private Cybersecurity,* 95 TEX. L. REV. 467, 536 (2017) ("As the operation of government-like power becomes more diffuse and more complicated, the actions of private sector actors can implicate the public law values that traditionally apply to governmental actions, and governmental actions may come into increasing tension with public law values.").

need for private companies to protect their data and cyber infrastructure. The private sector controls vast amounts of infrastructure that are vulnerable to cyberattacks, making the private sector's cybersecurity important not only to nations' economies, but also to their national security.[4]

The interconnected nature of cyber threats – in which an attack in one country could cause harmful spill-over effects in another country – provides policy-makers with a compelling reason to improve cybersecurity laws globally. To do so, nations should collaborate and articulate core principles for cybersecurity, just as the Organization for Economic Cooperation and Development (OECD) did for privacy law nearly four decades ago when it developed its Fair Information Practices.

This paper then draws on examples of successful cybersecurity laws and partnerships worldwide to outline some goals of a global cybersecurity legal framework:

- Modernization of cybersecurity laws to address current threats;
- Uniformity of legal and regulatory requirements;
- Coordination of cooperative cybersecurity programs and regulatory obligations; and
- Supply chain security.

Cybersecurity often involves an alignment of public-sector and private-sector interests. Accordingly, cybersecurity law should move from the outdated, purely punitive model of privacy law to a collaborative and cooperative framework. I refer to this model as "collaborative cybersecurity law," a mixture of incentives, public-private partnerships, and tailored regulations that is designed to improve cybersecurity as a whole.

For this paper, collaborative cybersecurity law has two equally important applications. First, the public and private sectors should collaborate to determine the most effective legal frameworks to build defenses and resilience. Second, governments should collaborate at the local, state/province, and national levels to ensure that their requirements and incentives are aligned to the common goal of protecting global cyber infrastructure. Cyberspace does not have clearly defined geographic or public/private boundaries. Nor should the defense of cyberspace.

I do not suggest the creation of a single set of cybersecurity laws to apply across all nations; such a task would be a fool's errand, as countries have a wide range of tort, constitutional, and administrative laws that would prevent a single law across all jurisdictions. Jurisdictions such as the United States tend to favor cybersecurity laws

4 *See* Roger Hurwitz, *The Play of States: Norms and Security in Cyberspace*, 36 AMERICAN FOREIGN POLICY INTERESTS 322 (2014) ("Our discussion suggests that efforts to establish a state-led comprehensive regime for cyberspace will not succeed, notwithstanding the illusion that it could provide a more stable order and block fragmentation of the Internet.").

that promote free expression over other interests, while jurisdictions such as those in Europe tend to favor privacy protection. Rather than attempt a uniform set of laws, countries should develop a set of shared core cybersecurity values to apply as they develop laws to address cybersecurity threats via laws, regulations, and government programs.

In short, this paper argues that nations must broaden their conception of the international cybersecurity dialogue. While the ongoing discussions regarding cyberwarfare norms are essential, it is only one piece of the much larger solution to improving the security of cyberspace. Nations must also develop a cohesive strategy to secure both public and private cyber information and infrastructure through regulations and incentives.

2. THE GLOBAL IMPACT OF INADEQUATE CYBERSECURITY

Cyber threats are not always confined to geographic borders. Many of the most damaging and persistent cyberattacks have targeted systems and data in multiple countries. The attacks target not only military systems or civilian government computers, but often also home systems that are operated by the private sector. With the private sector controlling critical infrastructure such as logistics, telecommunications, and financial systems globally, the cybersecurity of both the public *and* private sector is crucial to adequate defense.

The pervasive global nature of cyber threats can be seen in botnets, which use infected computers to amass power to launch devastating attacks. As botnets infect more computers, they cause more damage, such as forcing websites offline and interrupting critical services.[5] The Internet of Things era has exponentially increased the number of devices connected to the Internet. Botnets have commandeered these devices, in part due to the inadequate security measures on many IoT devices.[6]

For instance, in October 2016, the Mirai botnet, consisting of hundreds of thousands of infected devices, knocked some of the most popular websites in the world offline by targeting Dyn, a domain name system management service.[7]

Botnets demonstrate the international impact of inadequate cybersecurity. Consider, for example, a webcam that is manufactured in Germany with inadequate password protections. If a consumer in the United States uses that webcam, it could be used in a

5 *See* Elisa Bertino & Nayeem *Islam, Botnets and Internet of Things Security*, 50:2 COMPUTER 76-79 (Feb. 2017).

6 *See* Bernard Marr, *Botnets: The Dangerous Side Effects of The Internet of Things*, FORBES (Mar. 7, 2017).

7 Lorenzo Franceschi-Bicchierai, *Blame the Internet of Things for Destroying the Internet Today*, MOTHERBOARD (Oct. 21, 2016).

botnet that shuts down a website in New Zealand. New Zealand alone cannot address the botnet problem by regulating the security of Internet of Things devices.

Likewise, the WannaCry ransomworm demonstrates the interconnected nature of cyber threats. WannaCry was initially found on European businesses' computers on the early morning of May 12, 2017. The files on infected computers were encrypted, and the computer operators received a demand for bitcoin in exchange for the encryption key, though paying the ransom did not always guarantee decryption of the files. The ransomworm rapidly spread. In all, WannaCry infected more than 200,000 computers around the world.[8]

WannaCry was so malicious and pervasive because it spread using EternalBlue, an exploit that allows malware to spread in Windows operating systems. Hackers allegedly stole EternalBlue from the U.S. National Security Agency.[9] The U.S. and UK authorities have attributed WannaCry to North Korea.[10]

According to the European Union Agency for Network and Information Security, once a computer was infected with WannaCry, it would scan public Internet Protocol addresses for other external networks to infect.[11] Rather than merely spreading across a company's internal network, WannaCry used its infected computers to find and target other vulnerable networks.[12]

WannaCry and Mirai demonstrate the globally interconnected nature of harms associated with cyberattacks. The attacks demonstrate that an attack that initially focuses on one geographic region can have immediate and damaging spill-over effects into other countries. Therefore, it is in a nation's interests to secure not only the computers within its geographic boundaries, but the systems and networks across the globe.

3. THE NEED FOR LEGAL PRINCIPLES TO IMPROVE GLOBAL CYBERSECURITY

Enhanced cybersecurity of a nation's infrastructure plays two critical roles in cyber strategy. First, it reduces or eliminates the risk of harm from an attempted attack by bolstering defenses. Second, the known existence of the attack may deter the attacks from ever occurring.

8 Sam Jones, Timeline: *How the WannaCry Cyber Attack Spread*, FINANCIAL TIMES (May 14, 2017).
9 *Ibid*.
10 David E. Sanger, *U.S. Accuses North Korea of Mounting WannaCry Cyberattack*, N.Y. TIMES (Dec. 18, 2017).
11 European Union Agency for Network and Information Security, *WannaCry Ransomware Outburst* (May 15, 2017); Adam McNeil, *How Did the WannaCry Ransomware Spread?* MALWAREBYTES (May 19, 2017).
12 *See* Abishek Singh, *WannaCry Ransomware Analysis: Lateral Movement Propagation*, ACALVIO (May 16, 2017).

Deterrence strategy has two components: deterrence by punishment and deterrence by denial.[13] Deterrence by denial consists of strategies that both resist attacks and help recovery from attacks once they have occurred (known as "resilience.").[14] For effective cyber deterrence by denial, the private sector must both secure its own system and networks and develop secure products throughout the supply chain. As Dorothy Denning summarized in 2016:

> Cybersecurity aids deterrence primarily through the principle of denial. It stops attacks before they can achieve their goals. This includes beefing up login security, encrypting data and communications, fighting viruses and other malware, and keeping software updated to patch weaknesses when they're found.
>
> But even more important is developing products that have few if any security vulnerabilities when they are shipped and installed. The Mirai botnet, capable of generating massive data floods that overload internet servers, takes over devices that have gaping security holes, including default passwords hardcoded into firmware that users can't change. While some companies such as Microsoft invest heavily in product security, others, including many Internet-of-Things vendors, do not.[15]

Nations can promote such cybersecurity measures by enacting effective regulations and creating public-private partnerships. Defending against attacks helps to mitigate the overall harm.[16] However, a single nation's laws are likely to be insufficient to adequately shore up its cybersecurity. The cyber vulnerabilities in Country A may lead to negative consequences in Country B, and Country B has limited ability, acting alone, to impose consequences for inadequate cybersecurity in Country A. That is where an international dialogue on cybersecurity is vital.

Even to the extent that some cyberattacks are strictly local, an international dialogue about cybersecurity laws can allow nations to share lessons about their experiences with government programs, regulations, and laws. Unlike other areas of law that have centuries of empirical evidence to support or reject their adoption, cybersecurity law needs to address the rapidly evolving threat landscape. If, for instance, requiring a particular safeguard is effective, nations could share these experiences in determining best practices.

13 *See* A. Wess Mitchell, *The Case for Deterrence by Denial*, THE AMERICAN INTEREST (Aug. 12, 2015).
14 *See* Annegret Bendiek and Tobias Metzger, *Deterrence Theory in the Cyber Century*, in LECTURE NOTES IN INFORMATICS (LNI), GESELLSCHAFT FÜR INFORMATIK, BONN 2015.
15 Dorothy Denning, *Cybersecurity's Next Phase: Cyber Deterrence*, SCI. AMERICAN (Dec. 2016).
16 *See* Martin Libicki, Cyberdeterrence and Cyberwar 176 (2009).

In both the areas of cybercriminal law[17] and cyberwarfare,[18] international experts and policy-makers have at least attempted to find areas of broad agreement. However, criminal laws and warfare norms and guidelines often address *responses* to cyberattacks (i.e., criminal prosecutions or military action). While these are absolutely vital to a comprehensive cybersecurity framework, they are only part of the solution. Laws and regulations also should seek to bolster defenses to prevent attacks from succeeding in the first place.

The Council of Europe's Convention on Cybercrime (the Budapest Convention) sets minimum requirements for computer crime statutes in participating nations and provides for mutual assistance in investigating and prosecuting cybercrimes. This cooperation and harmonization is necessary because of the global nature of cybercrimes, and the criminal is often located in a different country from the target.[19] By harmonizing cybercrime laws, the Budapest Convention reduces the likelihood of some countries becoming "safe havens" for cybercriminals.[20] However, the Budapest Convention has been criticized for being unsuccessful and overall not helping to crack down on cybercrime.[21] It has not been adopted outside of a majority of Council of Europe members and the United States. When Russia, North Korea, Iran, China, and other non-members often are the sources of cyber-attacks, the Budapest Convention provides the target countries with little recourse. Moreover, criminal law alone is not always sufficient to prevent attacks in cyberspace due to the challenges of attributing attacks with certainty.[22] While the Budapest Convention plays an important role in harmonizing at least some cybercrime laws in some countries, it is not a panacea.[23]

In some respects, there are even more benefits to coming to a consensus on international cybersecurity law than in criminal law. The Budapest Convention is of limited utility because many of the most pernicious attacks are perpetrated from nations that are not parties to the Convention; laws that effectively promote the cybersecurity of public and private systems and networks, however, provide incremental worldwide benefits, even if they have not been adopted by the handful of nations that are the sources of the attacks. Consider, for example, a cybersecurity regulatory framework that bolsters resistance and reduces the spread of botnets by 75 percent in countries that have adopted its safeguards. If half of the nations were to adopt the framework, the overall

17 ETS 185 – Budapest Convention on Cybercrime, 23.XI.2001.

18 *See* TALLINN MANUAL 2.0 ON INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (Michael N. Schmitt ed., 2017) (hereinafter, *"Tallinn Manual"*).

19 *See*, e.g., Jonathan Clough, *A World of Difference: The Budapest Convention on Cybercrime and the Challenges of Harmonisation*, 40 MONASH L.R. 699, 700 ("Although many offences are transnational in nature – for instance trafficking in humans, weapons and drugs, money laundering and terrorism – cybercrime presents unique challenges due to the inherently transnational nature of the underlying technology.").

20 *Id*. at 700.

21 *See, e.g*., Jack Goldsmith, *Cybersecurity Treaties, A Skeptical View*, in FUTURE CHALLENGES IN NATIONAL SECURITY AND LAW (Feb. 2011).

22 Lily Hay Newman, *Hacker Lexicon: What is the Attribution Problem?* WIRED (Dec. 24, 2016).

23 *See* Kim-Kwang Raymond Choo, *The Cyber Threat Landscape: Challenges and Future Research Directions*, COMPUTERS & SECURITY 30:8 (Nov. 2011).

strength of a botnet likely would weaken because it would not be as successful in propagating.

Similarly, the growing body of scholarship that applies *jus ad bellum* and *jus in bello* to cyberwarfare is absolutely essential to our understanding of acceptable responses to cyberattacks and it helps to inform deterrence strategies. Understanding the application of *jus ad bellum* to cyberspace is essential in informing a deterrence by punishment strategy. The two editions of the *Tallinn Manual* have provided a forum for an International Group of Experts on the law of war to articulate both commonalities and differences in views about how their field applies in cyberspace.[24] Although the *Tallinn Manual* does not represent the official views of a single organization or state,[25] it is one of the greatest steps in articulating commonalities and differences in international cyber law.[26]

Likewise, from 2016-17, the United Nations Group of Government Experts attempted to reach an agreement on norms of cyber issues such as international humanitarian law and the right of self-defense. However, those discussions failed to lead to a consensus, as some participants had very different views on the fundamental international norms.[27] Indeed, such consensus will be difficult or impossible for norms related to *jus ad bellum* and *jus in bello*. But such issues should not be the only focus of international discussions. Global norms for *domestic* cybersecurity issues could play an equally vital role in securing cyberspace.

The cybersecurity of a nation's infrastructure may play a significant role in its response to a cyberattack, as the success or failure of cyberdefense often determines whether a cyber act constitutes an unlawful use of force.[28] Consider, for instance, a cyberattack by Iran on a portion of the U.S. power grid that is operated by a private company. If the utility has installed sufficient safeguards, the attack may be nothing more than a nuisance that causes little damage. If, however, the attack succeeds, it could cause significant economic loss, and perhaps even personal injury. Those two outcomes would warrant very different responses under international warfare norms. Just as

24 *See, e.g.*, Kristen Eichensehr, *Review of* The Tallinn Manual on the International Law Applicable to Cyber Warfare, 108 A. J. INT'L L. 585, 586 (2014) ("While the rules on which the IGE agreed are very useful in advancing thought and debate about international law regarding cyberwar, more valuable still are the instances in which the *Tallinn Manual* frankly acknowledges disagreement within the IGE.").

25 *See* TALLINN MANUAL at 2 ("Ultimately, *Tallinn Manual 2.0* must be understood only as an expression of the two International Groups of Experts as to the state of the law.").

26 *See* Gary Korn, *Tallinn Manual 2.0, Advancing the Conversation*, JUSTSECURITY (Feb. 15, 2017) ("[T]he advisory nature of Tallinn 2.0 should not detract from its immense value to legal practitioners and their clients in both the public and private sector as a quality compendium of the general framework of international rules and principles most pertinent to cyber operations.").

27 *See* Remarks of Michele G. Markoff, Deputy Coordinator for Cyber Issues, U.S. Department of State (June 23, 2017) ("It is unfortunate that the reluctance of a few participants to seriously engage on the mandate on international legal issues has prevented the Group from reaching consensus on a report that would further the goal of common understandings among UN Member States on these important issues.").

28 *See* Priyanka R. Dev, *'Use of Force' and 'Armed Attack' Thresholds in Cyber Conflict: The Looming Definitional Gaps and the Growing Need for Formal U.N. Response*, 50 TEX. INT'L L. J. 379 (2015).

the international legal community has attempted to develop common ground as to the application of the law of war to cyber, so too should the community develop principles that guide the protection of cyber infrastructure.

Efforts to develop transnational common ground on cybercrime law and cyberwarfare norms will not solve all of the complex international legal problems associated with threats, though they are necessary components of the overall approach to cybersecurity. Moreover, both efforts provide roadmaps for international dialogues about cybersecurity laws that deter by denial. The Budapest Convention and the *Tallinn Manual* demonstrate that it is possible for nations with different values to at least agree on some core principles for cyberspace. Both the Budapest Convention's formal attempts at proscribing specific cybercrime laws and the *Tallinn Manual'*s attempts to narrate common, nonbinding interpretations are essential as nations confront growing cyber threats.

Although there is not currently a universal set of cybersecurity principles outside of the cybercrime and cyberwarfare contexts, an analogue exists in the privacy arena and demonstrates the utility of setting forth a core set of shared legal values for technology law. In 1980, the OECD, an economic development organization consisting of 35 nations, published the *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, the centerpiece of which was the OECD Fair Information Practices.[29]

Drawing on robust discussions among participating countries, OECD developed the following eight general principles for information privacy: collection limitation; data quality; purpose specification; use limitation; security safeguards; openness; individual participation; and accountability.[30] The Guidelines have been revised only once, in 2013. Each of the eight principles provides a broad framework under which nations could choose how to best regulate privacy. For instance, the collection limitation principles state that "[t]here should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject".[31]

Broad principles such as this allow for some standardization across nations; yet they also provide countries with the flexibility to adhere to these principles within their existing legal systems and policy preferences.[32] The OECD Guidelines have helped

29 *See* Pam Dixon, *A Brief Introduction to Fair Information Practices*, World Privacy Forum, available at https://www.worldprivacyforum.org/2008/01/report-a-brief-introduction-to-fair-information-practices/.

30 OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA.

31 *Ibid*.

32 *Id*. at 48, Original Explanatory Memorandum to the OECD Privacy Guidelines ("On the whole, the Guidelines constitute a general framework for concerted actions by Member countries: objectives put forward by the Guidelines may be pursued in different ways, depending on the legal instruments and strategies preferred by Member countries for their implementation.").

to shape the contours of privacy laws around the world, even beyond the 34 OECD member nations.[33]

The OECD Guidelines are privacy-focused, though the document's Security Safeguards Principle states that personal data "should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data". The supplemental memorandum for the 2013 revisions suggests that these safeguards include data security breach notification requirements. Although this principle touches on a cybersecurity issue, it focuses on personal information security and does not adequately address the full range of cybersecurity threats, as discussed in the next section. Privacy and cybersecurity are often lumped into the same category of law and share some common issues, but they each present different challenges and should be individually addressed.[34] While the protection of personal information certainly is part of cybersecurity, other threats, such as the theft of trade secrets or attacks on cyber-physical systems, are not adequately addressed by privacy law.[35] Cybersecurity law should promote not only the privacy of personal data, but also the protection of systems and data from attacks that could interrupt economies or threaten national security.

This is not to suggest that the OECD framework has perfectly aligned the privacy laws and regulations of all member nations. Far from it. The European Union views privacy as a fundamental human right, and therefore its privacy laws, including the new General Data Protection Regulation, are often far more stringent than those of other jurisdictions. However, the OECD Principles, at the very least, give participating nations a basis on which to find some commonalities and a general framework for discussing and debating privacy issues.

4. GOALS FOR INTERNATIONAL CYBERSECURITY LEGAL PRINCIPLES

Because nations have had few robust and meaningful discussions about how to promote and regulate cybersecurity via legal frameworks, it would be impossible to propose a comprehensive set of principles to guide governments globally.

33 *See, e.g.*, Monika Kuschewsky, *What Does the Revision of the OECD Privacy Guidelines Mean for Businesses*, AB EXTRA (Oct. 22, 2013) ("Today, its basic privacy principles are essentially reflected in all relevant general data protection frameworks worldwide.").

34 *See, e.g.*, Bob Siegel, *What is the Difference Between Privacy and Security?*, CSO (May 26, 2016) ("A security program protects all the informational assets that an organization collects and maintains. A privacy program focuses on the personal information an organization collects and maintains.").

35 The OECD in 2002 adopted its Guidelines for the Security of Information Systems and Networks, which sets out nine general principles for information security. Although these Guidelines are useful, they do not address the problem that this paper seeks to address. The guidelines apply equally to government entities, businesses, and individual users, and focus more on ethical information security norms rather than guidelines for laws. They do not provide the same level of general principles for laws that OECD's privacy principles do. The guidelines are focused on the security of information, and do not address the comprehensive threats to cybersecurity that nations currently face.

This part sets out the goals of global cybersecurity legal standards and a few areas to begin discussions among nations as they determine how best to address cybersecurity challenges via laws and regulations. To be clear, I do not suggest that this should serve as the list of international cybersecurity principles. Such a framework would require significant multilateral discussion and assessments of both the cybersecurity threats and the legal capabilities and constraints to address those threats. Rather, these four goals are broad topic areas that serve as a starting point for an international discussion about common principles.

A. Modernizing Laws to Address Current Cybersecurity Threats

The laws in many nations do not adequately address some of the newer cybersecurity threats, as the laws are outgrowths of pre-Internet legal fields such as privacy torts and criminal law. International norms could help guide nations as they adjust their laws to the current threat landscape.

One of the core concepts in the cybersecurity field is the CIA Triad: confidentiality, integrity, and availability of data, systems, and networks.[36] Confidentiality protects information from unauthorized access.[37] Integrity ensures that the information is accurate and systems function as intended.[38] Availability guarantees uninterrupted access to information and systems.[39] An effective cybersecurity program will advance all three goals.

Unfortunately, cybersecurity law is often conflated with data security and privacy laws that have been on the books for many years or decades. This results in a focus on the confidentiality of personal information, which is the primary security-related concern of privacy law. Without a doubt, that is an important concern, but it overlooks the confidentiality of other critical but non-personal information, such as corporate trade secrets or classified government information. For instance, many jurisdictions require companies to notify individuals and regulators about disclosure of certain categories of personal information, and data security requirements often apply to particularly sensitive types of personal information such as medical records.

Privacy law cares little about integrity or availability, nor do any data security laws that are largely an outgrowth of privacy laws. Data security regulations, for example, often address the unauthorized access to or acquisition of data. These laws typically do little to address attacks on availability (such as ransomware) or attacks on integrity (such as website defacement or modifications to database systems that cause physical impacts, such as explosions in gas lines).[40]

36 *See* U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, FEDERAL INFORMATION PROCESSING STANDARDS PUBLICATION 199.
37 *Id*. at 2.
38 *Ibid*.
39 *Ibid*.
40 *See, e.g*., Derek E. Bambauer, *Schrodinger's Cybersecurity*, 48 U.C. DAVIS L. REV. 791 (2015).

Laws should, of course, continue to protect confidentiality. Protecting confidentiality and privacy is not mutually exclusive with protecting integrity and availability. Indeed, many of the concerns regarding interference in the 2016 American election boil down to breaches of confidentiality: the hacks of John Podesta's email account and the Democratic National Committee's servers. However, confidentiality should not be the exclusive focus, particularly in the age of cyber-physical systems and the Internet of Things, when everyday devices are increasingly connected to the Internet and could be vulnerable to attacks. A modern cybersecurity framework must address these threats as well as data breaches.

In addition to promoting all three prongs of the CIA triad, cybersecurity laws should be forward-looking and should minimize harm from future cyberattacks. Ideally, such laws would require companies and governments to bolster defenses to a point where the attacks do not succeed. However, it is highly unlikely that any legal system would entirely prevent all attacks. For that reason, a modern cybersecurity legal framework should also strive to improve resilience – the ability of a company or government to quickly recover after an attack has occurred.[41]

B. Uniformity of Regulations

Regulation of the private sector plays a key role in securing cyber infrastructure. Companies that have some of the most critical cyber infrastructure operate in many countries. Those companies, therefore, are subject to hundreds of legal regimes at the local, state/province, and national levels. To the greatest extent possible, cybersecurity regulations should be standardized across governments to improve the ease and likelihood of compliance. International norms could help to guide that uniformity.

For instance, companies are subject to dozens of data breach notification laws at the state/province and national levels, all varying in terms of the specific requirements that they impose as to what types of personal data trigger the notification requirements and the forms that the notices must take.[42] The breach notice laws apply based on the residency of the individuals whose data was breached. Thus, a company that has customers throughout the world must comply with all of these requirements in the days following a breach. Such compliance can be time-consuming, and can divert attention from efforts to remedy the harms caused by the breach and prevent further intrusions.[43]

Policy-makers at the international level could help strive toward such uniformity by adopting standards that could be the basis of private sector requirements, and jurisdictions should aim for uniformity among the regulations of state, provincial,

41 *See* Fredrik Hult & Giri Silvanesan, *What Good Cyber Resilience Looks Like*, J. OF BUS. CONTINUITY & EMERGENCY PLANNING 112 (2013-14).

42 *See* World Law Group, GLOBAL GUIDE TO DATA BREACH NOTIFICATIONS (2016).

43 *See* Brett V. Newman, *Hacking the Current System: Congress' Attempt to Pass Data Security and Breach Notification Legislation*, 2015 U. ILL. J.L. TECH & POL'Y 437, 442 (2015).

and local governments. The European Union's GDPR, for example, aims to improve uniformity among European Union members by imposing a single comprehensive set of requirements for privacy and security practices when dealing with European residents' personal information.[44]

Complete global uniformity of cybersecurity laws is impossible, as countries will differ in their legal constraints and values regarding issues such as privacy, expression, and security. For instance, in Europe, privacy is a fundamental human right, while the United States is more likely to balance privacy with other interests such as free expression.[45] However, even some movements toward similar cybersecurity regulations would be useful in providing companies with more effective pathways to comply with the global patchwork of laws.

C. Coordination of Coercive and Cooperative Laws

Cybersecurity laws should contain a mixture of punitive regulations and incentives to promote private sector security. Regulations will always play an important part in bolstering companies' cybersecurity. However, cybersecurity differs from other regulated areas in that the government's goals are often generally aligned with the goals of a company. A rational chief executive does not want their company to experience a denial of service attack or data breach, nor does a rational government official.

For that reason, there is great room for collaboration between the public and private sectors. Such collaboration should form part of a broader strategy for bolstering the cybersecurity of public and private infrastructure.

For instance, governments across the world are increasingly improving and expanding their cyber threat information sharing programs, which allow the private and public sectors to exchange information and collaborate to reduce the spread and damage of cyberattacks. In the European Union, the 2016 NIS Directive requires member states to establish Computer Security Incident Response Teams that monitor, share, and collect information about cyber threats and "establish cooperation relationships with the private sector".[46] Likewise, in late 2015, the U.S. Congress passed the Cybersecurity Information Sharing Act, which provides companies with limited legal immunity for sharing cyber threat information and defensive measures with other companies and the federal government's threat information sharing program. The statute has been called "the first major piece of cybersecurity legislation enacted into

44 *See* Terry Greer-King, *GDPR is Coming: 5 Things to Be Aware Of*, Cisco UK & Ireland Blog (Feb. 23, 2017) ("[E]ach country currently has their own ways of coming up with legislation to control data rights. GDPR is going to drive some uniformity, and make it easier to legislate.").

45 *See* Mark Scott & Natasha Singer, *How Europe Protects Your Online Data Differently Than the U.S.*, N.Y. TIMES (Jan. 31, 2016).

46 Annex I to Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union ("NIS Directive").

law that seeks to directly address the relationship between the private and public sectors".[47] An international dialogue on such efforts could establish best practice for such threat-sharing efforts and might also lead to more effective means of exchanging critical threat information internationally.

Cybersecurity education also requires collaborative efforts from both the public and private sectors. It includes general awareness campaigns to reduce the success of phishing and other social engineering attacks, as well as more advanced collegiate and graduate school training to build a cybersecurity workforce. For instance, the Israeli National Cyber Bureau has developed a plan both to build cybersecurity awareness among the general public,[48] and the EU's NIS Directive requires each member state to adopt a strategy that addresses "education, awareness-raising and training programs relating to the national strategy on the security of network and information systems".[49]

Governments could also provide financial incentives, such as tax credits and research and development funding, to encourage potential targets to invest large sums of money and staffing to bolster their cybersecurity. Because many high-profile targets are multi-national corporations, international coordination on incentives such as tax credits would be particularly useful in developing a global strategy.

International norms to improve cybersecurity education are particularly useful with a global information technology workforce. Nations could determine any particular skill shortages within cybersecurity and align educational programs accordingly. Moreover, international principles could help to guide and improve cybersecurity awareness campaigns to reduce the likelihood of cybersecurity attacks succeeding due to human error.

D. Secure Throughout the Supply Chain

Just as cybersecurity threats arise due to the global interconnection of networks and systems, they also often arise because products and services rely on a number of components developed around the world and inadequate security of a component can make an entire product or service vulnerable. Countries have individually begun addressing the supply chain in a thoughtful manner. For instance, in 2008, the United States began its Comprehensive National Cybersecurity Initiative, which recognized the need for "partnership with industry to develop and adopt supply chain and risk management standards and best practices".[50] However, the Initiative recognized that supply chain cybersecurity is not merely a problem that arises from U.S. companies:

47 Jamil N. Jaffer, *Carrots and Sticks in Cyberspace: Addressing Key Issues in the Cybersecurity Information Sharing Act of 2015*, 67 S.C. L. REV. 585, 586 (2016).

48 *See* Daniel Benoliel, *Towards a Cybersecurity Policy Model: Israel National Cyber Bureau Case Study*, 16 N.C. J.L. & TECH 435, 446 (2015).

49 NIS Directive at Art. 7(1)(d).

50 COMPREHENSIVE NATIONAL CYBERSECURITY INITIATIVE, *available at* https://obamawhitehouse.archives.gov/node/233086.

"Risks stemming from both the domestic and globalized supply chain must be managed in a strategic and comprehensive way over the entire lifecycle of products, systems and services."[51]

International standards for supply chain cybersecurity would be particularly useful, as products may rely on technology that is manufactured in many nations. A substantive dialogue between governments and industry could develop best practices for supply chain cybersecurity, which could be used as the basis for national or regional cybersecurity laws. Such standardization could improve the overall security of products and services while increasing the ease of compliance.

5. CONCLUSION

This paper argues that nations should broaden their cyber discussion beyond cyberwarfare and attempt to improve the patchwork of domestic laws that seek to improve the cybersecurity of public and private infrastructure and information. Nations cannot address cybersecurity threats merely by developing domestic legal rules that fail to account for the laws and programs in other nations. An international framework for cybersecurity would help nations to align their regulations and public-private partnerships to address threats that often know no borders. Effective cybersecurity laws require collaboration between governments worldwide and between the public and private sectors. Although nations will continue to carve out their own paths, a productive international dialogue would help policy-makers to find some common ground on effective cybersecurity laws and programs.

51 *Ibid*.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Utilizing Air Traffic Communications for OSINT on State and Government Aircraft

Martin Strohmeier
Department of Computer Science
University of Oxford
Oxford, United Kingdom
martin.strohmeier@cs.ox.ac.uk

Matthew Smith
Department of Computer Science
University of Oxford
Oxford, United Kingdom
matthew.smith@cs.ox.ac.uk

Daniel Moser
Department of Computer Science
ETH Zurich
Zurich, Switzerland
daniel.moser@inf.ethz.ch

Matthias Schäfer
Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
schaefer@cs.uni-kl.de

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Ivan Martinovic
Department of Computer Science
University of Oxford
Oxford, United Kingdom
ivan.martinovic@cs.ox.ac.uk

Abstract: In recent times, we have witnessed a trend in which communications data is increasingly collected and made open source by the public. A prominent example is the tracking of aircraft movements using unencrypted air traffic control (ATC) communication. This paper studies the implications of such new open source aircraft datasets on the operational privacy of military and government actors. We use publicly available aircraft metadata in conjunction with unfiltered ATC communication gathered from the collaborative sensor network OpenSky. We show that using these datasets, it is possible to collect, process and analyze large numbers of movements in an automated fashion, providing insights into potentially sensitive operations.

We use movement data collected from more than 580 identified aircraft used by 100 different governments and over 6,000 military aircraft to identify operations and relationships in the real world. We also provide case studies which show that potentially sensitive information appears in these open datasets in the clear from both military and government-operated aircraft, despite attempts at encrypting some of this information.

Considering these privacy violations, we establish which countries' militaries and governments take active steps in blocking the movements of their sensitive aircraft from online tracking websites. We find that overall more than 80% of all military aircraft and 60% of all government aircraft are filtered for reasons of privacy, with significant variation between different countries.
Finally, we study the main mitigation methods available to state aircraft operators and find that all currently existing options have significant downsides, which inhibit either their usability or their effectiveness.

Keywords: *OSINT, wireless security, air traffic communication, sensor networks, privacy*

1. INTRODUCTION

Nation states and military organizations have a long tradition of intelligence gathering for purposes such as national security, counter-terrorism or counter-proliferation. The public has often held these intelligence activities in contempt, as the associated data collection methods tend to be intrusive to personal privacy. In recent times, however, we have witnessed the opposite trend in which people themselves are increasingly collecting and analyzing intelligence data concerning state and military activities.

One of the most prominent examples is the tracking of military and government aircraft movements. As active communities surrounding affordable software-defined radios have brought previously hard-to-access communications into the reach of low-skilled observers, effective privacy no longer exists on unencrypted radio channels. Many avionics communications use such channels, transmitting messages for private, military, and governmental aircraft [1], [2]. Thus far, privacy, whilst used for civil air traffic communication, is ensured solely by means of policy.

This paper studies the implications of new open source aircraft data collection initiatives on the privacy of military and government actors. We used publicly

available aircraft metadata in conjunction with unfiltered air traffic communication data gathered from the collaborative sensor network, OpenSky [3]. We collected and examined messages sent via the ACARS and ADS-B protocols by military and government-operated planes over the period of one year. We show that it is possible to collect and process large amounts of data in an automated fashion, providing insights into potentially sensitive operations conducted by military and government aircraft. The novelty of this work is that such analysis is possible using open source data and is not restricted to professional intelligence services, but rather can be conducted by a wide range of actors.

In our work, we applied both large dataset analysis and case studies to illustrate the potential impact of air traffic data for intelligence purposes in several different areas. Our contributions in this paper are:

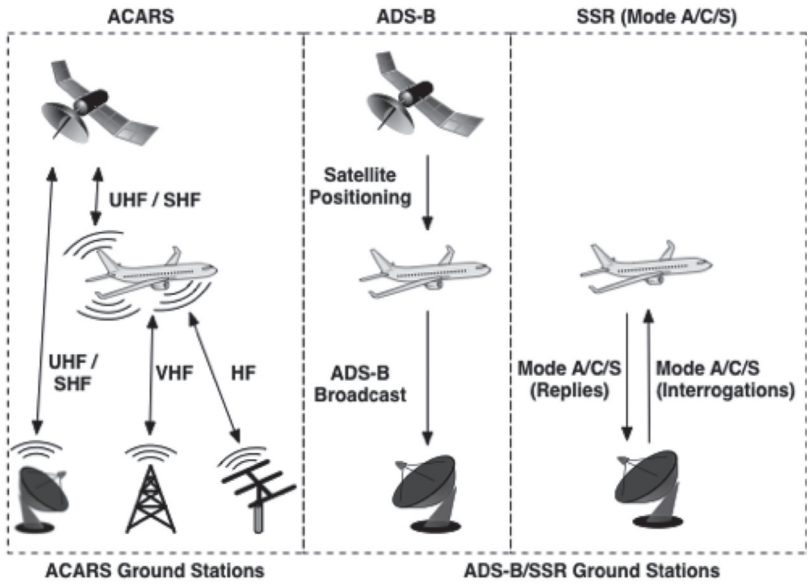
- We use movement data collected from more than 500 identified aircraft used by 100 different governments as well as over 6,000 military aircraft to identify operations and relationships in the real world.
- We provide evidence that potentially sensitive information is communicated in the clear by both military and government-operated aircraft using ACARS, despite attempts at encrypting some of this information.
- We establish which countries' militaries and governments are aware of the existence of large commercial air traffic sensor networks and take active steps to block the tracking of their sensitive aircraft on these websites.
- Finally, we examine the technical mitigation options open to state aircraft operators. Based on our analysis, we argue that all existing methods have severe drawbacks, which either inhibit their usability or their effectiveness.

In the remainder of this work, we first briefly describe the ATC technologies which we exploited in Section 2. Section 3 describes the crowdsourced system and the available public datasets which were used. Section 4 introduces our threat model, Section 5 presents the approach and the obtained results, and Section 6 analyzes the potential mitigations. Finally, Section 7 discusses the implication of our results and Section 8 concludes this paper.

2. BACKGROUND

Figure 1 provides an abstract overview and comparison of the wireless communication links of the three considered technologies, which are explained in the following sections.

FIGURE 1: REPRESENTATION OF ADS-B, SSR, AND ACARS SYSTEMS.



A. ACARS

The Aircraft Communications Addressing and Reporting System (ACARS) has been in use for over 20 years, providing a digital data link between the ground and the air [4]. It serves two main purposes: to administer ATC in order to decongest voice frequencies, and to improve efficiency for aircraft operations. As such, it can be used for safety critical procedures such as negotiating ATC clearance, as well as operational purposes including maintenance reports, engine data and weather information.

It is served over three bands: High Frequency (HF), Very High Frequency (VHF), and Satellite Communications (SATCOM). Most aircraft are equipped for all three, but may choose to not use one or more. VHF is further split into Plain Old ACARS (POA) and VHF Data Link mode 2 (VDLm2); the former is older and slower than the latter, though currently has wider coverage. SATCOM is offered by both Inmarsat and Iridium, which offer a range of packages depending on the use. ACARS messages are ASCII-based and are handled by a network provider, which maintains the network infrastructure and access to it. Two main providers exist – SITA and Rockwell Collins.

B. SSR and ADS-B

Secondary Surveillance Radar (SSR) is a cooperative ATC technology currently based on the so-called transponder Modes A, C, and S, which provide digital target information unlike traditional analog primary radar (PSR) [5]. Aircraft transponders

are interrogated on the 1030 MHz frequency and reply with the desired information on the 1090 MHz channel, as shown in Figure 1. With the newer Automatic Dependent Surveillance-Broadcast (ADS-B) protocol (see Figure 1), aircraft regularly broadcast their own identity, position, velocity, and additional information such as intent, status, or emergency codes. These broadcasts do not require interrogation; position and velocity are automatically transmitted twice a second [6].

C. Relationship to other ATC Technologies

Both ADS-B/SSR and ACARS are digital technologies, which send aircraft identification data (either the ICAO address, a registration, or both) with every message, enabling surveillance and data collection on a large scale. As security was not part of the design of these systems, neither includes any cryptography which could provide confidentiality for their users.

A large part of civil ATC is conducted with analog technologies such as traditional voice communication on the VHF band. It should be noted that the features used in this work could also be obtained through analyzing such analog communication (e.g., using automatic speech recognition [7]). However, focusing on unencrypted digital technologies has the key advantage of worldwide scalability, with easy manipulation and reliable extraction of relevant information using existing crowdsourced infrastructure.

D. Aircraft Identifiers in ATC Communication

A 24-bit address assigned by the International Civil Aviation Organization (ICAO) to every aircraft is transmitted via both ADS-B/SSR and partly on ACARS (on the SATCOM/VDLm2 data links). This identifier is different to an aircraft squawk or callsign. Squawks, of which there are only 4096, are allocated locally by ATC and are not useful for continuous tracking. The callsign can be set separately through the flight deck for every flight, and can include both letters and numbers. Callsigns of private aircraft typically consist of the aircraft registration number, commercial airliners use the flight number, and military and government operators often use special call signs depending on their mission.

In contrast, the ICAO identifier is unique providing address space for 16 million assignments, and enables the continuous tracking of the movements of particular aircraft; while the transponder can be re-programmed by engineers, the identifier is not easily (or legally) changed by the pilot. These characteristics make the ICAO identifier ideal for continuous tracking over a prolonged period of time.

E. Related Work

Open source information has been enjoying increased popularity, including by private and public intelligence services, which use it for OSINT purposes [8]. Much of the related OSINT literature concentrates on social media and the wider Internet as a source for information [9], [10]. To the best of our knowledge, no academic work has examined the true effect of wireless ATC communication for this purpose. However, the authors in [11] recently analyzed the current state of the transponder equipment of a sample of military and state aircraft, which is a pre-requisite for the present work. Similarly, several works have examined the state of privacy in aviation communication and highlighted the fundamental lack of confidentiality within the ADS-B and ACARS protocols [2], [12]–[15].

This is not limited to aviation; ships of various size and purpose use Automatic Identification System (AIS) to report their position in a similar way to ADS-B. AIS also suffers from basic security problems, much like ADS-B [16]. In recent years, its clear-text broadcast nature has been used to track illegal fishing [17] or monitor oil movements around the world [18].

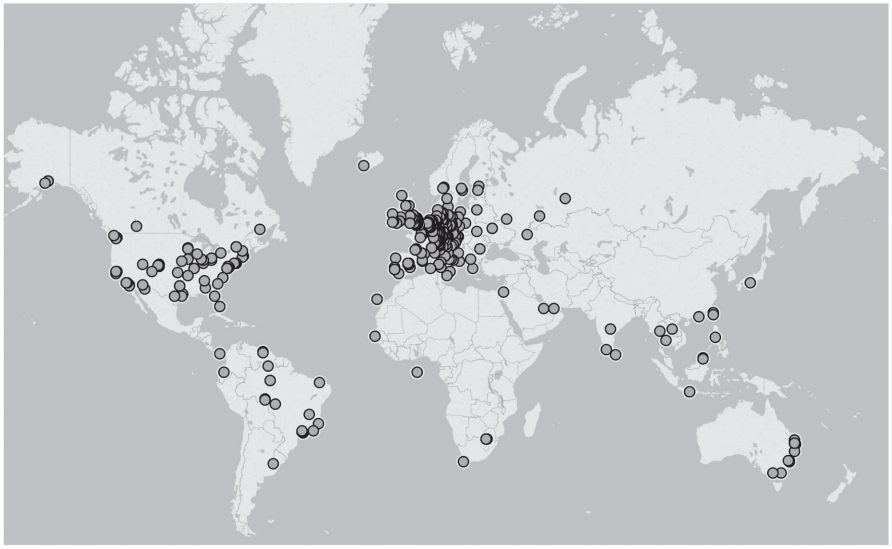
3. OVERVIEW OF PUBLICLY AVAILABLE AVIATION DATA

In this section, we present the data collection process. We first discuss the OpenSky Network as a representative example of a global sensor network available to passive threat actors. Following this, we analyze the potential sources from which to obtain metadata information about the observed aircraft. Finally, we illustrate the dataset that we use for our analysis in this paper.

A. The OpenSky Network

OpenSky is a crowdsourced network which is used as proof-of-concept for our OSINT collection. As of January 2018, the OpenSky Network consisted of 590 registered and about 450 anonymous sensors streaming data to its servers. Registered sensors are those operated by active members of the OpenSky Network community, and the operators of anonymous sensors are unknown. The network has currently received and stored over 4 trillion ATC messages, adding over 15 billion messages by more than 50,000 different aircraft every day.

FIGURE 2: A MAP OF SENSORS REGISTERED TO THE OPENSKY NETWORK (JANUARY 2018).



B. Public Metadata Sources

Besides the pure movement data, we require metadata about the aircraft to contextualize their behavior for OSINT purposes. We discuss the available sources of aircraft and airport metadata below.

1) Aircraft Metadata

Several public data sources exist which provide aircraft meta-information based on different identifiers. These identifiers include aircraft registration or the unique 24-bit ICAO Mode S transponder address. The data usually includes type and the owner or operator, which can then be used for further in-depth analysis and stakeholder identification. We used several of these third-party databases in our analysis of aircraft metadata:

- The plane spotting and aviation community actively maintains and shares database files with spotted aircraft using the BaseStation format for this [19].
- Junzi Sun maintains a database of aircraft seen on Flightradar24. The version used in this work is of 24 months and amounting to 136,637 rows [20].
- Aircraft registered in the US are logged on a daily-updated FAA database containing owner records. This is online and available for download, but excludes any sensitive owner information. Even so, the data set used for this work contained 312,162 records in December 2017 [21].

Besides these offline databases, which amounted to data of more than 2 million aircraft, we used several online sources to identify and verify aircraft as being operated by the government and military. These sources include the two major private flight tracking websites FlightAware [22] and Flightradar24 [23] and the popular database website airframes.org. Further leads and insights on more obscure aircraft identifications can also be gained on social media (Twitter, Flickr), a Wikipedia article on the topic [24], specialized aviation forums and aircraft photo websites such as JetPhotos [25].

2) Airport Metadata

To relate the actual destinations (countries and cities) of the tracked aircraft, we obtained the open airport database from Openflights.org [26]. As of December 2017, it contained 12,057 different airports around the globe, including name, ICAO and IATA (International Air Transport Association) short codes and precise location.

C. Overview of the Analyzed Datasets

For our work, we created two ADS-B datasets for further analysis, one for government aircraft and one for military aircraft. For government movements, we looked at a period of one year from 1 July 2016 to 30 June 2017, while for the significantly larger military dataset, we considered the period of one month in April 2017 for a more straightforward analysis. Regarding the ACARS data, we were able to obtain separate datasets for the three data links spanning 9 months in total, which we combined to analyze both government and military aircraft together.

1) Government Aircraft Movements

Using the public data sources described above, we created a list of 590 verified government aircraft from 113 different states. Table 1 shows the distributions of these aircraft and their operating governments per world region and whether OpenSky has tracked their position using ADS-B in the observed time frame of one year.

TABLE 1: OVERVIEW OF KNOWN AND TRACKED GOVERNMENTS IN THE DATASET.

	Europe	Americas	Africa	Asia	Oceania	Mid. East
A/C	172	78	119	79	8	134
Tracked A/C	157	73	76	66	7	113
Gov's	33	14	33	18	3	12
Tracked Gov's	33	13	30	16	3	11
Flights	8,915	1,775	399	706	248	2,115

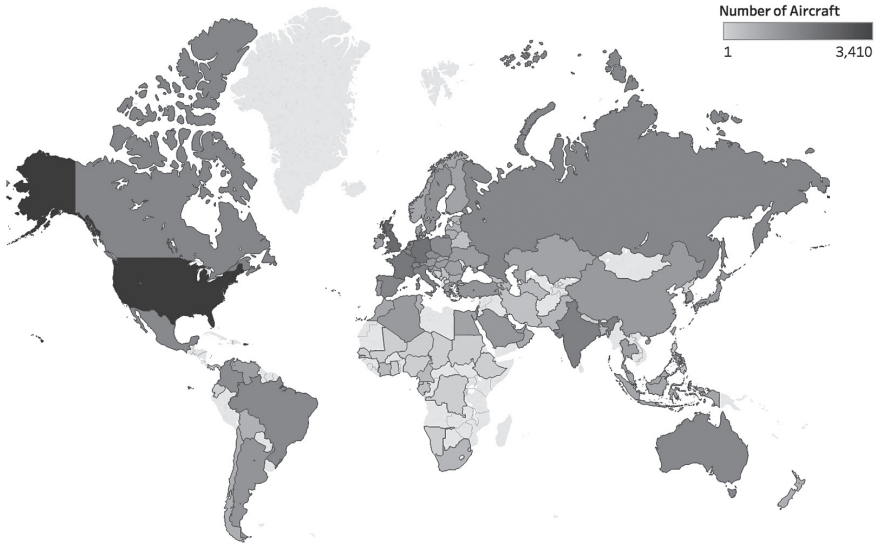
2) Military Aircraft Movements

Unlike government aircraft, military aircraft are not limited to those contained in the

public data sources. Air forces typically reserve a block in their country's ICAO ID range for military transponders; for example, identifiers used by the US Air Force tend to begin with 'AE'. Any aircraft with an ICAO ID matching this pattern can be identified as being used for military purposes. Exploiting this information, we can identify aircraft not in our public metadata sources – including the country and operator – though in these instances we lack additional meta information such as aircraft type. Overall, this approach resulted in a list of about 520,000 potential military aircraft transponder IDs.

In order to analyze the movements of military aircraft, we combined this list with all 1090 MHz downlink transponder transmissions recorded by OpenSky in April 2017. In this set of about 290 billion transmissions, we detected 6,024 unique military aircraft that broadcast unencrypted Mode S or ADS-B messages within range. Figure 3 shows the distribution of countries these aircraft were registered to.

FIGURE 3: DISTRIBUTION OF MILITARY AIRCRAFT SEEN IN OPENSKY BY ORIGIN COUNTRIES (APRIL 2017).



3) ACARS Collection

We further used the data from an ACARS receiver set up for the OpenSky Network in Central Europe, which collected 2,760,141 messages from 9,924 different aircraft on three data links (SATCOM, POA and VDLm2) over a period of 2 months for SATCOM and 7 months for VHF and VDLm2. While this ACARS data is not currently open source, there are existing platforms such as AVDelphi [27] which make such ACARS data publicly available.

In this dataset, we received 6,149 ACARS messages sent by 200 unique government aircraft and 24,923 messages sent by 438 aircraft operated by the military. The majority of messages from these groups were received via SATCOM (60% for the government and 97% for the military), indicating a strong preference for this data link.

4. THREAT MODEL

We consider a purely passive attacker as described in [14]. In our model, these are interested observers who exploit the open nature of air traffic communication protocols to obtain open source intelligence. This threat actor does not actively interfere with any of the observed technologies. Instead, they use public tracking services such as FlightRadar24 or ADS-B Exchange [28] in conjunction with public metadata sources to gather intelligence about government or military aviation movements. A more powerful version of this threat actor uses their own network of cheap SDR receivers to gather an unfiltered air traffic picture in real time which can be stored for historic analysis. This enables them to listen to a wider range of technologies such as ACARS and is within the capabilities of practically any determined attacker today [2].

5. EXPLOITING OPEN SOURCE ATC DATA FOR INTELLIGENCE PURPOSES

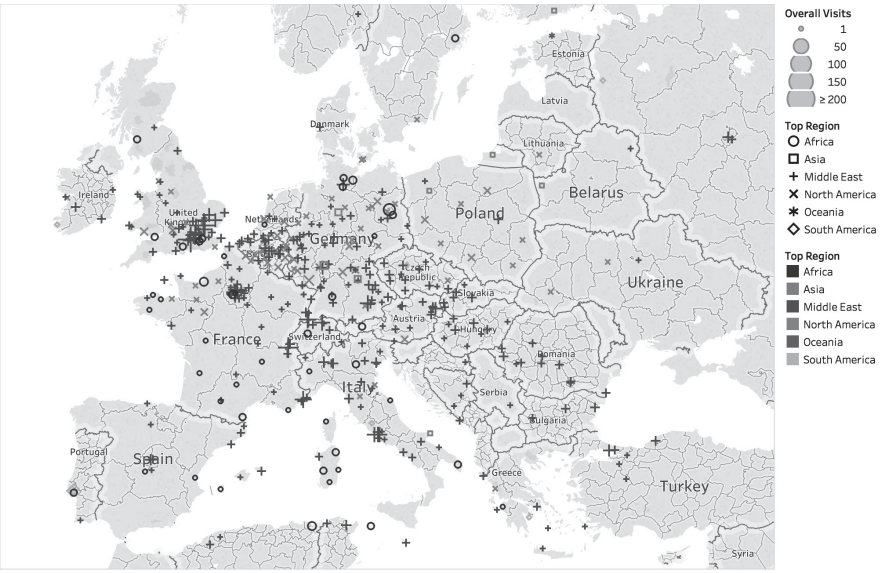
In this section, we provide examples of the type and scope of intelligence that can be gleaned from ATC data. We first discuss the government dataset, followed by the military dataset and an exemplary case study of a government jet operated by the military.

A. Government

We assume that governments are less secretive by nature than the military. At least in democratic countries, the electorate should be able to hold the government accountable, which requires an element of transparency. Whilst there are instances in which government transport might need to be kept private momentarily, most day-to-day government operations may not be secret in order to provide said accountability. However, this is evidently not true for all government missions from all countries. Thus, in the following, we analyze the quantitative possibilities a passive observer has with regards to the tracking of government aircraft.[1] Figure 4 illustrates the scope of our observations by showing the number and distributions of non-European government aircraft in Europe during the observation period.

1 Analyzing the reasons and motivations for specific relationships and government movements is out of the scope of this paper.

FIGURE 4: AIRCRAFT USED BY NON-EUROPEAN GOVERNMENTS VISITING EUROPE DURING JULY 2016 – JUNE 2017.



1) Meetings

During the one-year observation period, we observed 164 meetings of groups of at least three aircraft from different governments at the same destination.[2] As would be expected, the majority of these meet-ups happened at the major European capitals: Paris (44 times), Brussels (23), Rome (10), London (9), and Berlin (8).

The largest meetings with the most participants are naturally large global summits, such as the World Economic Forum (21 tracked governments), the Nuclear Security Summit (20), or the Munich Security Council (13). While these gatherings are not secret, their list of participants is not always published, and if it is, it may not be complete. Indeed, we found several government aircraft which landed in the vicinity of the World Economic Forum that were absent from the official list of participants [29].

While large multinational meetings such as the EU or NATO summits are well known, most smaller gatherings of three or four countries are not easily attributable. We acknowledge that every such occurrence may be due to simple chance, however, they can provide a heuristic starting point for further investigations.

2) Relationships

While there is a possibility of coincidence for every time that government aircraft are in the same location, this becomes much less likely for the consistently high

2 We define a potential multilateral meeting as three or more aircraft, which have landed within 50 km range within the same 48h period and not left again.

numbers of meetings we have seen over a prolonged time frame for many government pairs. Table 2 shows the top relationships between all tracked government aircraft in OpenSky's sensor range. The top three relationships have seen two governments at the same airport for 133 times (France/Saudi Arabia), 127 times (France/Morocco), and 102 times (Dubai/Qatar), respectively. Overall, we detected 7,106 pairwise meetings over 994 different relationships with a median of 3 meetings/relationship.

TABLE 2: RELATIONSHIPS BETWEEN MOST SEEN GOVERNMENTS BASED ON ADS-B DATA.
Note: We counted the Emirates of Dubai and Abu Dhabi as separate entities due to their prevalence.

	Qatar	Saudi Arabia	US	UK	Nether-lands	Morocco	Total
France	65	133	4	4	13	127	346
Germany	35	19	91	20	76	10	251
Dubai	102	23	17	71	9	2	224
Belgium	9	6	38	32	72	-	157
Bahrain	49	16	11	46	5	8	135
Abu Dhabi	28	40	33	13	2	13	129
Total	288	237	194	186	177	160	

Besides looking at the spatio-temporal correlation of two or more government aircraft, we can also investigate the most popular destinations of any single aircraft over time to infer public or private relationships of the operator. Table 3 lists the most visited destinations by the top eight observed governments. Considering OpenSky's core coverage area in Europe and the US, it is unsurprising that the most observed government aircraft are those from European countries and the US. Their preferred foreign destinations reflect the close diplomatic ties between these countries, or special commitments as in the case of Slovakia's EU presidency (Jul-Dec 2016), which necessitated a large amount of flights to the EU's headquarters in Brussels.

TABLE 3: MOST POPULAR NON-DOMESTIC DESTINATION COUNTRIES AND AIRPORTS OF THE EIGHT MOST SEEN GOVERNMENTS.
Note: Numbers in brackets indicate the number of times an aircraft was observed visiting the destination. Note, that country and airport are measured separately and can be unrelated.

Government (seen)	Top Destination Country	Top Destination Airport
Germany (2,345)	United States (57)	Washington (44)
United States (1,221)	Germany (48)	Brussels (9)
Russia (972)	Germany (54)	Rome (16)
Italy (740)	Belgium (17)	Brussels (15)
France (717)	Germany (19)	Basel (9)
Qatar (554)	United Kingdom (148)	London (75)
Czech Republic (536)	Germany (28)	Brussels (8)
Slovakia (472)	Belgium (39)	Brussels (32)

3) ACARS Analysis

Of the government aircraft considered in this section, 29.9% were observed sending ACARS messages. This in turn means that they often leak both their existence (their identification) and their intent (where they are going).

In Table 4 we see the position leakage for government aircraft as a result of using ACARS across the different subnetworks. Explicit position is simply a set of coordinates, whereas indicated position is when the aircraft is sending messages which reveal the area it is in. These could be airport information requests, for example. Note that we see at least 20% of government aircraft leak indicated position leakage on each link. Some of these aircraft were observed transmitting clear text e-mail messages via the ACARS satellite link. The nature of these messages was mainly flight status related, but some included names and e-mail addresses of fleet operators or government employees.

TABLE 4: POSITION-RELATED MESSAGES SENT OVER ACARS BY GOVERNMENT AIRCRAFT (AC). PERCENTAGES ARE OF ALL GOVERNMENT AIRCRAFT SEEN ON THAT SUB-NETWORK.

Sub-network	Number of Messages	Number of Aircraft	Explicit Position	Number of Aircraft	Indicated Position	Number of Aircraft
POA	1,491	66	169	26 (39.4%)	47	15 (22.7%)
VDLm2	275	54	31	13 (24.1%)	11	11 (20.4%)
SATCOM	3,654	117	218	13 (11.1%)	480	41 (35.0%)

B. Military

Compared to the identified government aircraft, military aircraft are much less likely to be equipped with ADS-B. Nonetheless, of the 6,024 unique military aircraft observed in April 2017, 42.9% were equipped with ADS-B and broadcast their positions at least some of the time. This varies greatly between different aircraft categories and also between countries as previous research has shown [5], [11]. Compared to the government aircraft, clusters of military aircraft on the ground are not as obviously insightful to an observer, as most operational missions are normally airborne and do not require landing. Yet, visits to foreign countries are interesting nonetheless and can support analyzes of military strategy and troop movements.

To prove that valuable OSINT can be collected on military aircraft, we offer some additional approaches: we analyze the ACARS messages sent by these aircraft and also look at the prevalence of military UAV movements in the dataset.

1) ACARS Analysis

Of all military aircraft we investigated, we observed 462 or 7.7% sending ACARS messages. Table 5 shows the distribution of these messages by subnetwork. It

illustrates that satellite communication is by far the most popular data link, making up about 98% of all traffic received by aircraft of this category. One might speculate that this preference indicates concern about the operational security of the ground-based links; however, the difficulty of eavesdropping on SATCOM with software-defined radios is broadly similar in practice.

As can be seen, 118 of the observed 462 military aircraft explicitly sent their position in the clear using ACARS at least once. Furthermore, 269 aircraft broadcast data that would give away their position by, for example, requesting weather reports for their destination airport.

TABLE 5: POSITION-RELATED MESSAGES SENT OVER ACARS BY MILITARY AIRCRAFT (AC). PERCENTAGES ARE OF ALL MILITARY AIRCRAFT SEEN ON THAT SUB-NETWORK.

Sub-network	Number of Messages	Number of Aircraft	Explicit Position	Number of Aircraft	Indicated Position	Number of Aircraft
POA	305	19	19	6 (31.6%)	26	7 (36.8%)
VDLm2	165	25	25	3 (12.0%)	9	4 (16.0%)
SATCOM	24,124	418	1,183	109 (26.1%)	2,011	258 (61.7%)

2) UAV Detection

Unmanned Aerial Vehicles (UAV) are fast becoming a major presence in civil airspace, and many UAVs are operated by governments or the military. Some of these drones carry ADS-B or Mode S transponders to cooperate with ATC and detect and avoid other aircraft. Hence, their presence and movements are visible to flight trackers and ATC receivers in general.

Using the metadata described in Section 3, we obtained the identifiers of 74 military-operated UAVs. We analyzed the complete historical data of OpenSky to find evidence of these Mode S and ADS-B-equipped UAVs, which returned sightings for 31 or 41.9% of the complete set.

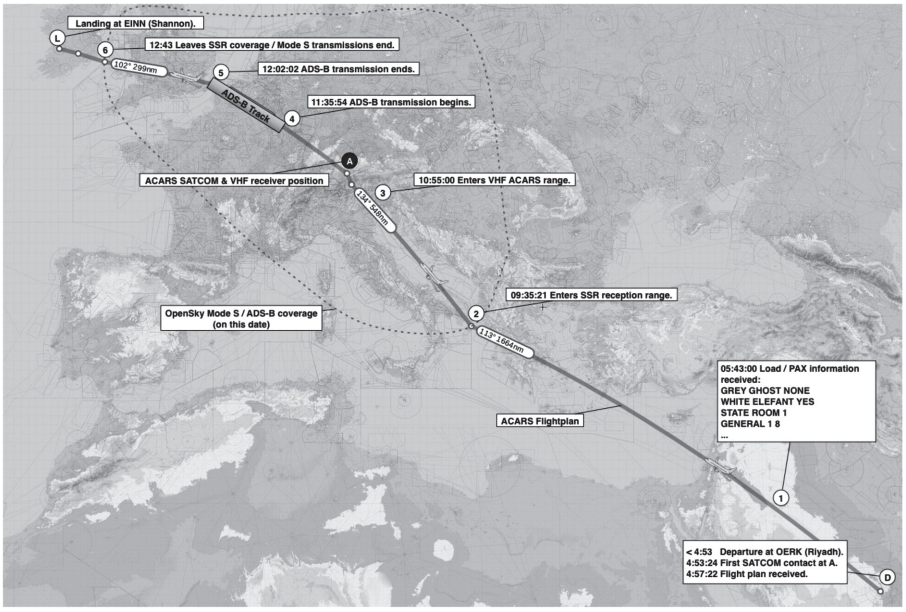
“ADS 95 Ranger Drones” operated by the Swiss Air Force to patrol borders and for general surveillance purposes provided the most striking evidence of such UAVs. Overall, we encountered messages from 14 of these drones, which use Mode S to communicate their identification and altitude.

Additionally, we received ATC messages from four General Atomics MQ-9 Reaper UAV and 10 Northrop Grumman RQ-4 Global Hawks. Some of these sightings have also been reported in aviation and military blogs on the Internet, showing that gathering OSINT by eavesdropping on air traffic communication is becoming more and more widespread [30].

C. Case Study

Figure 5 provides a case study on typical OSINT that can be gleaned from a government flight operated by a country's air force. It illustrates that, even with limited sensor coverage, the pieces put together via different technologies can provide a detailed picture of the whole flight.

FIGURE 5: A CASE STUDY OF OPEN SOURCE FLIGHT INFORMATION OBTAINABLE ABOUT A GOVERNMENT FLIGHT.



At the time of flight in December 2016, the OpenSky Network had comprehensive ADS-B and SSR coverage in the area within the dotted red line. A satellite ACARS receiver was placed centrally within this area, which was able to pick up the uplink part of the satellite communication; i.e., the one sent out by aircraft and addressed to the ground network.

Figure 5 shows the complete flight from the departure (D) in Riyadh to the landing (L) in Shannon. Around departure, the flight plan was sent out via ACARS by the aircraft and picked up by the receiver in Europe, detailing the precise route and waypoints the aircraft was planning to take. Several other ACARS messages containing potentially sensitive information about load and passengers were also picked up within an hour of departure (1). At (2), the aircraft reached the ground sensor coverage of OpenSky, which received 18,348 messages, providing the altitude of the aircraft and positional

information within the range of the receivers. At (3) it entered the range of the ground ACARS receiver, which captured all information provided via this channel only. While still at cruising altitude between (4) and (5), the aircraft activated its ADS-B transponder, broadcasting its exact position, call sign and velocity. It switched off the positional broadcasts again before leaving OpenSky's SSR range at (6) during the approach to Shannon (as verified by the Mode S altitude messages).

This behavior shows that ADS-B can and is turned on and off by military-operated aircraft. Turning it on at least sometimes indicates a general willingness to use ADS-B and, by doing so, facilitate tracking with civil surveillance technologies. However, turning it on only at cruising altitude and turning it off again before descending most likely aims at concealing the airport of departure and/or arrival.

6. EXISTING MITIGATION OPTIONS

There are several potential mitigation options for both government and military aircraft to prevent the information leakages discussed in the previous section. Here, we analyze the effectiveness of blocking information from web trackers, the use of pseudonyms, encryption, and attempts at forgoing civil ATC communication completely.

A. Web Tracker Blocking

One approach to limiting the privacy leaks of aircraft tracking is through block lists, which instruct the companies operating aircraft tracking websites to hide the aircraft on the list from public view. The most popular example of such a list is the Blocked Aircraft Registration Request (BARR) program, originally run by the National Business Aviation Association (NBAA) but now maintained by the FAA [31]. A BARR block places a restriction on the feed of aircraft send out by the FAA, which is used as a source by flight trackers. Table 6 shows that in our sample 85.0% of all military aircraft and 61.6% of all government aircraft were being filtered on the most popular flight tracking website (FlightRadar24). This indicates a clear awareness of a privacy impact through flight tracking by a majority of these state actors.

TABLE 6: PERCENTAGE OF IDENTIFIABLE MILITARY AND GOVERNMENT AIRCRAFT BLOCKED FROM POPULAR WEB TRACKERS. PERCENTAGES ARE OF THE NUMBER OF AIRCRAFT TRACKED.

		Europe	Americas	Africa	Asia	Oceania	Mid. East
Gov.	Tracked	157	73	76	66	7	113
	Blocked	93 (59.2%)	61 (83.6%)	38 (50.0%)	31 (47.0%)	6 (85.7%)	74 (65.5%)
Mil.	Tracked	1,851	3,646	45	268	73	78
	Blocked	1,359 (73.4%)	3,418 (93.7%)	36 (80.0%)	157 (58.6%)	38 (52.1%)	56 (71.8%)

Despite the popularity of the blocking approach, it is wholly ineffective against our threat model. As illustrated in the previous section, any passive actor with control over the raw data obtained from ATC sensors has full access to an unfiltered view of the airspace, including any government and military aircraft. Yet, for unknown reasons, 18 of all 106 tracked governments (17%) do not ask any of their aircraft to be blocked, forgoing even these basic mitigations.

B. Pseudonyms

A more comprehensive solution to the described tracking problem consists of pseudonymous identifiers that thwart an attacker's ability to correlate flight tracks with each other and with a specific aircraft.

For aircraft call signs, this is generally feasible for all considered technologies; changing a call sign before or during a flight is technically straightforward and often legally possible. For example, there are online services such as FltPlan.com [32], which offer randomized call signs to private operators, and both commercial and military operators are known to change their call signs regularly depending on an aircraft's mission. For the ICAO 24-bit identifier, the case is very different, as the pilot or operator cannot easily change it. The ICAO allows for a manual change in case of sensitive missions [33], yet we do not see this option in wide operation by government or military aircraft as our results in the previous section show.

ADS-B can alternatively be served over a newly developed data link, the Universal Access Transceiver (UAT), which offers a built-in privacy mechanism that generates a non-conflicting, random, temporary ICAO 24-bit identifier to avoid third-party tracking. However, it has been shown that this implementation is flawed and does not successfully disable aircraft tracking over time [34]. Furthermore, it is only in use by general aviation aircraft within the US airspace and as such not a quick fix for any other operator.

Finally, regardless of identifier, it has been shown that it is possible to fingerprint ADS-B transponders on the physical and link layer levels, which, in sufficient granularity, would circumvent even properly implemented pseudonyms [35].

C. Encryption

As mentioned previously, the use of encrypted communication would be the most effective countermeasure to the described data leakages. Unauthorized access to both movement data and other information can be stopped through the use of symmetric or asymmetric encryption as it is in current use in many wireless communication technologies.

As with any distributed security solution, implementing a public-key infrastructure is costly and requires thoughtful, security-conscious design. Especially in the case of aircraft, which must be able to communicate with unexpected ground stations, keeping credentials up-to-date for all communications partners is a challenge. Secure ACARS, available since 2001 [36], provides such an option, and not only to military and government operators. However, we have not seen Secure ACARS in use in the wild; in our data set of 1,749,142 messages from all three data links, we never recorded a single message of this type.[3] We speculate that the fact that it comes at a surcharge to the ACARS service impedes its adoption.

This assumption is supported by the fact that there are several proprietary encryption solutions in use for ACARS, which are not standardized, but potentially come at a cheaper running cost. Unfortunately, many such solutions are insecure, quickly broken and provide no more security than clear-text messages against any interested adversary. One such example is discussed by Smith et al. [12], who show that it is in wide use even in government and military aircraft. In our dataset, we found that 1.78% of the observed military and 11.36% of the observed government aircraft used this obfuscation method, a serious lapse of operational security. In principle, however, there is no fundamental obstacle to developing a secure proprietary ACARS solution for exclusive use by a state's sensitive aircraft as long as compatibility with the existing system is ensured.

While ACARS messages can be encrypted by the user's choice, this is not possible for both ADS-B and SSR. As has been analyzed previously, the current technological lock-in does not allow for a quick encryption solution for these protocols [15]. While there are military equivalents to civil SSR and ADS-B in use and under development (NATO STANAG 4193, SSR Modes 4 and 5), due to obvious secrecy requirements, very few details are publicly available. As Mode 5 is believed to provide full confidentiality using strong encryption, its use would indeed fully mitigate the information leakage of ATC movement data. However, due to the lack of independent scrutiny, it is not possible to make any reliable statements on the security of the system.

Unfortunately, even for those military operators with access to encrypted protocols, the preference of civil ATC authorities for open systems and maximum compatibility precludes any proprietary solutions as long as they are flying in civil airspace [14]. In short, all operators must be aware that using any current civil ATC technology will leak information immediately and widely.

D. Switch off civil ATC communication

The final mitigation option for military and government aircraft operators is to not use civil ATC communications. For ACARS, this is fairly simple, as it is not a required

[3] A distinct set of message labels is reserved in the ACARS standard for Secure ACARS messages, enabling us to detect their presence even where it is not possible to decrypt them.

technology in controlled airspace and some operators choose to forgo ACARS for cost reasons, including entire airlines. Yet, as shown above, many sensitive aircraft use unencrypted ACARS, presumably for operational reasons.

When considering ADS-B and SSR, the picture is much more complex. Aircraft are still not required to broadcast their precise position using ADS-B. As long as the technology is not mandated for state aircraft in (mostly Western) civil airspaces, there are many operators who choose to delay the upgrade in the first place for reasons of cost, convenience, or indeed privacy. Overall, only around 6.7% of all government aircraft but 57.1% of the military aircraft in our sample did not yet use ADS-B, which is in line with previous research [11]. Naturally, this is only a solution in the very short term and the consequences of upgrading will have to be addressed in the very near future.

7. DISCUSSION

We have demonstrated that tracking aircraft using civil ATC systems allows us to glean significant intelligence that the aircraft operators or users might not be interested in sharing. Indeed, with a relatively low level of skill and equipment used by a purely passive attacker, this combination of public data sources can reveal much more than where an aircraft is. Even though options exist to mitigate the problem, they are largely ineffective against a reasonably persistent attacker. Naturally, this generates some recommendations for how to improve the state of privacy in aviation. In the short term, regulation provides a possible key to allowing relevant actors to protect their privacy. Governments would have to legally restrict and regulate those entities (private and commercial) that are sharing data about aircraft movements for which a reasonable effort at privacy has been made. This would need to be a more concerted effort than the BARR scheme, which is, to some extent, opt-in.

In the longer term, technical solutions should be developed to provide guarantees of privacy. For example, a robust pseudonym system would go a long way to limiting the ability to track aircraft over time, similar to the concept of Temporary Mobile Subscriber Identity (TMSI) in cellular networks. There is no critical technical or procedural need to have a consistent, publicly known identifier for aircraft — there is in fact evidence of aircraft being prescribed alternative ICAO identifiers by the authorities in situations such as sensitive military flights [33]. Doing away with the inflexible current system in favor of a more transient one would in turn de-correlate consecutive flights by a given aircraft. This measure alone would greatly reduce the impact of ATC-based flight tracking.

Hence, in our opinion, the only way to effectively create the opportunity for privacy in ATC systems is through the combination of technical and regulatory measures. Regulatory measures can cover the case of data generated by state entities, but technical measures are needed to stop passive observers from easily collecting significant amounts of data.

As discussed in [14], there is currently a preference for open systems in aviation, but this is not necessarily wise if a good level of security and privacy is required. Parallels can be drawn to the creation of the Internet in that, initially, open systems allowed easy integration and global interaction between different networks. However, in the longer term, malicious parties have resulted in both a desire and need for securing all communications. Aviation networks carry bigger safety risk, so should aim for similar, if not greater, levels of security than the Internet currently uses.

8. CONCLUSION

The findings we have presented in this work conclusively prove that it is possible to collect, process, and ultimately exploit, a trove of open source air traffic communication data for intelligence purposes. While examining all potential use cases for such data is out of the scope of a single paper, we believe that our proof of concept is sufficient to raise awareness of the issue among all concerned stakeholders.

It has also become clear that traditional ways of protecting the privacy of aircraft owners are all but obsolete in the era of cheap software-defined radio receivers, and relying on them should be done with extreme caution. Military and nation state actors have superior means and resources to protect their operational privacy and security in some cases, as evidenced by the existence of encrypted communications solutions. However, the requirement to be able to communicate with civil ATC negates at least some of this advantage as illustrated in this work. Consequently, only a change to those civil communication technologies will lead to comprehensive privacy improvements for those who seek it. In the meantime, many actors will be able to exploit the openly available information gained in this domain for their purposes.

REFERENCES

[1] M. Strohmeier, M. Smith, V. Lenders, and I. Martinovic, “The Real First Class? Inferring Confidential Corporate Mergers and Government Relations from Air Traffic Communication,” in *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

[2] M. Smith, D. Moser, M. Strohmeier, V. Lenders, and I. Martinovic, “Analyzing Privacy Breaches in the Aircraft Communications Addressing and Reporting System (ACARS),” no. arXiv:1705.07065v1 [cs.CR], 2017.

[3] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, “Bringing up OpenSky: A large-scale ADS-B sensor network for research,” in *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (IPSN)*, 2014, pp. 83–94.
[4] R. T. Oishi and A. Heinke, “Air-Ground Communication,” in *Digital Avionics Handbook*, Third., C. R. Spitzer, U. Ferrell, and T. Ferrell, Eds. Taylor & Francis Group, 2015, p. 2.1-2.3.
[5] C. R. Spitzer, U. Ferrell, and T. Ferrell, *Digital Avionics Handbook*, 3rd ed. CRC Press, 2014.
[6] RTCA Inc., “DO-262 - Minimum Operational Performance Standards (MOPS) for 1090 MHz Automatic Dependent Surveillance - Broadcast (ADS-B).” 2000.
[7] D. Hoffman and S. Rezchikov, “Busting the BARR: Tracking ‘Untrackable’ Private Aircraft for Fun & Profit,” in DEF CON 20, 2012.
[8] R. Steele, “Open Source Intelligence,” in *Handbook of Intelligence Studies*, Routledge, 2007, pp. 129–147.
[9] D. Gritzalis and V. Stavrou, “Exploiting Open Source Intelligence capabilities for the benefit of the Hellenic Air Force,” in *4th Air Power Conference*, 2016.
[10] C. Weinbaum, S. Berner, and B. McClintock, “SIGINT for Anyone - The Growing Availability of Signals Intelligence in the Public Domain.” 2017.
[11] M. Schäfer, M. Strohmeier, M. Smith, M. Fuchs, V. Lenders, M. Liechti, and I. Martinovic, “OpenSky Report 2017 : Mode S and ADS-B Usage of Military and other State Aircraft,” in *Digitial Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th*, 2017.
[12] M. Smith, D. Moser, M. Strohmeier, V. Lenders, and I. Martinovic, “Economy Class Crypto: Exploring Weak Cipher Usage in Avionic Communications via ACARS,” in *21st International Conference on Financial Cryptography and Data Security*, 2017.
[13] K. Sampigethaya and R. Poovendran, “Security and privacy of future aircraft wireless communications with offboard systems,” in *Third International Conference on Communication Systems and Networks (COMSNETS 2011)*, 2011, pp. 1–6.
[14] M. Strohmeier, M. Smith, M. Schäfer, V. Lenders, and I. Martinovic, “Assessing the Impact of Aviation Security on Cyber Power,” in *8th International Conference on Cyber Conflict (CyCon)*, 2016, pp. 223–241.
[15] M. Strohmeier, V. Lenders, and I. Martinovic, “On the Security of the Automatic Dependent Surveillance-Broadcast Protocol,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 1066–1087, 2015.
[16] M. Balduzzi, K. Wilhoit, and A. Pasta, “A Security Evaluation of AIS,” 2014.
[17] Image Sat International (iSi), “Optimizing fish production with space intelligence,” 2017. [Online]. Available: https://www.imagesatintl.com/optimizing-fish-production-space-intelligence/. [Accessed: 18-Dec-2017].
[18] S. Madani and L. Ward, “TankerTrackers.com,” 2017. [Online]. Available: http://tankertrackers.com. [Accessed: 18-Dec-2017].
[19] D. Taylor, “Databases,” *Planeplotter*, 2016. .
[20] J. Sun, “World Aircraft Database,” 2017. [Online]. Available: http://junzisun.com/adb/. [Accessed: 11-Dec-2017].
[21] Federal Aviation Administration, “Aircraft Registry - Releasable Aircraft Database Download,” 2017. [Online]. Available: https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/. [Accessed: 11-Dec-2017].
[22] FlightAware, “FlightAware,” 2017. [Online]. Available: https://www.flightaware.com/. [Accessed: 06-Mar-2017].
[23] Flightradar24 AB, “Flightradar24,” 2017. [Online]. Available: https://www.flightradar24.com. [Accessed: 06-Mar-2017].
[24] “Air Transports of Heads of State and Government,” *Wikipedia*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Air_transports_of_heads_of_state_and_government. [Accessed: 11-Dec-2017].
[25] O. A. Saffe and S. De Rudder, “JetPhotos,” 2017. [Online]. Available: 2017-12-11.
[26] J. Patokallio, “OpenFlights.org,” 2017. [Online]. Available: https://openflights.org. [Accessed: 11-Dec-2017].
[27] D. R. Crocker, “AvDelphi,” 2018. [Online]. Available: https://www.avdelphi.com/. [Accessed: 06-Jan-2018].
[28] D. Streufert, “ADS-B Exchange,” 2017. [Online]. Available: https://www.adsbexchange.com/. [Accessed: 11-Dec-2017].
[29] World Economic Forum, “World Economic Forum Annual Meeting - List of Public Figures,” Davos-Klosters, 2017.
[30] D. Cenciotti, “U.S. Air Force RQ-4 Global Hawk drone flew over Ukraine with transponder turned on for everyone to see,” *The Aviationist*, 2016. [Online]. Available: https://theaviationist.com/2016/10/18/u-s-air-force-rq-4-global-hawk-drone-flew-over-ukraine-with-transponder-turned-on-for-everyone-to-see/. [Accessed: 18-Dec-2017].

[31] National Business Aviation Association, "Block Aircraft Registration Request (BARR) Program," 2011. [Online]. Available: https://www.nbaa.org/ops/security/barr/background/. [Accessed: 24-Oct-2017].
[32] FltPlan.com, "Flying in Private," 2017. [Online]. Available: https://flttrack.fltplan.com/FltPlanInfo/DCMCallSigns.htm. [Accessed: 13-Dec-2017].
[33] Directorate of Air Traffic Management, "Automatic Dependent Surveillance-Broadcast (ADS-B)," New Delhi, 2014.
[34] K. Sampigethaya, S. Taylor, and R. Poovendran, "Flight Privacy in the NextGen: Challenges and Opportunities." 2013.
[35] M. Leonardi, L. Di Gregorio, and D. Di Fausto, "Air Traffic Security: Aircraft Classification Using ADS-B Message's Phase-Pattern," *Aerospace*, vol. 4, no. 4, p. 51, 2017.
[36] A. Roy, "Secure aircraft communications addressing and reporting system (ACARS)," *20th Digital Avionics Systems Conference*, vol. 2, p. 7A2/1--7A2/11 vol.2, 2001.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

FeedRank: A Tamper-resistant Method for the Ranking of Cyber Threat Intelligence Feeds

Roland Meier
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Cornelia Scherrer
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
cornelia.scherrer@alumni.ethz.ch

David Gugelmann
Exeon Analytics
Zürich, Switzerland
david.gugelmann@exeon.ch

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Laurent Vanbever
Department of Information Technology and Electrical Engineering
ETH Zürich
Zürich, Switzerland
lvanbever@ethz.ch

Abstract: Organizations increasingly rely on cyber threat intelligence feeds to protect their infrastructure from attacks. These feeds typically list IP addresses or domains associated with malicious activities such as spreading malware or participating in a botnet. Today, there is a rich ecosystem of commercial and free cyber threat intelligence feeds, making it difficult, yet essential, for network defenders to quantify the quality and to select the optimal set of feeds to follow. Selecting too many or low-quality feeds results in many false alerts, while considering too few feeds increases the risk of missing relevant threats. Naïve individual metrics like size and update rate

give a somewhat good overview about a feed, but they do not allow conclusions about its quality and they can easily be manipulated by feed providers.

In this paper, we present FeedRank, a novel ranking approach for cyber threat intelligence feeds. In contrast to individual metrics, FeedRank is robust against tampering attempts by feed providers. FeedRank's key insight is to rank feeds according to the originality of their content and the reuse of entries by other feeds. Such correlations between feeds are modelled in a graph, which allows FeedRank to find temporal and spatial correlations without requiring any ground truth or an operator's feedback.

We illustrate FeedRank's usefulness with two characteristic examples: *(i)* selecting the best feeds that together contain as many distinct entries as possible; and *(ii)* selecting the best feeds that list new entries before they appear on other feeds. We evaluate FeedRank based on a large set of real feeds. The evaluation shows that FeedRank identifies dishonest feeds as outliers and that dishonest feeds do not achieve a better FeedRank score than the top-rated real feeds.

Keywords: *cyber threat intelligence, intelligence feeds, cyber attacks, malware, botnets, situational awareness*

1. INTRODUCTION

States, organizations, companies and individuals are faced with ever-growing cyber threats. The most prominent among these threats include phishing or spam campaigns, malware distribution and DDoS attacks [1] [2]. To mitigate these threats, Cyber Threat Intelligence Feeds (CTIFs, also known as blacklists or block lists) are a major source of information for most network defenders [3]. The CTIF ecosystem is currently very large and complex [4] and for reliable protection, network defenders need to correlate data from multiple CTIFs [1].

However, while selecting the best set of CTIFs is crucial to maximizing efficiency, it is also difficult as there is no easy way to objectively compare CTIFs. In fact, network defenders often only evaluate feeds individually based on naïve metrics such as the feed's size. While these metrics allow for a rough assessment, they do not allow conclusions about the combination of multiple feeds and – as we will show in this paper – they are easy to manipulate for a dishonest CTIF provider in order to pretend a higher quality and thus increase its impact and revenue.

Problem statement. In this paper, we address the problem of finding an objective, tamper-resistant ranking algorithm that allows well-grounded selections of high quality CTIFs. We determine the quality of a feed by three key properties: *completeness, accuracy* and *speed*. That is, an ideal CTIF lists all malicious entities, does not list non-malicious entities and updates its entries promptly. In particular, we address the following research questions:

- How can the quality of a CTIF be estimated in a robust and scalable way? Achieving this is challenging because there is no ground truth to compare it with. Hence, one cannot rely on standard metrics such as precision and recall.
- How can the structure of the CTIF ecosystem be evaluated and how do existing CTIFs differ in terms of completeness, accuracy and speed? In particular, do CTIF providers cluster in groups or is there a large diversity regarding the reported threats among the different providers?
- Can we identify individual CTIFs that consistently outperform others and CTIFs that seem to lack behind or borrow information from other feeds? Specially, what are good metrics to identify outperformers and tampering-attempts by a subset of the feeds?

FeedRank. We present FeedRank, a novel metric for the ranking of CTIFs. The key idea behind FeedRank is to model the correlations between CTIFs as a graph and to obtain the ranking by applying algorithms to this graph. This way, FeedRank quantifies the relative performance among CTIFs and is able to evaluate the quality of feeds without a ground truth. At its core, FeedRank bears similarities with collective intelligence approaches or PageRank [5], an algorithm to rank websites that is used by Google.

The setting for ranking CTIFs bears similarities with the ranking of websites by search engines in the following aspects:

- Websites can contain arbitrary content (including dummy keywords to improve their ranking).
- Websites can contain links to any other website.
- There is no ground truth for the quality of websites.
- A website to which many other websites refer to is likely to be important.

Similar properties hold for CTIFs:

- CTIFs can contain arbitrary entries.
- CTIFs can copy entries from any other CTIF.
- There is no ground truth for the quality and validity of CTIF entries.
- A CTIF whose entries appear in other CTIFs is likely to be of high quality.

Despite these similarities, applying website ranking algorithms (such as PageRank) to CTIFs is challenging because of the particular semantic of the CTIF application domain. The key idea to apply website ranking algorithms to CTIFs is to model correlations between CTIFs in a graph. In particular, while PageRank uses the web graph (a graph that models the links between websites), FeedRank uses a correlation graph to model common entries in CTIFs and the time at which they appear in each of the considered CTIFs. The correlation graph provides us with the foundation to assess a CTIF's quality as we argue that a CTIF whose entries later appear on many other feeds has a high quality (like a website with many incoming links is assumed to be important). However, since the correlation graph alone does not allow conclusions about the completeness of a particular CTIF, FeedRank also performs an analysis of the contribution of each CTIF.

Contributions. The main contributions of this paper are:

- A tamper-resistant approach to rank CTIFs at scale (Section 3) based on:
  - correlations between CTIFs (Section 3B); and
  - the individual contribution of each CTIF (Section 3C).
- A comprehensive evaluation based on large sets of freely available CTIFs (Section 4).
- Two case-studies to demonstrate useful use-cases of FeedRank (Section 5).

2. EVALUATING CYBER THREAT INTELLIGENCE FEEDS

In this section, we provide an overview over Cyber Threat Intelligence Feeds (CTIFs) and identify key properties that characterize good CTIFs, sketch traditional evaluation metrics and identify strategies for how dishonest CTIF providers can tamper with them.

A. Cyber Threat Intelligence Feeds

In general, CTIFs are collections of *Indicators of Compromise* (IOC) that characterize malicious or non-malicious endpoints or activities. In this paper, we focus on feeds that list IP addresses associated with malicious activities (such as sending spam or hosting phishing sites). However, the obtained results are also applicable to other types of feeds.

CTIFs are available from a variety of commercial and non-commercial providers and can cover one or multiple types of threats (e.g. spam or phishing). The feeds are typically provided in real time; that is, the contents are updated continuously or with a certain frequency. New entries may be added when, for example, an endpoint is

found to behave maliciously and removed if the malicious activity has stopped. CTIFs obtain information about malicious endpoints in various ways. For example, malicious activity can be detected by email providers, honeypots, CERTs or by manual reports from users. CTIFs can also copy or fuse information from other CTIFs.

B. Properties of High Quality Feeds

An ideal CTIF is complete, accurate and fast. To be *complete*, the CTIF needs to contain all malicious endpoints at a given time. To be *accurate*, the CTIF must not list benign endpoints. To be fast, the completeness and accuracy property must hold at any given point in time, i.e., an endpoint must appear exactly during the time it behaves maliciously. This ideal state is obviously difficult to reach in practice, as there always exist malicious endpoints that have not yet been identified as such.

C. Individual Feed Metrics

Naïve metrics which evaluate each CTIF individually are easy to calculate and widely used. Examples of such individual metrics include the feed's size, the update frequency and the number of entries that are added or removed (cf. Table I). However, a major problem with individual metrics is that they provide little insight about the quality of a CTIF without a ground truth (i.e. a way to objectively verify the correctness of the feed's contents). Even worse, all the listed individual metrics can easily be manipulated by adding or removing entries to/from a CTIF (cf. Table I).

With FeedRank, we present an advanced and tamper-resistant ranking metric that does not require a ground truth. As we will describe in the following sections, analyzing the correlations of CTIFs allows reasoning about the feed's completeness, accuracy and speed.

TABLE I: EXAMPLES OF INDIVIDUAL FEED METRICS. THESE METRICS DO NOT ALLOW CONCLUSIONS ABOUT A FEED'S QUALITY AND CAN BE MANIPULATED BY THE FEED PROVIDER.

Metric	Description	Manipulation strategy
Size	Number of entries in the CTIF	Add random entries
Insertion rate	Number of entries that are added to CTIF per time unit	Add random entries
Removal rate	Number of entries that are removed from a CTIF per time unit	Remove random entries
Update rate	Rate at which entries are added or removed	Frequently replace random entries

3. FEEDRANK

In this section, we present the design goals and an overview of FeedRank. Further, we describe the two core components of FeedRank in more detail and explain why FeedRank is robust against tampering attempts.

A. Overview

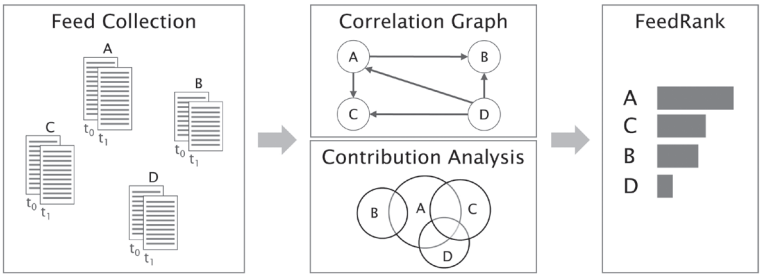
FeedRank allows us to identify high quality CTIFs, while at the same time being robust against tampering attempts from CTIF providers, by combining the contribution analysis and the correlation graph (see Table II).

TABLE II: KEY PROPERTIES OF HIGH QUALITY CTIFS AND HOW THEY ARE REPRESENTED IN FEEDRANK.

Property	Represented in
Completeness	Contribution analysis
Accuracy	Correlation graph (weighted edges model the entries that are confirmed by other CTIFs)
Speed	Correlation graph (directed edges between CTIFs represent the order in which common entries were listed)

FeedRank operates in three steps (see Figure 1). First, it collects snapshots of considered CTIFs; second it builds a feed correlation graph and performs a contribution analysis; and third, it computes a score for each considered CTIF.

FIGURE 1: FEEDRANK OPERATES IN THREE STEPS: IT COLLECTS SNAPSHOTS OF CTIFS, COMPUTES A CORRELATION GRAPH AND A CONTRIBUTION ANALYSIS AND OUTPUTS A RANKING.



I. Feed Collection. As an input, FeedRank requires at least two snapshots of each considered feed. A snapshot consists of the timestamp and all entries of a CTIF. For the most accurate results, the time between the two snapshots should be long enough such that all CTIFs provide an update of the entries. The set of considered feeds can

be specified by the operator who uses FeedRank. It should contain all the CTIFs that the operator considers using in their environment.

II. a) Correlation Graph. Based on the snapshots, FeedRank builds a correlation graph. The nodes in this graph correspond to the CTIFs and the (directed) edges represent correlations between them.

II. b) Contribution Analysis. For each CTIF, FeedRank computes a contribution metric that measures the CTIF's contribution to the total number of listed entries.

III. Feed Rating. FeedRank runs an algorithm similar to PageRank on the correlation graph. This, together with the results from the contribution analysis, assigns each feed a score and allows to rank them.

B. Correlation Graph

The correlation graph is used to model correlations between CTIFs. It is a directed graph where the vertices represent feeds and the weighted edges describe correlations between them. Two CTIFs (X and Y) are connected with a directed edge from X to Y if X contains entries that were contained in Y before they appeared in X. This means that X implicitly confirms the respective entries from Y. In other words: both feeds classify the entries as malicious, and Y was faster in listing them, which makes it more likely that Y is accurate with respect to these entries.

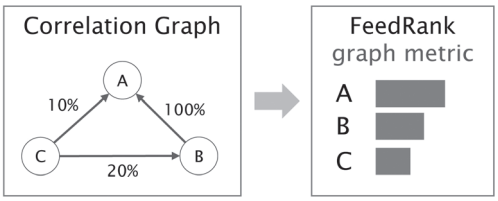
The weight of this edge is determined by the percentage of entries that appear first in Y and are later mentioned by X. For example, if Y contains 20 entries and 10 of them appear later on X, the weight of the edge would be 50% as this is the percentage of entries in Y that were confirmed by X.

Figure 2 illustrates an example of a correlation graph with 3 feeds with the following characteristics:

- B confirms 100% of the entries in A (i.e. every entry that appears in A later appears in B)
- C confirms 10% of the entries in A and 20% of the entries in B (i.e. 10% of the entries in A and 20% of the entries in B appear later in C)

In this example, feed A achieves the highest score according to the correlation graph and would thus be considered as the most valuable feed. The intuitive explanation for this is that all of A's entries are confirmed by B and no other feed is faster than A.

FIGURE 2: EXAMPLE OF THE FEEDRANK GRAPH METRIC. IT RANKS FEEDS ACCORDING TO THE AMOUNT OF ENTRIES THAT ARE CONFIRMED BY OTHER FEEDS (E.G. B CONFIRMS 100% OF THE ENTRIES IN A).



To determine a ranking of CTIFs in the correlation graph, we apply the PageRank algorithm [5]. PageRank is a ranking algorithm for websites (famously used by Google) and is based on a graph that models the hyperlinks between websites. Besides the web graph, PageRank requires two additional parameters: the damping factor and a convergence condition.

The damping factor d in PageRank describes the probability with which a user browsing at a certain website will click on any of the links to visit another website. For FeedRank, we calculate the damping factor depending on the average path length l (i.e., the average number of CTIFs that subsequently list an entry) of all entries that appear in at least two CTIFs within the analyzed dataset. From l, we calculate the probability that an entry "propagates" to another feed – along the lines of a user that clicks on a link to move to another website – as:

$$d = P(continue) = 1 - P(stop) = 1 - \frac{1}{l}$$

Being an iterative algorithm, PageRank further requires a convergence condition. The convergence condition in PageRank specifies the maximal delta between the graph score of all nodes (i.e. the precision of the result).

C. Contribution Analysis

The contribution metric is the result of the contribution analysis and measures the relative contribution of a single CTIF compared to the complete set of analyzed feeds. Therefore, it provides the foundation to select a subset of the analyzed feeds that together have a maximal contribution.

FeedRank's contribution analysis works as follows. First, it computes the complete set of all listed entries, i.e., the union of all entries listed in the considered feeds at any of the recorded snapshots. For each entry, it determines the feed that listed the entry first, and assigns the entry to that feed. In case multiple feeds add an entry at the same time, the entry is assigned to the biggest feed. The resulting contribution metric is then computed as the percentage of entries that each feed contributes to the complete set.

D. Tamper-resistance

In this section, we explain why FeedRank is robust against an unfair CTIF which tries to manipulate its rank. Since a CTIF provider can arbitrarily choose the contents of its feed, there are no guarantees about the validity of entries.

At a high level, we distinguish between the following tampering strategies:

- Adding entries that are not contained in the original CTIF.
- Removing of entries from the original CTIF.
- Replacing existing entries by other values (i.e. pretend updates).

For each of these strategies, the dishonest CTIF provider needs to choose the entries that will be added or removed. This can be done in at least the following ways:

- At random: New entries are generated randomly and randomly chosen entries are removed from the feed.
- Based on the contents of another CTIF: A dishonest CTIF can copy a subset or all entries from one or multiple other CTIFs and thus copy the behavior of these CTIFs.

The manipulation strategies mentioned above work well for individual metrics (as described in Section 2. C) but, as we explain in the following, they do not work with FeedRank.

A dishonest CTIF that tries to manipulate its FeedRank score by adding entries is not successful because: *(i)* if the added entries are chosen randomly, they will not be confirmed by other feeds with very high probability; *(ii)* if the added entries are copied from another CTIF, this is considered as if the dishonest feed confirms the other feed's entries and can therefore help the other feed, but not the dishonest feed. Obviously, a feed that copies entries from another feed is always slower in listing these entries. If a dishonest CTIF tries to improve its score by removing entries, each of the removed entries is either of high quality (i.e. it is confirmed by other feeds) or of low quality (it does not appear on other feeds). If a CTIF removes high quality entries, this lowers its ranking because a smaller percentage of its entries are confirmed. If it removes low quality entries, its overall quality increases and it (deservedly) obtains a better ranking.

A dishonest CTIF that both adds and removes entries faces the union of the limitations mentioned above. FeedRank does not measure the update frequency of CTIFs and therefore a higher update frequency does not help to improve the score. Instead, FeedRank is run with a certain frequency and based on the feed's contents at the time of execution. Therefore, as long as the update frequency of a CTIF is larger than or

equal to the execution frequency of FeedRank, increasing the update frequency does not change the FeedRank score.

While FeedRank is robust against a small percentage of dishonest CTIFs, it can be susceptible to manipulation attempts by many colluding CTIFs. However, this is hardly feasible in practice because it would require many CTIFs to become dishonest and it only works if the user considers all of them when running FeedRank (it is easy for a single entity to publish a large number of dishonest feeds, but it is unlikely that a user would consider all of them). Such a set of colluding CTIFs can be identified by doing basic cluster analysis based on the considered feeds and the correlation graph (we show this in Section 4B).

4. EVALUATION

In this section, we use real CTIFs to compare FeedRank with individual metrics and show its tamper-resistance. In the following subsections, we describe and visualize the dataset and show the evaluation results.

A. Dataset and Methodology

To evaluate FeedRank, we use both real CTIFs which we collected over a timespan of almost 12 days and synthetic CTIFs with which we simulated the impact of tampering strategies. Below, we provide more details about both types of feeds.

1) Collecting Real Feeds

For a comprehensive dataset, we fetched the feeds listed in Table III at regular intervals (60 min) during almost 12 days in 2017. In this way, we obtained 277 snapshots representing the activity of 27 feeds with a total of around 40 million entries. These snapshots allowed us to analyze correlations between feeds at a granularity of an hour. Some of the snapshots were incomplete because our collection infrastructure was unable to fetch them due to connectivity issues, database overload or rate limiting by the CTIF provider. The feed collection functionality was implemented in Python on top of the stix [6] and libtaxii [7] modules. The collected feeds were normalized and stored in an Elasticsearch database to facilitate analysis. We anonymized the feeds as it is not our goal to provide a ranking of particular feed providers, but to demonstrate the practicality of our algorithm.

TABLE III: EVALUATED FEEDS. WE ANALYZE 27 FREELY AVAILABLE CTIFS (LISTED IN ALPHABETICAL ORDER HERE).

Feed △	Source
AlienvaultReputationIP	reputation.alienvault.com
Autoshun	www.autoshun.org
BinaryDefense	www.binarydefense.com
CIArmyBadGuys	www.cinsscore.com
CymonBlacklist	www.cymon.io
CymonBotnet	www.cymon.io
CymonMaliciousActivity	www.cymon.io
CymonMalware	www.cymon.io
CymonPhishing	www.cymon.io
CymonSpam	www.cymon.io
Cymondnsbl	www.cymon.io
EmergingThreatsBlockRules	rules.emergingthreats.net
EmergingThreatsCompromised	rules.emergingthreats.net
FeodoIpBlocklist	feodotracker.abuse.ch
MalcodeIP	www.malc0de.com
MalwareDomainIp	mirror1.malwaredomains.com
NoThinkDNS	www.nothink.org
NoThinkHTTP	www.nothink.org
NoThinkMalwareIRC	www.nothink.org
NoThinkSNMPWeek	www.nothink.org
NoThinkSSH	www.nothink.org
NoThinkSSHWeek	www.nothink.org
NoThinkTelnetWeek	www.nothink.org
OpenBLBase	www.openbl.org
PhishTankJSON	data.phishtank.com
SSLIPBlacklist	sslbl.abuse.ch
ZeusTracker	zeustracker.abuse.ch

2) Generating Dishonest Feeds

To capture the effect of dishonest feeds, we considered two strategies: listing random entries and imitating high-ranked feeds.

I. Adding random entries: Adding random entries is a straightforward approach for a dishonest feed to improve its rank because it makes the feed appear larger and more up-to-date. Particularly because random entries are unlikely to be contained in other feeds, thus the tampering feed is the first to report them.

Adding random entries can be risky for a CTIF provider as it can increase the false positive rate, especially if an entry maps to a popular non-malicious service. However, by choosing unused (or rarely used) IP addresses or domains, a dishonest CTIF provider can cheat with a low risk of being detected.

We call a synthetic feed that follows such a strategy "RandomFeed" and generate it by choosing 50k IP addresses uniformly at random at each time unit (i.e. 1 hour).

II. Copying entries from high-reputation feeds: For this case, we generate "CopyFeed" by assuming that it copies all entries from the two best-ranked feeds with a delay of one time unit (i.e. 1 hour). By doing so, CopyFeed becomes the most complete feed but it lacks speed as it is never the first to announce any entry.

3) Parameters

PageRank, which is part of the graph ranking, requires the specification of a damping factor and a convergence condition. As we explained in Section 3B, we compute the damping factor as d=1-1/l where l is the average path length. For the evaluated dataset, the average path length is 2.87, which leads to a damping factor of 0.65. For the convergence condition, we choose a maximum delta (i.e. the precision of the results) of 10^{-6}.

B. FeedRank Dataset Baseline

In this section, we illustrate our dataset and the input of FeedRank with Figure 3 and Figure 4 after listing basic properties of each analyzed CTIF in Table IV.

TABLE IV: SIZE OF THE EVALUATED CTIFS.

	Number of entries				Number of entries		
Nr.	average ▽	max	min	**Nr.**	average ▽	max	min
F1	49125	51349	46931	F15	771	772	765
F2	22997	24397	21523	F16	764	1988	37
F3	16092	16591	15717	F17	500	500	500
F4	16085	16948	15247	F18	464	522	252
F5	15587	15826	15305	F19	444	444	444
F6	12719	18467	2362	F20	184	283	85
F7	8260	8861	7640	F21	127	133	121
F8	7956	11618	7	F22	127	133	123
F9	2556	2807	1165	F23	115	115	115
F10	2134	3491	6	F24	43	43	43
F11	1756	1761	1750	F25	40	43	37
F12	1277	1330	1213	F26	28	33	20
F13	1034	2695	27	F27	21	25	1
F14	1029	1033	498				

For a first insight in correlations in our dataset, we use Figure 3 to visualize a clustering of the evaluated feeds according to the number of common entries. That is, we run the Stoer-Wagner HCS (highly connected subgraphs) clustering algorithm [8] on a graph where the nodes represent feeds and the edges connect feeds with common entries

and are assigned a weight that equals the number of common entries. In Figure 3, we observe four clusters:

- One big cluster containing 7 (out of 27) feeds of different providers.
- Two smaller clusters consisting of 2 and 3 feeds from the same provider.
- One small cluster of two feeds (F11 and F15) where the number of common elements corresponds to the size of the smaller feed. This depicts an example of a feed (F15) that most likely contains a subset of the entries from another feed (F11).

Even though this clustering is not directly contained in the FeedRank algorithm, it shows that there are indeed correlations between the analyzed feeds.

In Figure 4, we show the correlation graph for the evaluated feeds. As explained in Section 3. B, this graph consists of nodes representing the feeds and directed, weighted edges that describe the percentage of confirmed entries from another feed.

FIGURE 3: EVALUATED FEEDS CLUSTERED BY THE NUMBER OF COMMON ELEMENTS. ABOUT 25% OF ALL FEEDS ARE CONTAINED IN ONE CLUSTER. FEEDS FROM THE SAME PROVIDERS ARE CONTAINED IN SMALLER CLUSTERS AND F15 DOES NOT LIST ELEMENTS THAT ARE NOT IN F11.

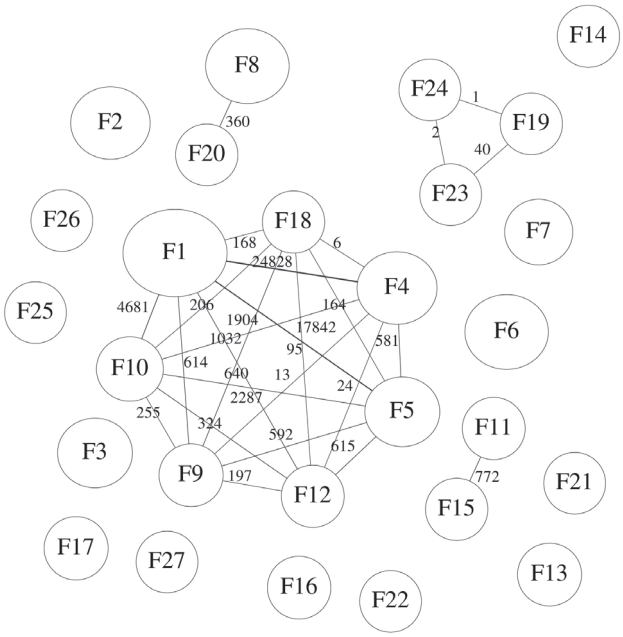


FIGURE 4: CORRELATION GRAPH FOR THE EVALUATED FEEDS. AS AN EXAMPLE, THE LARGEST FEED (F1) AND ALL ITS IN- AND OUTGOING EDGES ARE HIGHLIGHTED. THE EDGE LABEL DENOTES THE PERCENTAGE OF ENTRIES THAT A FEED CONFIRMS.

C. FeedRank vs. Individual Metrics

In this experiment, we compare the ranking obtained by individual metrics with the ranking according to FeedRank (see Table V). In Table VI, we show the ranking for all real feeds. The listed overall rank corresponds to the ranking according to the combination of all individual (or FeedRank) metrics. In this non-malicious case, we observe that the ranking according to the two metrics are strongly correlated (with a Spearman correlation coefficient of $\rho = 0.81$).

TABLE V: EVALUATED FEED METRICS.

Individual metrics	
Size	The average number of entries contained in the feed (more is better).
New	The average number of new entries per hour (more is better).
Removed	The average number of removed entries per hour (more is better).
FeedRank	
Contribution	A measure of how many additional entries a feed contributes.
Graph	The score obtained from the correlation graph.

TABLE VI: RANKING WITH INDIVIDUAL METRICS COMPARED WITH FEEDRANK. THE RANKINGS ARE STRONGLY CORRELATED ($\rho = 0.81$).

	Individual metrics				**FeedRank**		
Feed	Size	New	Removed	**Overall**	Contribution	Graph	**Overall** △
F1	1	5	7	3	1	10	1
F4	4	6	5	4	6	6	2
F13	13	4	4	6	9	5	3
F16	16	7	6	9	11	3	3
F8	8	2	2	2	2	14	5
F12	12	16	13	12	16	1	6
F9	9	15	21	17	10	8	7
F3	3	10	11	7	4	15	8
F10	10	3	3	5	13	7	9
F7	7	13	21	16	7	13	9
F2	2	9	21	11	3	18	11
F6	6	1	1	1	5	17	12
F5	5	11	9	8	14	9	13
F27	27	17	14	19	24	4	14
F21	21	20	16	18	18	11	15
F18	18	14	12	14	25	2	15
F17	17	8	8	10	8	21	15
F25	25	21	18	22	21	12	18
F11	11	18	15	14	12	22	19
F14	14	24	20	19	15	22	20
F20	20	12	10	13	19	20	21
F22	22	22	19	22	25	16	22
F19	19	25	21	25	17	24	23
F15	15	22	21	24	25	19	24
F23	23	25	21	26	20	24	25
F24	24	25	21	27	21	24	26
F26	26	19	17	21	23	24	27

D. Tamper-resistance

In this experiment, we evaluate the impact of RandomFeed and CopyFeed on the ranking. As the results in Table VII show, the dishonest feeds can obtain very good ranks (rank 1 for RandomFeed and rank 3 for CopyFeed) according to individual metrics, but not for FeedRank (rank 16 and 20).

TABLE VII: RANKING IN THE PRESENCE OF DISHONEST FEEDS. RANDOMFEED AND COPYFEED CAN TAMPER WITH WITH INDIVIDUAL METRICS, BUT NOT WITH FEEDRANK.

	Rank with individual metrics			**Rank with FeedRank**		
Feed	initial	+RandomFeed	+CopyFeed	initial △	+RandomFeed	+CopyFeed
F1	3	4	4	1	3	3
F4	4	5	5	2	2	2
F13	6	7	7	3	3	3
F16	9	10	10	3	3	5
F8	2	3	2	5	1	1
F12	12	13	13	6	7	7
F9	17	18	18	7	7	9
F3	7	8	8	8	11	11
F7	16	17	17	9	10	9
F10	5	6	5	9	6	5
F2	11	12	12	11	12	12
F6	1	2	1	12	7	8
F5	8	9	9	13	14	12
F27	19	20	20	14	18	17
F21	18	19	19	15	15	15
F17	10	11	11	15	13	14
F18	14	15	15	15	16	16
F25	22	23	23	18	19	19
F11	14	15	15	19	20	18
F14	19	20	20	20	21	20
F20	13	14	14	21	21	20
F22	22	23	23	22	23	24
F19	25	26	26	23	24	23
F15	24	25	25	24	25	27
F23	26	27	27	25	25	25
F24	27	28	28	26	27	27
F26	21	22	22	27	28	25
RandomFeed	n/a	1	n/a	n/a	16	n/a
CopyFeed	n/a	n/a	3	n/a	n/a	20

5. CASE-STUDY

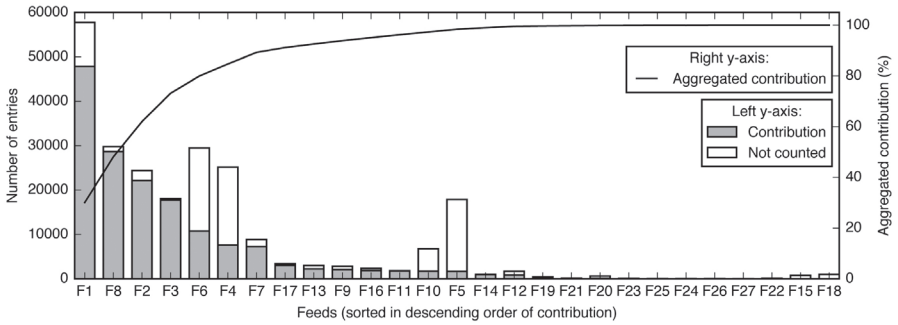
In this section, we come back to the two use-cases mentioned initially – network defenders that want to: *(i)* select the best feeds that together contain as many distinct entries as possible; and *(ii)* select the best feeds that list new entries before they appear on other feeds.

A. Prioritizing Completeness

To find a set of CTIFs that covers as many entries as possible (i.e. is as complete as possible) while not being susceptible to tampering attempts, FeedRank is used as follows. First, we compute the ranking solely according to the contribution. Since this ranking ignores the correlations, it is not tamper resistant and a CTIF that adds random entries can achieve a good rank. In a second step, we ensure tamper-resistance by excluding CTIFs whose graph score is below a user-defined percentile. Intuitively, the choice of this percentile reflects how many dishonest feeds the user expects. Here, we use the fifth percentile; that is, we assume that feeds whose graph metric is in the upper 95% are non-malicious.

In Figure 5, we plot the contribution of all collected CTIFs. As the figure shows, considering a small subset of all feeds is enough to cover a large percentage of all entries (e.g. the best 5 feeds together cover 80% of all entries). The figure also shows that by only looking at a feed's size it is not possible to derive the feed's contribution.

FIGURE 5: CONTRIBUTION OF ALL EVALUATED CTIFS. SELECTING 5 FEEDS IS ENOUGH TO COVER 80% OF ALL REPORTED IPS.



In Tables VIII and IX, we show the ranking in the presence of dishonest feeds.

RandomFeed has a high contribution score because its entries are most likely not listed on any other feed. However, because the vast majority of them are not confirmed by

any other feed, the graph score is very low. In particular, the graph score is below the 5th percentile, which is why the feed is not eligible to be used. CopyFeed has a poor contribution score because it is never the first feed to list any entry. However, because the copied entries originate from highly ranked feeds, CopyFeed deservedly achieves a good graph score.

TABLE VIII: RANKING ACCORDING TO THE CONTRIBUTION METRIC IN THE PRESENCE OF RANDOMFEED. THE GRAPH METRIC IS USED TO EXCLUDE POTENTIALLY DISHONEST FEEDS.

	FeedRank				
Feed	Contribution	Graph	Percentile	Rank △	Eligible
RandomFeed	28	1	0	1	No
F1	27	16	54	2	Yes
F8	26	21	71	3	Yes
F2	25	9	29	4	Yes
F3	24	11	36	5	Yes
F6	23	14	46	6	Yes
F4	22	22	75	7	Yes
F7	21	15	50	8	Yes
F17	20	13	43	9	Yes
F13	19	24	82	10	Yes
F9	18	19	64	11	Yes
F16	17	26	89	12	Yes
F11	16	7	21	13	Yes
F10	15	27	93	14	Yes
F5	14	17	57	15	Yes
F14	13	6	18	16	Yes
F12	12	25	86	17	Yes
F19	11	1	0	18	No
F21	10	20	68	19	Yes
F20	9	10	32	20	Yes
F23	8	1	0	21	No
F25	6	18	61	22	Yes
F24	6	1	0	22	No
F26	5	1	0	24	No
F27	4	23	79	25	Yes
F15	1	8	25	26	Yes
F18	1	28	96	26	Yes
F22	1	12	39	26	Yes

TABLE IX: RANKING ACCORDING TO THE CONTRIBUTION METRIC IN THE PRESENCE OF COPYFEED. THE GRAPH METRIC IS USED TO EXCLUDE POTENTIALLY DISHONEST FEEDS.

	FeedRank				
Feed	Contribution	Graph	Percentile	Rank △	Eligible
F1	28	16	54	1	Yes
F8	27	21	71	2	Yes
F2	26	8	25	3	Yes
F3	25	10	32	4	Yes
F6	24	13	43	5	Yes
F4	23	22	75	6	Yes
F7	22	14	46	7	Yes
F17	21	12	39	8	Yes
F13	20	24	82	9	Yes
F9	19	17	57	10	Yes
F16	18	25	86	11	Yes
F11	17	6	18	12	Yes
F10	16	27	93	13	Yes
F5	15	19	64	14	Yes
F14	14	5	14	15	Yes
F12	13	26	89	16	Yes
F19	12	1	0	17	No
F21	11	20	68	18	Yes
F20	10	9	29	19	Yes
F23	9	1	0	20	No
F25	7	15	50	21	Yes
F24	7	1	0	21	No
F26	6	4	11	23	Yes
F27	5	23	79	24	Yes
F15	1	7	21	25	Yes
F22	1	11	36	25	Yes
CopyFeed	1	18	61	25	Yes
F18	1	28	96	25	Yes

B. Prioritizing Speed

In this case study, a network defender wants to select CTIFs such that new entries are available as early as possible. For this, we rank the feeds according to the graph metric (that is, we ignore the contribution). FeedRank's correlation graph models the order in which entries appear in the feeds. Therefore, a feed that scores well in the

graph metric is one that is fast in including new entries. In contrast to computing the added entries for each feed individually, FeedRank ensures that it is impossible for a dishonest feed to tamper with the ranking.

In Table X, we show the resulting ranking with and without the dishonest feeds. RandomFeed appears at the very end of the ranking because its entries are not confirmed by other feeds. CopyFeed achieves a better rank because it copies the entries of highly ranked feeds with only a one-hour delay. By doing so, it is faster in listing these entries than other feeds that confirm the entries later.

TABLE X: RANKING ACCORDING TO THE GRAPH METRIC TO SELECT THE FASTEST FEEDS. THE DISHONEST FEEDS CANNOT ACHIEVE TOP RANKINGS.

	Rank with FeedRank		
Feed	initial △	+RandomFeed	+CopyFeed
F12	1	4	3
F18	2	1	1
F16	3	3	4
F27	4	6	6
F13	5	5	5
F4	6	7	7
F10	7	2	2
F9	8	10	12
F5	9	12	10
F1	10	13	13
F21	11	9	9
F25	12	11	14
F7	13	14	15
F8	14	8	8
F3	15	18	19
F22	16	17	18
F6	17	15	16
F2	18	20	21
F15	19	21	22
F20	20	19	20
F17	21	16	17
F14	22	23	24
F11	22	22	23
F19	24	24	26
F23	24	24	26
F24	24	24	26
F26	24	24	25
RandomFeed	n/a	24	n/a
CopyFeed	n/a	n/a	11

6. DISCUSSION

In this section, we first summarize the answers to the research questions, then discuss additional aspects of and choices that we made in the design of FeedRank.

A. Research Questions

Our research questions listed in Section 1 relate to the estimation of the quality of CTIFs, the CTIF ecosystem and the tamper-resistance of the evaluation metrics.

CTIF quality estimation. We use a graph-based correlation analysis together with a contribution analysis to measure correlations between CTIFs and the individual contribution of each CTIF. This allows us to estimate the relative quality of each CTIF with respect to all other analyzed CTIFs without requiring a ground truth.

CTIF ecosystem. Our correlation analysis allows finding clusters of highly correlated CTIFs and shows that most of the evaluated feeds are contained in the same cluster (i.e. most of the feeds overlap in terms of their entries but differ in terms of speed).

Tamper resistance. Our evaluation shows that correlation and contribution are tamper-resistant metrics for ranking CTIFs. While FeedRank produces a ranking that is strongly correlated with the ranking according to individual metrics in the absence of dishonest feeds, only FeedRank allows to identify dishonest feeds and to prevent them from achieving a good rank.

B. Speed of Dishonest Feeds vs. Execution Interval of FeedRank

For our evaluation, we use hourly snapshots and assume that the dishonest CopyFeed copies entries with a delay of one hour. If the delay were to be shorter than the snapshot interval, FeedRank could not determine whether CopyFeed and the two copied (legitimate) feeds listed the entries simultaneously or not. To prevent this inaccuracy, we envision the following mechanisms:

- The time between two snapshots can be decreased, which makes it more likely to be faster than dishonest feeds.
- CTIF providers can provide FeedRank with exclusive access to updates shortly before they are published.
- CTIF providers can add a few random (non-malicious and inactive) entries to their feeds to detect if another feed copies them (if these entries appear on another feed, it is highly likely that they were copied).

C. Evaluating CTIFs Instead of Evaluating Entries in CTIFs

With FeedRank, we assess the quality of CTIFs as a whole and not the quality of individual entries. With this, FeedRank provides the foundation to select CTIFs for deployment. The problem of evaluating the quality of particular entries is orthogonal to our work, but could be approached with a similar technique (e.g. by building a graph that models individual entries). From a network defender's point of view, the advantages of scoring threat intelligence at the level of CTIFs instead of individual entries are that: *(i)* reporting an IOC is delayed if an entry first needs to be verified by multiple feeds; and *(ii)* scoring CTIFs can be done once before deciding which feeds to use, which reduces operational and potential subscription costs.

D. Choice of the Evaluated Feeds

For our evaluation, we used a generic threat model and included a large set of freely available feeds covering different domains (e.g. generic, malware or phishing). Generally speaking, the set of considered CTIFs should contain all feeds that are suitable for the network defender's purpose. For example, a network defender that wants to select CTIFs for a spam filter should only consider feeds in this domain to get the most meaningful results.

Evaluating CTIFs for a more specific threat model or including commercial feeds is possible without modifying FeedRank but it is out of the scope of this paper.

7. RELATED WORK

To the best of our knowledge, we are the first to rank CTIFs based on their correlations and to consider potential manipulation strategies from CTIF providers. However, there has been previous work in the area of evaluating CTIFs and applying graph-based ranking algorithms in other domains.

A. Analysis and Evaluation of CTIFs

Sheng et al. study the effectiveness of phishing blacklists [9] and find that blacklists are ineffective when protecting users against phishing attacks because most phishing campaigns only last for a short time and blacklists are too slow in reacting.

Kührer and Holz describe a CTIF parser system [10] that records a large number of CTIFs and allows users to compute intersections between feeds and to query entries (e.g. domains). Entries that are contained in a large number of feeds are considered as being dishonest with high certainty. In later work [11], Kührer et al. propose a mechanism to identify parked domains and sinkholes (i.e. malicious domains that are identified and mitigated) in CTIFs.

Metcalf and Spring present an analysis of CTIFs over multiple years [12]. Similar to our approach, they analyze individual and combined features of CTIFs. However, they do not address the issue of dishonest CTIF providers that attempt to manipulate the rankings.

The Ponemon Institute found in a survey [3] that the application of threat intelligence is considered as very important for running secure systems but they did not investigate in the quality or the ecosystem of threat intelligence providers.

B. Graph-based Ranking

Page and Brin developed PageRank to rank websites [5]. They showed that the problem of ranking websites can be transferred to a graph problem and thus provided the foundation of transferring problems with several connected parties to graph problems.

Since then, concepts similar to PageRank have been applied to various problems, including to:

- Predict future relevance of scientific articles [13].
- Rank authors and publications [14].
- Rank correspondents according to their degree of expertise [15].
- Find influential users [16] and important content [17] in social networks.

8. CONCLUSION AND FUTURE WORK

The core concept of FeedRank is to model temporal correlations between feeds in a graph structure and to rank feeds based on this graph and the individual contribution of each feed. In contrast to traditional metrics that are applied to feeds individually, FeedRank is robust against tampering attempts by potentially dishonest feed providers.

In the evaluation and two case studies, we use data from 27 real feeds and show that FeedRank allows a reliable ranking even in the presence of dishonest feeds. For future work, we suggest using FeedRank to track the rankings of CTIFs over time. This will provide insights in the long-term behavior of CTIFs. Further, FeedRank could be extended by additional metrics and applied to related problems such as the evaluation of single entries in CTIFs.

REFERENCES

[1] D. Shackleford, “Who’s Using Cyberthreat Intelligence and How?,” SANS Survey, 2015.
[2] Symantec, “Internet Security Threat Report,” Bd. 22, 2017.
[3] “The Value of Threat Intelligence: The Second Annual Study of North American & United Kingdom Companies,” Ponemon Institute, 2017.
[4] H. Slatman, “awesome-threat-intelligence,” [Online]. Available: https://github.com/hslatman/awesome-threat-intelligence.
[5] L. Page, S. Brin, R. Motwani and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web.,” Stanford InfoLab, 1999.
[6] “python-stix,” [Online]. Available: https://github.com/STIXProject/python-stix.
[7] “libtaxii,” [Online]. Available: https://github.com/TAXIIProject/libtaxii.
[8] M. Stoer and F. Wagner, “A Simple Min Cut Algorithm,” *Journal of the ACM (JACM)*, Bd. 44, Nr. 4, pp. 585-591, 1997.
[9] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong and C. Zhang, “An Empirical Analysis of Phishing Blacklists,” in *Proceedings of Sixth Conference on Email and Anti-Spam (CEAS)*, 2015.
[10] M. Kührer and T. Holz, “An Empirical Analysis of Malware Blacklists,” *PIK-Praxis der Informationsverarbeitung und Kommunikation 35.1*, pp. 11-16, 2012.
[11] M. Kührer, C. Rossow and T. Holz, “Paint It Black - Evaluating the Effectiveness of Malware Blacklists,” in *International Workshop on Recent Advances in Intrusion Detection*, 2014.
[12] L. Metcalf and J. M. Spring, “Blacklist ecosystem analysis: Spanning Jan 2012 to Jun 2014,” in *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security*, 2015.
[13] H. Sayyadi and L. Getoor, “FutureRank: Ranking Scientific Articles by Predicting their Future PageRank,” in *Proceedings of the 2009 SIAM International Conference on Data Mining*, 2009.
[14] D. Zhou, S. A. Orshanskiy, H. Zha and C. L. Giles, “Co-ranking Authors and Documents in a Heterogeneous Network,” in *Seventh IEEE International Conference on Data Mining*, 2007.
[15] B. Dom, I. Eiron, A. Cozzi and Y. Zhang, “Graph-based ranking algorithms for e-mail expertise analysis,” in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003.
[16] Q. Wang, Y. Jin, S. Cheng and T. Yang, “ConformRank: A conformity-based rank for finding top-k influential users,” *Physica A: Statistical Mechanics and its Applications*, Bd. 474, pp. 39-48, 2017.
[17] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, “Finding high-quality content in social media,” *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

HTTP Security Headers Analysis of Top One Million Websites

Artūrs Lavrenovs
NATO CCD COE
Tallinn, Estonia
arturs.lavrenovs@ccdcoe.org

F. Jesús Rubio Melón
Spanish Joint Cyber Defence Command
Madrid, Spain
jrubio@isdefe.es

Abstract: We present research on the security of the most popular websites, ranked according to Alexa's top one million list, based on an HTTP response headers analysis.

For each of the domains included in the list, we made four different requests: an HTTP/1.1 request to the domain itself and to its "www" subdomain and two more equivalent HTTPS requests. Redirections were always followed. A detailed discussion of the request process and main outcomes is presented, including X.509 certificate issues and comparison of results with equivalent HTTP/2 requests.

The body of the responses was discarded, and the HTTP response header fields were stored in a database. We analysed the prevalence of the most important response headers related to web security aspects. In particular, we took into account Strict-Transport-Security, Content-Security-Policy, X-XSS-Protection, X-Frame-Options, Set-Cookie (for session cookies) and X-Content-Type. We also reviewed the contents of response HTTP headers that potentially could reveal unwanted information, like Server (and related headers), Date and Referrer-Policy.

This research offers an up-to-date survey of current prevalence of web security policies implemented through HTTP response headers and concludes that most popular sites tend to implement it noticeably more often than less popular ones. Equally, HTTPS sites seem to be far more eager to implement those policies than HTTP only websites. A comparison with previous works show that web security policies based on HTTP response headers are continuously growing, but still far from satisfactory widespread adoption.

Keywords: *web security, HTTP headers, top one million websites survey, X.509 certificate, HTTP/2, HTTPS, HTTP Strict Transport Security, Content Security Policy*

1. INTRODUCTION

The main goal of this research is to assess the current adoption rate of security policies based on HTTP response headers on the most popular Internet websites. Declarative web security through HTTP response headers constitute a powerful and easy way to enhance website security, while relatively little effort is required from website operators. It has been a recurrent research topic, aided by the fact that the nature of the World Wide Web makes data publicly accessible to any interested party and that the WWW itself is continuously growing and evolving.

Besides measuring security headers adoption in popular websites, we set out to understand it in a deeper way by trying to find correlations between adoption rates and variables like HTTPS usage and popularity rank position. We want to gain insight into why and how policies based on HTTP headers are adopted. As will be shown, the most popular a website is, the more likely it will apply security through HTTP headers. Those sites also tend to be more prone to favouring HTTPS protocol over HTTP.

Regarding the structure of this paper, in the first section we present a brief literature review concerning different past security analysis and current online efforts. Next, we proceed to describe in detail the data set that served as the basis for this research. We then show our results for all analysed HTTP response headers, and we conclude with a "conclusions" section where we summarize our findings and a last section on planned future work.

2. RELATED WORK

Extensive analysis of Content Security Policy (CSP) adoption among the top one million websites is provided by Ying et al. (2015). It was found that CSP is used in less than 0.2% of the sites, and oftentimes incorrectly. They also investigated other relevant security related headers. In particular, they found that *X-XSS-Protection, X-Frame-Options* and *Strict-Transport-Security* headers were implemented, respectively, in about 4.4%, 4.1% and 1% of the websites they analysed. Despite the low adoption rate of HTTP security related headers found by Ying et al., their results show a noticeable

increase in the adoption rates observed over research done previously by Weissbacher et al. (2014). In fact, they conducted the first CSP adoption study of the Top One Million websites in 2012-2014 and found that CSP was used in less than 0.1% of sites. Other security-related HTTP headers, like *X-XSS-Protection*, *X-Frame-Options* and *Strict-Transport-Security* were seen on 4.6%, 4.1% and 0.3% of the websites, respectively. Although both of these papers primarily concentrate on CSP adoption rate and related implementation issues, they analysed other security headers as a by-product.

Chang et al. (2017) investigated the "redirection trail", which basically consists of a set of pairs, each one formed by the Location header combined with redirection HTTP status codes. Combining this redirection trail with other data readily available, like protocol and host, allowed the researchers to evaluate the security of the Top One Million websites. They found that 20.5% of them contained some configuration inconsistency related to redirection requests that could be exploited by the adversary. Sood et al.(2011) conducted a research in 2011 among the world's top 43 banks. They found that none of them implemented the HTTP security related headers available at that time.

Response HTTP header analysis from a security standpoint is also present outside academic literature. Scott Helme's (2017) website has published multiple times research on security headers prevalence and HTTPS adoption in the Alexa Top One Million websites. The latest one we know of, at the time of this writing (October 2017), is from August 2017. He has been reporting positive trends of adoption rates of most common HTTP headers. Additionally, his website (IO) provides a public tool that enables checking of security headers for any website. Based on these results, the tool assigns a given grade, from A to F, for the provided website. A similar tool is provided by Mozilla Observatory that also gathers statistics from executed checks and estimates that about 10% of the checked websites follow good practices regarding security header configuration (Mozobs). April King (2017) has conducted similar research on Alexa Top One Million websites and found similar results about positive trends.

3. THE DATA SET

A. URL Set

This research makes use of Alexa "top one million" website list (1M) as the source for domains to be analysed. For each domain contained in the list, we made four HTTP/1.1 requests: to http://domain, https://domain, http://www.domain and https://

www.domain. Timeout for connection establishment was set to 60 seconds, and response timeout was also set to 60 seconds.

B. Data collection approach

We developed a custom Python tool based on Python requests library (Pyreq). After an HTTP/1.1 response arrived, only HTTP response headers and status code were saved to a relational database. The response body was disregarded. In order to mimic real users, we set *User-Agent* and other request headers to match exactly those of the Mozilla Firefox browser (version 50.0 on Ubuntu 17.04). For all the requests we followed redirections, saved them all, and created convenient relationships between them. Finally, duplicate URL requests that arose from redirections were removed from the dataset.

Preliminary testing revealed that using Certification Authorities (CA) and Intermediaries lists bundled inside Ubuntu were not sufficient for HTTPS requests. Therefore, we made use of public CA lists Mozilla CA (Mozca) and Mozilla Intermediaries (Mozinter) (they are both internally used by the Firefox browser). Several full scans were performed during August and September 2017, and we always updated website and CA lists right before the scanning process. In this paper we will exclusively refer and analyse data gathered between September 1st and 4th, 2017. After the scan was completed, we retried those websites that had failed all of our four requests, since it just might indicate temporary network issues.

C. Data Overview

Our final dataset contained 3.135.962 recorded responses with unique URLs (either the protocol, the domain or the subdomain was different). At least one response was received from 975.729 websites (97.5% of all one million domains). Only 2.558 websites were successfully processed during the retrying process (to allow for network issues). We observed large amounts, up to 1.4 million, of duplicate URL records caused by redirection to an already visited URL. They were all removed from the database. We obtained about 27% more responses from www-subdomains than from direct domain requests.

For our current analysis we have considered only unique URL responses with HTTP response status code 200, which amounts to 1.478.750 records.

D. Data quality

The Alexa top one million list was chosen because it is a large list, but not “too large”, thus data collection can still be achieved in a few days or less. Moreover, the list contains the most popular websites, an attractive target for attackers and security researchers alike. As pointed out previously, it has been repeatedly used in various

web security surveys. Alternative lists, like the ones by Majestic (Majestic) or Cisco (Umbrella) also provide one million most popular sites, although to compare our work with previous results we have adhered to Alexa's list.

However, Alexa's list, despite its usefulness, has some caveats. It makes use of proprietary ranking and domain processing algorithms not fully disclosed and we have observed inconsistencies within the list: it contains many domains that cannot be accessed directly (typically because there is no DNS entry for them, like in *cloudfront.net*), but can be through the www-subdomain. Additionally, a significant set of entries in the list are actually subdomains (most common websites are *tumblr.com*, *blogspot.com* and *wordpress.com* with 5.698, 2.904 and 2.696 respective subdomains). Although content is different on these subdomains, these platforms usually provide little to no control for header configuration to final website authors. In fact, all of their subdomains will share the same security headers. Even some apparently unrelated domains will share the same headers configuration because they are hosted by these providers, although their domain name is totally unrelated to the hosting server.

E. Response Codes

Status code distribution observed in the responses for both HTTP and HTTPS requests are presented in Tables 1 and 2. As clearly seen from these data, most websites seem to prefer the www subdomain to the plain domain name, regardless of the protocol (47% of HTTP sites and up to 63% for HTTPS ones). As for redirections, we have observed that 45.7% of HTTP domain requests redirect to the corresponding www subdomain, and 15.5% of HTTPS domain requests point to the www subdomain. Most of the remaining requests are server-side errors either intermittent or permanent (e.g., web servers which are not properly configured to handle the domain or subdomain requests).

TABLE 1. STATUS CODES FOR HTTP REQUESTS

Domain responses		WWW subdomain responses	
status	**count**	**status**	**count**
301	46,8%	200	47.3%
200	38.9%	301	39.5%
302	12.0%	302	11.0%
403	0.7%	404	0.6%
404	0.7%	403	0.6%

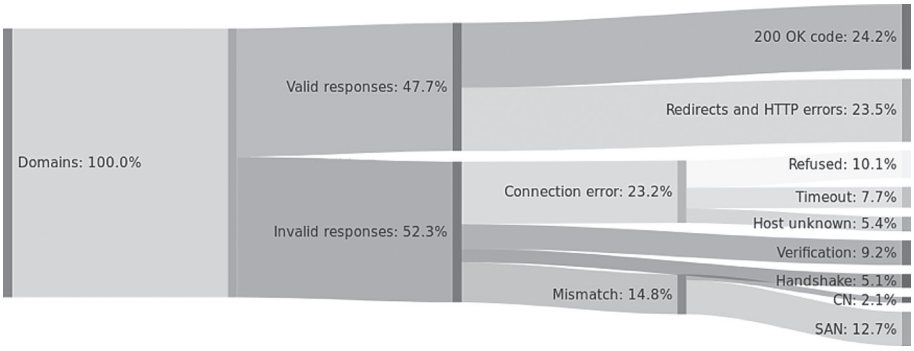
TABLE 2. STATUS CODES FOR HTTPS REQUESTS

Domain responses		WWW subdomain responses	
status	**count**	**status**	**count**
200	47.3%	200	62.7%
301	39.5%	301	26.0%
302	11.0%	302	8.4%
404	0.6%	403	0.7%
403	0.6%	404	0.6%

F. HTTPS Subset

Our data set allowed for a detailed analysis of HTTPS deployment since we stored all failed requests and their associated error messages. The result for over two million HTTPS requests are presented in the Figure 1.

FIGURE 1. HTTPS RESPONSES



Only about half of the scanned domains and www subdomains (47.7%) are properly configured for HTTPS. That number includes 24.2% of sites that are actually responding with 200 OK status code and a substantial number of redirects and HTTP errors (23.5%).

Many sites, 23.2% of all HTTPS requests, do not respond to HTTPS at all, either because of the TCP connection being refused, timeout or missing DNS records.

We found a sizeable number of cases, 29.1% of all HTTPS requests, where it was possible to establish a TCP connection on port 443, but HTTPS ultimately failed. The reason for failure is related to verification errors (mostly expired certificates, self-signed, signed by untrusted CA's or malformed), handshake errors (usually outdated

protocols) and hostname mismatching. A 29.1% rate is remarkable: it implies that a large number of web servers are already somehow configured to handle HTTPS requests, but have mostly missed the step of acquiring and installing the correct X.509 certificate (even though nowadays it is possible to quickly obtain a certificate for free, for example from the highly popular "*Let's Encrypt*" (Lets) online certification authority).

Regarding handshake errors, 5.1% of all HTTPS requests, we found that there is a small number of websites that work properly when requested by the real Mozilla Firefox browser but not by our software. We traced back that behaviour to outdated and misconfigured server sites that are still supported by the browser for backwards compatibility, but not by the OpenSSL 1.0.2g library we used in our scanning software.

Host name mismatching happens in about 14.8% of all HTTPS requests we made. That can happen because there is no Subject Alternative Name (SAN) and the requested host does not match certificate's Common Name (CN) or because, even though CN and SAN are both present, the hostname does not match either of them. This latter cause is more common (12.7%) than the former (2.1%). Name mismatching is typically found in shared environments where several websites run on a single server that oftentimes issues a "default" SSL certificate (as with the well-known shared hosting provider *Hostgator*). The most common reason for name mismatching is that CN is set to either *www.domain* or *.*domain*, and therefore certificate validation fails for the https://domain request (like the high ranked website *ups.com*).

G. HTTP/2 Analysis

Our data gathering procedure, and the subsequent response headers analysis, is entirely based on HTTP/1.1 requests. However, HTTP/2 is quickly growing in popularity and it may replace HTTP/1.1 as the main web protocol in the near future. Different protocol versions might be somehow correlated with different security settings (due, for example, to different security awareness). That raises the question whether there are different response headers, or different headers values, in HTTP/1.1 and HTTP/2 data subsets. To answer this question, we used the same Alexa top one million websites list (1M) and followed the same data gathering approach, but this time making an additional HTTP/2 request to each website's domain and www-subdomain. We made use of Python hyper library (Hyper). If an HTTP/2 request was successful, we made the equivalent HTTP/1.1 request to the same URL and compared the HTTP headers and their values for both responses. To simplify HTTP header comparison, we did not follow redirects and did not analyse response status codes. Furthermore, we did not take into account the fact that multiple backends can serve a single domain or www subdomain (in principle, those servers could have different configurations and that might produce different HTTP headers).[1]

1 Those backends could be either serving requests using single IP address or multiple IP addresses, but we chose not to manipulate to which IP addresses HTTP requests are being sent.

The resulting dataset consists of 746.758 records, totalling 211.638 unique domains that support HTTP/2 protocol (21% of all Alexa top one million websites). This percentage is a bit higher than the figure reported by the *w3techs* portal, 17%, as the HTTP/2 support rate across all the world wide web (W3tech). However, it is coherent with the fact that we are analysing most popular websites, not all existing ones. The failure rate of HTTP/1.1 requests to same domain following successful HTTP/2 requests is insignificant (0.26%).

1) Missing headers

We analysed the top 10 HTTP headers missing from HTTP/2 responses but present in HTTP/1.1 responses, and vice versa. The results, header names and their missing count in their counterpart protocol requests, are presented in Table 3. As might be expected, most significant differences are related to headers used in establishing and maintaining the HTTP/1.1 connection (those headers are unneeded in HTTP/2). Fortunately, none of these missing headers are related to any security issue.

Regarding security related headers, some may be missing in one version of the protocol, but present in the other, although the numbers are insignificant in all cases. For example, *X-XSS-Protection* response header is missing in 28 HTTP/2 requests that issue it in the equivalent HTTP/1.1 requests. Similarly, 45 HTTP/1.1 requests did not contain that header, although it was present in the equivalent HTTP/2 ones. We found that no common misconfiguration pattern is distinguishable, and most common cause could be attributed to responses coming from different backends serving the same domain name, but different protocol.

TABLE 3. RESPONSE HEADERS NAMES COMPARISON

Missing in HTTP/2	count	Missing in HTTP/1	count
connection	350152	content-length	11082
transfer-encoding	262147	link	3076
keep-alive	28559	pragma	1080
upgrade	5816	set-cookie	1058
cache-control	3014	vary	672
content-length	2983	cache-control	640
last-modified	1987	expires	614
x-nananana	1137	x-pingback	415
content-encoding	1107	accept-ranges	405
vary	1084	x-litespeed-cache-control	297

2) Different values in HTTP headers

The top 20 response headers that carry different values in equivalent HTTP/1.1 and HTTP/2 requests are presented in Table 4.

TABLE 4. RESPONSE HEADERS VALUES COMPARISON

Missing in HTTP/2		Missing in HTTP/1.1	
Header	**count**	**Header**	**count**
set-cookie	215265	last-modified	5229
cf-ray	183755	x-served-by	4921
date	181046	x-timer	4838
expires	23789	vary	4756
x-cache	14628	content-encoding	4706
server	10732	x-contextid	4434
content-length	9244	x-servedby	4384
x-varnish	6902	x-request-id	4295
via	6865	x-via	4283
x-amz-cf-id	6308	x-cache-hits	3664

Set-Cookie differences are due to different session identifiers. The differences in the response headers *Cf-Ray, X-Cache, X-Varnish, Via, X-amz-cf-id, X-Served-By, X-Timer, X-Contextid, X-Servedby, X-Request-Id, X-Via* and *X-Cache-Hits* are due to debug information, usually set by cloud providers and caching frontend servers. *Date, Expires* and *Last-Modified* response headers contain timestamps that are usually one second apart, in agreement with the fact that the requests are made consecutively. *Content-Encoding* differences are due to *Brotli*, the compression algorithm used for HTTP/2. Differences in the *Vary* header value are related to different compression algorithms. *Content-Length* variations are caused by the dynamic nature of the generated content. Nevertheless, none of these headers can be related to any security risk, and the variations are meaningless from the point of view of our security analysis.

Regarding security related headers, only *Content-Security-Policy* and *X-XSS-Protection* showed any significant count difference in 401 and 358 of the requests, respectively. Almost all of the CSP differences lie either in nonce tokens or report URI identifiers. *X-XSS-Protection* differences can be always traced back to different report URL's found in the value of the header.

Server header values show some differences between the protocol versions and in most cases it is irrelevant (variations of nginx server identified by names like

openresty, Tengine, kinsta-nginx). For example, 68.6% of HTTP/2 requests carried nginx as the Server value, but openresty for the equivalent HTTP/1.1 requests (77.2% of these cases correspond to tumblr.com subdomains). However, in about 400 cases we identified obvious attempts to try to conceal the server name by removing the header or changing it to an undescriptive one in one of the protocol version, but not in the alternative one.

In summary, regarding versions 1.1 and 2 of the HTTP protocol, no significant HTTP response headers variations were found from the security perspective. The only noticeable risk that shows some correlation with protocol version is information leakage via *Server* response header. Additionally, a potential security risk could arise due to inconsistent configuration management across sets of backend servers (which still could be useful to an attacker). However, this issue requires further investigation and lies outside the current research.

4. HTTP RESPONSE HEADERS ANALYSIS

As stated previously, the evaluation of the security of the websites is done through an analysis of the HTTP headers sent from the web server. Some HTTP headers, among all possible server-side headers, were devised to instruct the web browser to protect the web application against certain security threads. Accordingly, their analysis constitute the basis of our current research. Additionally, a few HTTP server-side headers may carry information about the web application that potentially can help an attacker to perform malicious actions. They will also be analysed as part of our research.

The headers involved in each group are the following:

Security headers:

- *Strict-Transport-Security*
- *Content-Security-Policy* (and related *Content-Security-Policy-Report-Only*)
- *X-XSS-Protection*
- *X-Frame-Options*
- *Set-Cookie*
- *X-Content-Type*

Information revealing headers:

- *Server* (and related headers)
- *Date*
- *Referrer-Policy*

We have deliberately excluded HTTP Public Key Pinning (HPKP) from our research. Standardized in IETF 7469, HPKP provides a mechanism by which the TLS protocol is protected against Certification Authority (CA) attacks and spoofed certificates. However, it is well known that its implementation poses considerable risks for website operators. It is currently supported by Chrome, Firefox and Opera. Nevertheless, Google has recently announced that it will deprecate it in Chrome in May 2018, and soon thereafter it will be completely removed Palmer 2017. Research by Scott Helme (2017) regarding his own analysis of Alexa Top One Million sites indicates that it is usually implemented wrongly by website operators. Security expert Ivan Ristic (2016) has also pointed out similar concerns about HPKP.

Subresource Integrity (SRI) has been sometimes taken into account in the context of top one million analysis (see Mozilla Observatory (Mozobs) or recent April King results (l2017). By specifying a hash token together with the URL of any given resource on a web page, a browser can check that the resulting content obtained from actually downloading the resource has not been unexpectedly altered. This technique is effective, for example, against attackers manipulating JavaScript libraries located in Content Delivery Networks. Chrome, Firefox and Opera already implement this feature. SRI is a relatively new protective mechanism, and current recommendation is from 2016 (SRI). Despite the undeniable interest in measuring its current adoption rate among the top one million websites, it entails parsing the HTML content found in the body of the HTTP responses, and it lies outside the scope of the current research, centred around HTTP response headers analysis.

A. Security Headers

1) Strict-Transport-Security Header

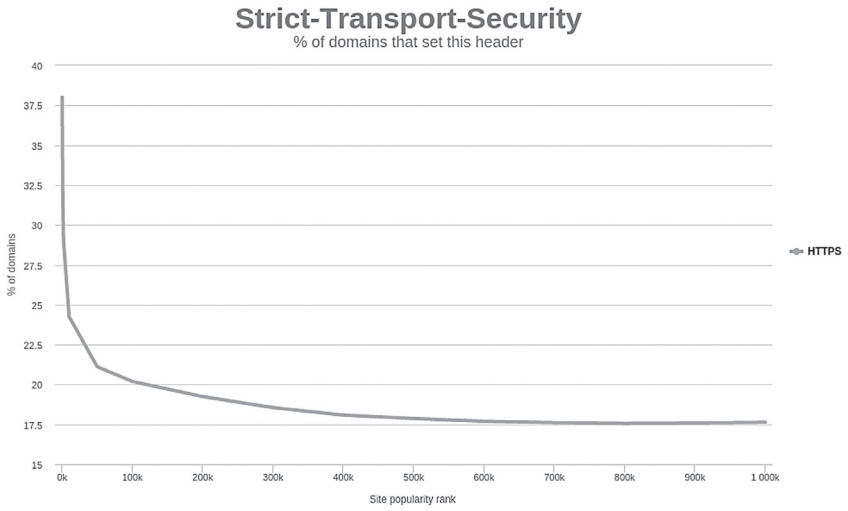
HTTP *Strict-Transport-Security* header (or HSTS, for short) allows a web server to inform the browser that all subsequent connections for all requests should be established exclusively through HTTPS, never through HTTP, using a valid certificate. It helps prevent several man-in-the-middle (MITM) attacks that may arise in different situations. Some common vulnerable situations are the following:

- A user types in a URL in the browser address bar. By default, this URL will be requested by the browser through an HTTP connection, not an HTTPS one.
- By means of social engineering techniques, a user is tricked into clicking on an HTTP link, instead of an HTTPS one, therefore initiating the HTTP request that can be captured by the MITM.
- An attacker sends a fake certificate, hoping that the user will accept it by clicking through the warnings displayed by the browser.
- Forgotten HTTP links scattered throughout the web pages.

All of these vulnerable situations can be avoided by the web application just by issuing this header. In fact, no other server header or web application configuration exists that can prevent these kind of MITM attacks (at least, regarding the first two cases). That makes HSTS a key protective server-side header. The header specification was published in 2012, (RFC 6797).

By parsing the scanning data obtained from the top one million web sites we have found that most websites do not issue any HSTS header. The aggregated results can be seen in Fig 2.

FIGURE 2. HSTS IMPLEMENTATION RATE AS A FUNCTION OF WEBSITE POPULARITY



It is readily appreciated that highly popular sites tend to implement HSTS more often than those sites that are less popular. Nearly 38% of top one thousand sites implement HSTS, while only 17.5% of top one million HTTPS websites implement it. This trend will be recurrent for all headers analysed in this research: the most popular a site is, the more security headers it will tend to implement.

Our numbers are comparable to the ones published by Helme (2017), who reports a 7.3% penetration rate. The difference is due to the fact that Helme's results are referred to the whole dataset, just not to the HTTPS sites (about 40% of all websites). Our 17.5% HSTS implementation rate becomes 7.0% when referred to the whole dataset. April (2017) reports a 4.4% adoption rate in June 2017 (also referred to the whole dataset). We believe, however, that HSTS rates should be referred to the HTTPS subset, since HSTS does make sense in HTTP only sites.

Only about 2% of HTTP websites redirect to an HTTPS site while simultaneously enforcing HSTS policy. Finally, a small number of sites (0.7%) make use of HTTP protocol and respond with a status code of 200, instead of responding with a redirection 300 code.

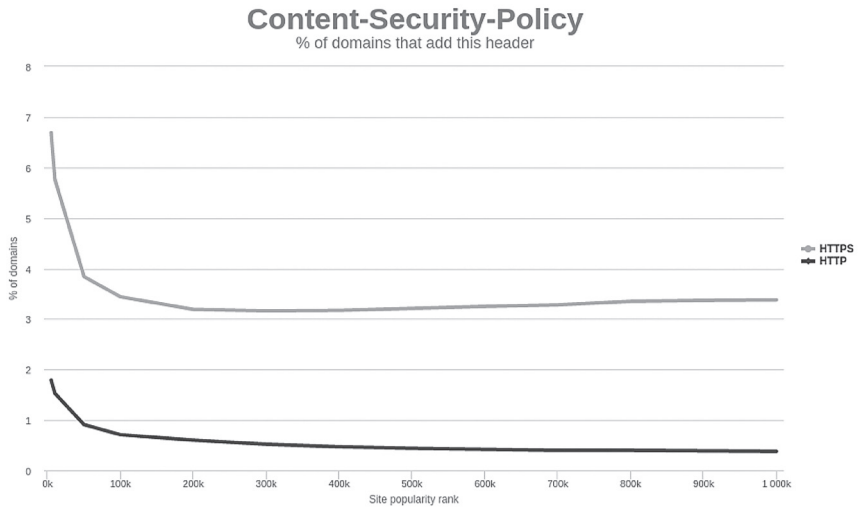
3) Content-Security-Policy Header

Content-Security-Policy (CSP) is a key response header that provides strong defence mechanisms against Cross Site Scripting (XSS) and other client-side injection attacks by whitelisting allowed sources and disabling certain insecure JavaScript features. It can also be used to prevent attacks against HTTPS, mostly those related to inadvertent HTTP links within HTTPS web pages. It has been standardized by W3C, originally in 2012 (CSP Level 1), then revised and augmented in 2015 (CSP Level 2) and currently undergoing a third revision (CSP Level 3). CSP is currently supported by all major browsers, with the exception of Microsoft Internet Explorer which uses the alternative *X-Content-Security-Policy* header.

The header directives, up to 16 in CSP2, offer the possibility of a fine-grained configuration, although at the cost of having to deal with non-trivial setup choices. In fact, due to the growing complexity of client-side scripting code and the large number of different assets handled by web applications (up to hundreds or even thousands of different resources requested from within a given page), the adoption of a CSP policy may result in unexpected glitches. Therefore, most implementation guidelines recommend starting to implement CSP by making use of the related *Content-Security-Policy-Report-Only* response header that allows web administrators to test their CSP policies before they are fully enforced without risking unwanted web application behaviour.

Our results show that CSP is scarcely implemented in HTTPS sites (3.4%) and hardly in HTTP sites (0.4%). The figures for Content-Security-Policy-Report-Only usage are even smaller (0.3% and 0.1% respectively). Globally, including both HTTP and HTTPS sites, CSP is implemented in 1.6% and CSP report only version in just 0.2% of sites. On the other hand, the implementation rate of CSP with respect to the popularity rank follows the same pattern as with other headers: more popular sites choose to issue the CSP header more often than less popular ones. These findings can be easily appreciated in Figure 3:

FIGURE 3. CSP IMPLEMENTATION RATE AS A FUNCTION OF WEBSITE POPULARITY



Our results are similar to the ones by Helme (2017), from August 2017, about 2.0% globally, while significantly higher than April's (l2017), 0.04% in June 2017.

We have also observed that there are significant differences between the directives used in HTTP and HTTPS sites. In fact, for HTTP sites, *frameAncestors* (48.27%), *scriptSrc* (35.96%) and *defaultSrc* (35.72%) are the most common directives. However, HTTPS sites typically issue different directives. The most common ones being: *upgradeInsecureRequests* (61.79%), *reportUri* (53.34%) and *defaultSrc* (20.37%).

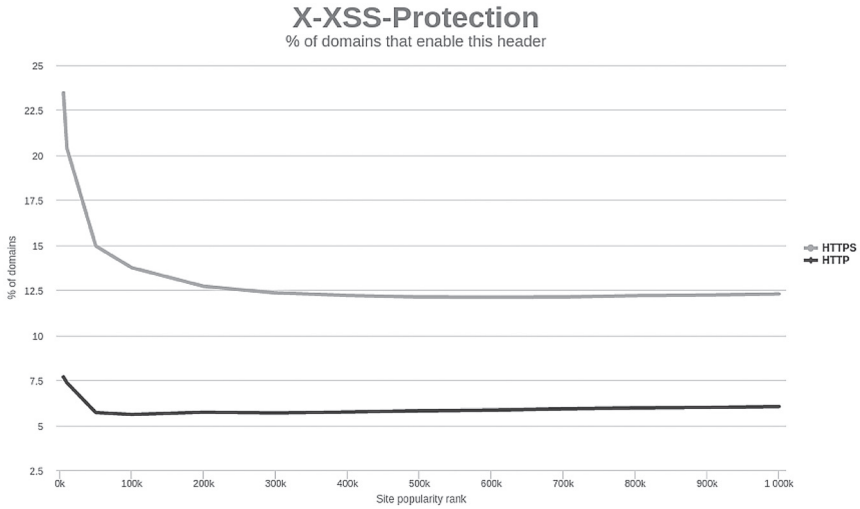
On the other hand, it is interesting to note that the CSP report only version seems to differ from fully enforcing CSP. In fact, the most common directives in HTTP sites are *reportUri* (94.68%), *blockAllMixedContent* (81.73%), *defaultSrc* (13.53%). And for HTTPS sites, most common CSP report only directives are *reportUri* (94.68%), *blockAllMixedContent* (81.73%), *defaultSrc* (13.53%).

4) X-XSS-Protection Header

This header is responsible for toggling off the XSS filter implemented by most current browsers (except, notably, Firefox). By default, the XSS filter is enabled, but website administrators can disable it by setting its value to zero (*X-XSS-Protection: 0*), possibly to prevent the browser from interfering with the desired behaviour of the web application. Web sites that issue that header, and set its value to zero, risk being vulnerable to reflected XSS attacks. Content Security Policy, and in particular CSP level 2 contains a directive, "*reflected-xss*", that completely replaces this header.

Our scanning results for the full one million set show that about 12% of HTTPS sites and 6% of HTTP sites set this header. Most of the times the header is issued so that the browser is granted the right to apply its XSS filter, but in 3% of HTTP sites, and nearly 2% of HTTPS sites, the configuration is such that the sites deny permission to apply the filter. Therefore, as expected, HTTPS sites tend to be more concerned with security. In a similar way, the more popular a site is, the more it will tend to set the header, and will mostly do it so that the browser is granted the right to enable its filter. Both trends can be appreciated in Figure 4.

FIGURE 4. X-XSS-PROTECTION IMPLEMENTATION RATE AS A FUNCTION OF WEBSITE POPULARITY



Our findings are in agreement with a global implementation rate of [Helme2017], 9.3%, and [April2017], 8.1% (none of them break up implementation rates by protocol or popularity of website).

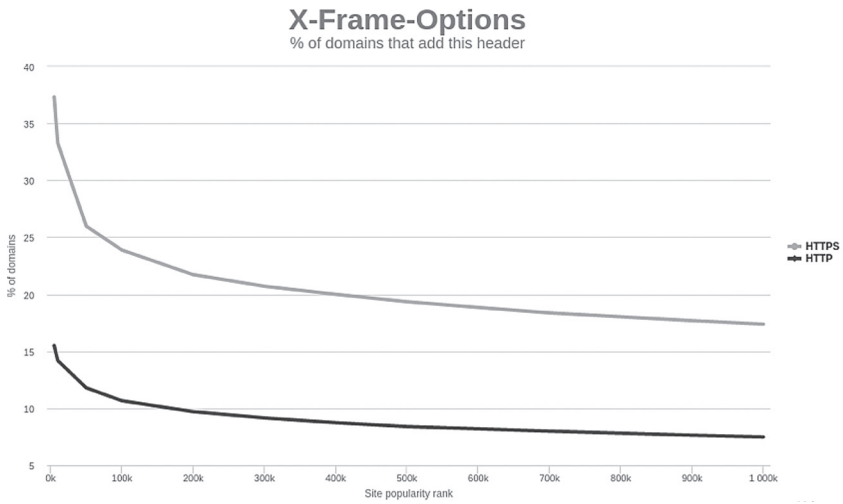
5) X-Frame-Options Header

This header, standardized in RFC 7034, is used to instruct a browser whether a given web page or resource is allowed to appear within a Frame, iFrame or Object, thereby avoiding frame based attacks, like "clickjacking" (for example, rydstedt or OWASPxfo). As with *X-XSS-Protection*, this header is superseded by CSP, which contains a directive, "*frame-ancestors*", that completely replaces *X-XSS-Protection* header. There are three "options", or directives, defined for this header: *deny, same-origin* and *allow-from*.

The results from our scanning survey show once more that HTTPS sites are prone to add this header more often than HTTP sites, 17.38% and 7.48%, respectively. For the

sites that choose to issue this header, the most common directives found are "same-origin" (86% in HTTP and 91% in HTTPS) and "deny" (12% in HTTPS and 7% in HTTP). Again, highly popular sites make use of this protection more often than less popular sites as can be readily appreciated in Figure 5:

FIGURE 5. X-FRAME-OPTIONS IMPLEMENTATION RATE AS A FUNCTION OF WEBSITE POPULARITY



These results do not essentially deviate from those of Helme (2017), 12.4% or April (2017), 11%.

6) Set-Cookie Header

This header is used by web sites to send cookies to the client side as part of the response message. Supported by all browsers, its current syntax was standardized by IETF RFC 6265. From the security perspective, the interest on this header lies on the "session cookies", i.e., those cookies that are set from the server side with the purpose of establishing a "session" between client and server (the stateless HTTP protocol was devised without any built-in session mechanism). In principle, cookies can be sent from the server to the browser without any particular security risk, unless they are session cookies. These cookies constitute a major target of many web application attacks, and therefore, we have tackled their study as part of the current research.

In order to prevent session hijacking and other web attacks that usually proceed through Cross Site Scripting or MITM attacks, it is generally agreed that session cookies should, at least, carry the directives "*HttpOnly*", for both HTTP and HTTPS sites, and "*Secure*", for HTTPS sites. *HttpOnly* offers protection against cookies being

accessed from client-side scripts (and therefore stolen under XSS attacks) and the *Secure* flag prevents the cookies from being captured through an unintended HTTP connection.

However, not all cookies need to be protected by *HttpOnly* or Secure flags, only session ones. Given the fact that session cookies need not carry any flag or follow any rules that distinguish them from non-session cookies, we have tried to tell them apart, and therefore assess the presence of the mentioned flags, by parsing the Set-Cookie value and search there for the token "sess" (case insensitively). While this is far from being a satisfactory criterion, our research shows that most of "highly probable" session cookies can be identified this way. Table 5 shows most frequent cookie names observed in the responses obtained from our data set:

TABLE 5. MOST COMMON COOKIE NAMES

Cookie name	Frequency
__cfuid	24.3%
PHPSESSID	20.3%
ASP.NET_SessionId	4.5%
JSESSIONID	2.5%

Cloud Flare ID cookie, __cfuid, cannot be considered properly as a web site session cookie (and indeed does not meet our criteria), while PHP sessions, ASP.NET sessions and Java based sessions are identified using this "sess" token technique. Taking all together, we can assume that 53.6% of all cookies received from server side are properly identified as being, or not, a session cookie. A further inspection analysing the 250 most popular cookie names proved that the "sess" token technique was enough to tell apart session cookies from ordinary cookies, up to that level of "cookie name popularity".

Our results show that, regarding HTTP sites, about 49.4% of them set a session cookie within the response to our first request, but 55.4% of those cookies do not set the *HttpOnly* flag. Regarding HTTPS sites, 42.7% do not set *HttpOnly* flag and up to 80.7% of them do not make use of the *Secure* flag. The following graphs exhibit the same pattern found in other security headers: HTTPS sites and popular sites seem to be more security concerned than HTTP or less popular sites.

FIGURE 6. SESSION COOKIES. *HTTPONLY* DIRECTIVE AS A FUNCTION OF WEBSITE POPULARITY

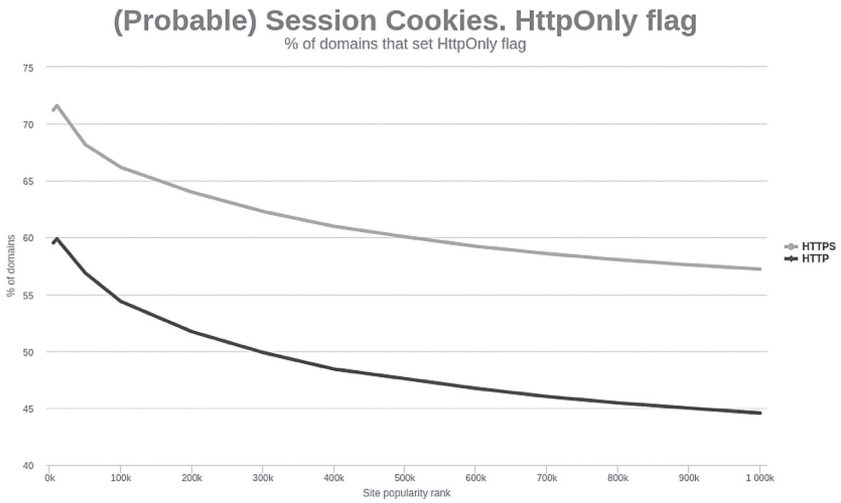
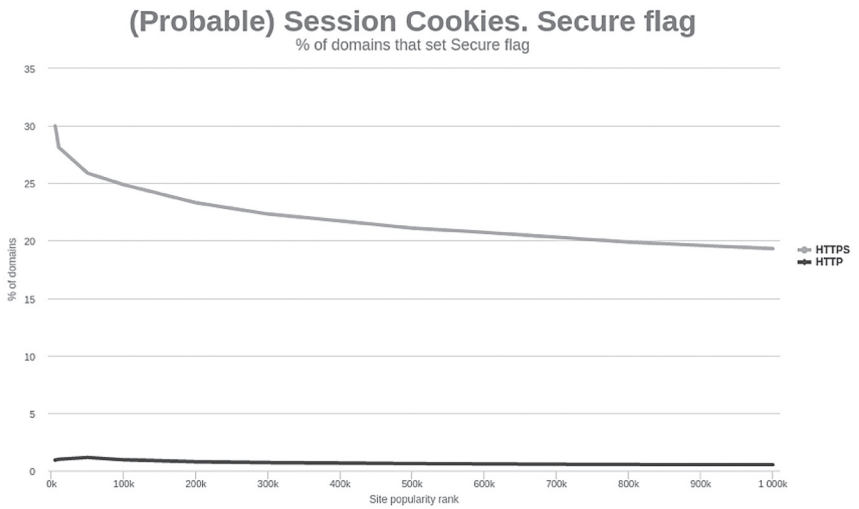


FIGURE 7. SESSION COOKIES. *SECURE* DIRECTIVE AS A FUNCTION OF WEBSITE POPULARITY



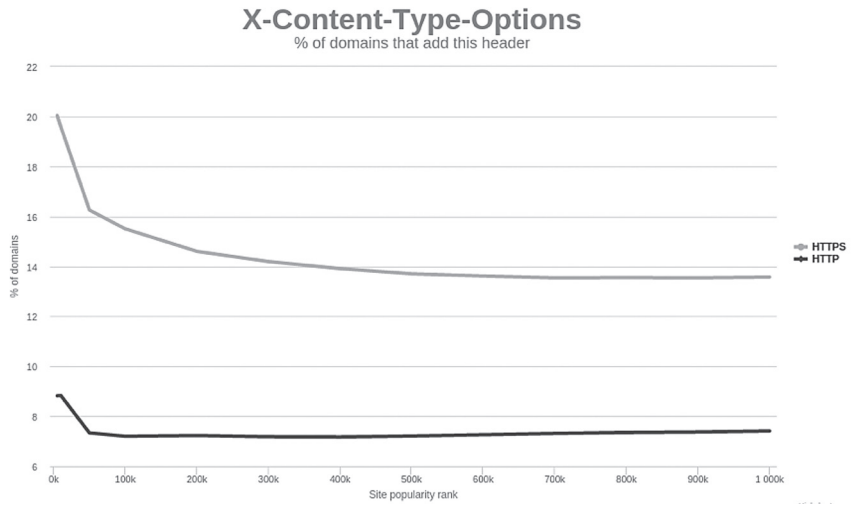
Finally, the "*SameSite*" cookie attribute recently implemented by Chrome and Opera, but still lacking in all other browsers, is an interesting flag currently defined under IETF draft (2016). It helps prevent Cross Site Request Forgery and cookie hijacking by instructing the browser not to send a cookie with that attribute to any request other than same-site requests. Although a very promising attribute, given its novelty and lack of widespread implementation, it is understandable that only 0.05% of HTTPS sites and 0.01% of HTTP sites make use of the flag.

7) X-Content-Type-Options Header

This header was defined to protect browsers from MIME sniffing vulnerabilities, by which an attacker may trick the browser into executing content that was not meant to be executed by the web application. These kinds of attacks make use of the fact that, under some circumstances, browsers do not follow the MIME type indicated in the *Content-Type* header. It is implemented by all major browsers, after Microsoft introduced it in IE8. The only allowed directive for this header is "*nosniff*".

The results follow the same pattern observed in other security headers: HTTPS sites set *X-Content-Type-Options* more often that HTTP ones (roughly, 16% vs 8%) and popular web sites do it also more often than less popular ones, as shown in the next figure:

FIGURE 8. X-CONTENT-TYPE-OPTIONS ADOPTION RATE AS A FUNCTION OF WEBSITE POPULARITY



Helme (2017) reports an adoption rate of 11.6%, whereas April (2017) finds 9.4%, global rates.

C. Information Revealing Headers

1) Server Header (and other related server-side headers)

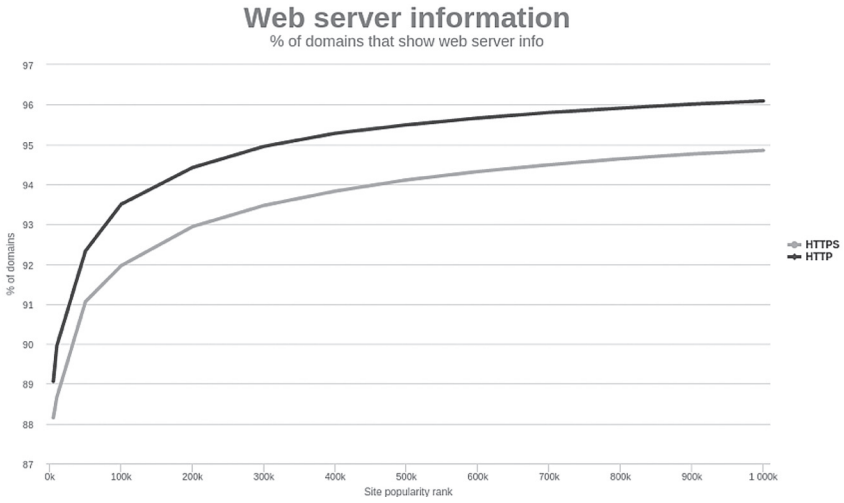
The Server header, defined as part of the RFC 7231 for HTTP/1.1 protocol, is a server-side header originally devised to inform a browser about software used in the web application. Although it is not mandatory, it is issued by most web sites (according to our scanning results, more than 90% of sites set this header). It typically contains the name and version of the web server on which the web application is running.

By itself, the presence of this header as part of an HTTP response does not pose any security thread. However, it may help an attacker to easily obtain the web server name, version and additional information, for example, the name of the CMS supporting the web application. There are currently many "fingerprinting" tools that can be used to obtain that information, regardless of the presence of the Server header. They include well known utilities like command line command nmap, dedicated tool httprint or web utilities like Netcraft that can reveal valuable information to any attacker willing to make use of known vulnerabilities and their corresponding exploits.

The interest of this header from the point of view of the current research is twofold: on the one hand, the *Server* header provides fast and valuable information to those attacks that rely on large scale Internet web site scanning to find potential victims. On the other hand, we want to study statistical correlation between this header and other web site variables, specifically domain popularity and protocol (HTTP / HTTPS) in order to help obtain a more accurate picture of the security of the sites we have analysed.

It should be taken into account that other HTTP headers, besides *Server*, can carry information regarding web server and other relevant software. In particular, we have taken into account the following additional headers: *X-Powered-By, X-AspNet-Version, X-AspNetMvc-Version* and *X-Varnish*. We have combined the information carried by these headers, if present, with the one in the *Server* header, nearly always present, in order to try to find the web server name and version. The results are shown in next figure.

FIGURE 9. WEB SERVER INFORMATION AS A FUNCTION OF WEBSITE POPULARITY



From last figure it is clearly appreciated two tendencies: a) HTTPS sites tend to be more restrictive than HTTP served sites on the information they provide, at least regarding Server and related headers, and b) the more popular a given domain is, the less information it will probably leak. The overall picture, however, shows that a huge amount of Internet web sites (at least, over 85% of them) expose their web server info through their HTTPS headers. Our statistics indicates that, within those sites with recognizable web server, most popular web server is Apache HTTP server (46% of sites) followed by nginx (38% of sites) and IIS (14%).

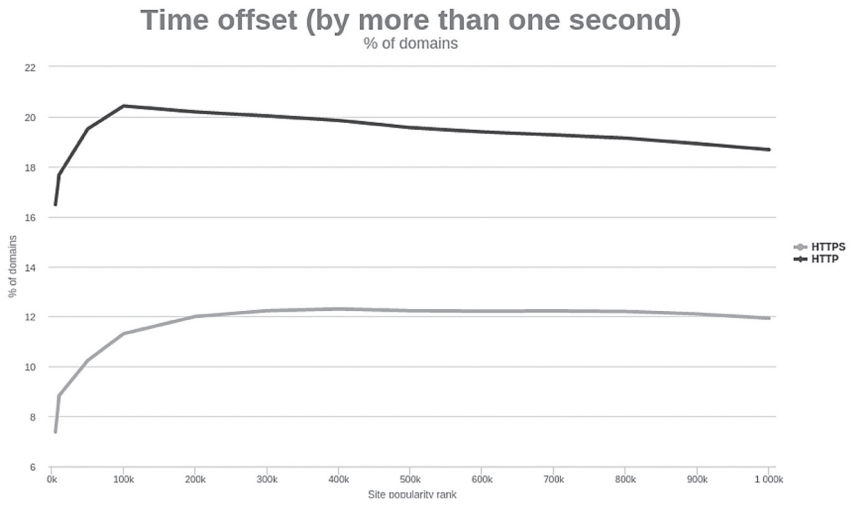
Another common header, *X-Powered-By* header, appears in 48% of responses from sites and typically (66% of cases) contains the PHP version used to develop the web site.

2) Date Header

This header, defined as part of HTTP/1.1 specification, RFC 7231, contains the date and time at which the response message was originated. In our research this header is set in over 99% of all responses. It is, in fact, the most common header seen in responses. Although it is not related to any significant security attack, server-side date and time play an important role as part of the logging information needed to analyse security incidents (see, for example, Prodromou 2016). Inaccurate timestamps will yield unreliable logging records, and therefore making then inappropriate for forensics tasks.

We see the general trends observed previously: a) HTTPS sites seem to run on more precisely configured servers than HTTP ones and b) the more popular a site is, the more secure it tends to be configured.

FIGURE 10. SITES THAT SHOW TIME OFFSET IN *DATE* RESPONSE HEADER AS A FUNCTION OF WEBSITE POPULARITY



3) Referer and Referer-Policy Headers

The *Referer* -sic- header, specified by RFC7231 allows the browser to inform the web server to which the request is made about the URI from where the user made that request. It is meant to provide information that can be processed on the server side for logging or commercial analysis, for example, getting to know where customers typically come from when reaching a given site. This header is also commonly used as a key component of web tracking technologies.

In principle, *Referer* header poses a privacy concern, not a security one, since it reveals information to a third party that a user might not want to be revealed. Sometimes, however, a URL may carry sensitive information, for example, a session token or a capability indicator [Cap2014], and under such circumstances the Referer header may pose a security risk. Both, privacy and possible security risks, have led to the proposal of a *Referrer-Policy* header (see W3C Editor's draft at 2017). This header, currently implemented by all major browsers, allows a website to control the information carried by the *Referer* header in a rather fine-grained manner. It defines up to eight different directives.

Our scanning database indicates that Referrer-Policy header is scarcely implemented. Only 0.05% of HTTP responses, and 0.33% of HTTPS responses, contain some form of a valid *Referrer-Policy*. The distribution of the different policies can be seen in Table 6.

TABLE 6. REFERER DIRECTIVES

Referrer-Policy Directives	HTTP Requests	HTTPS Requests
no-referrer	26.97%	14.86%
no-referrer-when-downgrade	23.63%	29.25%
origin	14.32%	7.06%
origin-when-cross-origin	11.46%	17.73%
same-origin	5.97%	9.18%
strict-origin	5.01%	2.87%
strict-origin-when-cross-origin	7.64%	10.77%
unsafe-URL	0%	0%

Finally, the P3P header, related to users' privacy settings, was not included as part of this study since it is not implemented by any major browser other than Microsoft IE and Edge. However, it is still being issued by 7.5% of the sites (6.9% of HTTP sites and 8.4% of HTTPS ones).

5. CONCLUSIONS

We have presented a new analysis of implementation rate in Alexa's top one million websites of web security policies based on HTTP response headers. A careful data gathering process was carried out to collect HTTP response headers from four different requests for each domain in the list: http://domain, http://www.domain, https://domain and https://www.domain. Redirections were followed. HTTPS issues were examined, finding in particular that a sizeable number of sites, 29.1% of all HTTPS requests made, exhibit some incorrect TLS configuration. They are typically X.509 certificate errors, as the leading causes for TLS misconfiguration are name mismatching and verification errors (self-signed certificates, untrusted CA's or expired certificates). We also compared HTTP response headers obtained from HTTP/1.1 and HTTP/2 equivalent requests and found that, besides some connection related headers, response headers show no significant differences.

We repeatedly showed that security policies based on HTTP response headers are always far more common in HTTPS websites that in HTTP sites. Those policies are also noticeably more commonly implemented among highly popular sites than not so popular ones. In fact, for all security headers analysed here, when implementation rates are depicted against website popularity the resulting curve follows an exponential decline pattern.

In particular, we have found that HTTP Strict Transport Security policy is implemented in about 38% among top one thousand HTTPS sites, but only 17.5% considering all top one million websites. Content Security Policy, despite its powerful prevention capability against Cross Site Scripting and other vulnerabilities, remains poorly implemented at a global 1.6% among all one million websites. HTTPS sites show a markedly larger adoption rate, 3.4%, whereas HTTP sites hardly implement this policy, only 0.4% of them. Session cookies were also analysed and we found that about 50% of sites do not set their *HttpOnly* flag (55.4 % of HTTP sites and 42.7% of HTTPS sites) and and the *Secure* directive is issued for the session cookies in about 19.3% of all HTTPS sites. Although not so relevant as these headers, other security-related response headers were analysed (*X-Frame-Options*, *X-XSS-Protection* and *X-ContentType-Options*). We also analysed information leakage from web servers through their Server and other related response headers and, again, we found that information leakage is more common among less popular and HTTP sites than in highly popular and HTTPS sites.

All in all, security policies based on HTTP headers remain low. They are slightly increasing when compared to the figures reported by previous researches during 2017 (Helme 2017, April 2017), but still well below satisfactory rates. Notably higher implementation rates observed in the most popular sites suggests that security awareness could be influenced by factors like business size. Alternatively, it may be argued that security-aware websites tend to thrive better.

6. FUTURE WORK

The authors plan to expand the current research in several ways. The survey offers a picture of certain web security policies implemented by the top one million websites at a given time (September 2017) and a periodical repetition of the scanning process will be interesting, as it will show how the adoption of these policies are evolving. Following similar initiatives like the ones by Mozilla Observatory and Scott Helme, the authors plan to assign a "global" scoring (e.g., from A to F) for each website and generate the corresponding global statistics and their correlation to HTTPS and site popularity ranking. However, our initial work on this area shows that it is far from obvious how to assign relative weights to each of the analysed HTTP headers and we firmly believe that further work is needed, taking into account, at least, web vulnerability prevalence statistics.

Additionally, we are currently exploring the possibility of considering more variables in our work, like website country and Content Distribution Network usage and how

it relates to web security policies based on HTTP response headers. Finally, we deem interesting to study Subresource Integrity current adoption resources.

7. REFERENCES

Alexa's Top One Million Websites, https://s3.amazonaws.com/alexa-static/top-1m.csv.zip.

Li Chang, Hsu-Chun Hsiao, Wei Jeng, Tiffany Hyun-Jin Kim and Wei-Hsi Lin, "Security Implications of Redirection Trail in Popular Websites Worldwide", in *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*, 2017, pages 1491-1500. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland. DOI: https://doi.org/10.1145/3038912.3052698.

Cisco Umbrella 1 Million, https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/.

Scott Helme, "I am giving up on HPKP", blog post, 24 August 2017, https://scotthelme.co.uk/im-giving-up-on-hpkp/.

Scott Helme, "Alexa Top 1 Million Analysis - August 2017", blog post, 29 August 2017, https://scotthelme.ghost.io/alexa-top-1-million-analysis-aug-2017/.

Scott Helme, "Daily scans of the top one million sites", https://scans.io/study/scott-top-one-million.

Httprint web server fingerprinting tool, http://www.net-square.com/httprint.html.

Hyper: HTTP/2 for Python, https://python-hyper.org/en/latest/.

Internet Engineering Task Force, "HTTP Header Field X-Frame-Options", IETF Informational, RFC 7034, October 2013, https://tools.ietf.org/html/rfc7034.

Internet Engineering Task Force, "HTTP State Management Mechanism", IETF Standard, RFC 6265, April 2011, https://tools.ietf.org/html/rfc6265.

Internet Engineering Task Force, "HTTP Strict Transport Security (HSTS)", IETF Standard, RFC 6797, November 2012, https://tools.ietf.org/html/rfc6797.

Internet Engineering Task Force, "Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content", IETF Standard, RFC 7231, June 2014, https://tools.ietf.org/html/rfc7231.

Internet Engineering Task Force, "Public Key Pinning Extension for HTTP", IETF Standard, RFC 7469, April 2015, https://tools.ietf.org/html/rfc7469.

Internet Engineering Task Force, "Same-Site Cookies", IETF Internet-Draft Standard, 20 June 2016, https://tools.ietf.org/html/draft-ietf-httpbis-cookie-same-site-00.

April King, "Analysis of the Alexa Top 1M sites (June 2017)", 13 June 2017, https://pokeinthe.io/2017/06/13/state-of-security-alexa-top-one-million-2017-06/.

Let's Encrypt - Free SSL/TLS Certificates, https://letsencrypt.org/.

Majestic Million database, https://blog.majestic.com/development/majestic-million-csv-daily/.

Mozilla HTTP Observatory Website, https://mozilla.github.io/http-observatory-website/.

Mozilla Included CA Certificate List, https://wiki.mozilla.org/CA/Included_Certificates.

Mozilla Intermediate Certificates, https://wiki.mozilla.org/CA/Intermediate_Certificates.

Netcraft Site Report, http://toolbar.netcraft.com/site_report.

Nmap, https://nmap.org/.

OWASP, Clickjacking Defence Cheat Sheet, https://www.owasp.org/index.php/Clickjacking_defence_Cheat_Sheet.

Chris Palmer, "Intent to deprecate and remove HPKP", forum post, 27 October 2017, https://groups.google.com/a/chromium.org/forum/#!msg/blink-dev/he9tr7p3rZ8/eNMwKPmUBAAJ.

Agathoklis Prodromou, "Using logs to investigate a web application attack", blog post, 11 May 2016, https://www.acunetix.com/blog/articles/using-logs-to-investigate-a-web-application-attack/.

Requests: HTTP for humans. v2.18.4 Python library, http://docs.python-requests.org/en/master/.

Ivan Ristic, "Is HTTP Public Key Pinning dead?", blog post, 6 September 2016, https://blog.qualys.com/ssllabs/2016/09/06/is-http-public-key-pinning-dead.

Gustav Rydstedt, Elie Bursztein, Dan Boneh and Collin Jackson, "Busting Frame Busting: a Study of Clickjacking Vulnerabilities at Popular Sites", in *IEEE Oakland Web 2.0 Security and Privacy (W2SP 2010)*, https://crypto.stanford.edu/~dabo/pubs/papers/framebust.pdf.

[IO] Security Headers Website, https://securityheaders.io/.

Aditya Sood and Richard Enbody, "The state of HTTP declarative security in online banking websites", in *Computer Fraud & Security*, Volume 2011, Issue 7, July 2011, pages 11-16. DOI: https://doi.org/10.1016/S1361-3723(11)70073-2.

Usage of HTTP/2 for websites, W3Techs, https://w3techs.com/technologies/details/ce-http2/all/all.

Michael Weissbacher, Tobias Lauinger and William Robertson, "Why Is CSP Failing? Trends and Challenges in CSP Adoption", in *Research in Attacks, Intrusions and Defenses*. RAID 2014. Lecture Notes in Computer Science, vol 8688. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-11379-1_11.

World Wide Web Consortium, "Content Security Policy 1.0" (CSP1), W3C Working Group Note, 19 February 2015, discontinued, https://www.w3.org/TR/CSP1/.

World Wide Web Consortium, "Content Security Policy Level 2" (CSP2), W3C Recommendation, 15 December 2016, https://www.w3.org/TR/CSP2.

World Wide Web Consortium, "Content Security Policy Level 3" (CSP3), W3C Working Draft, 13 September 2016, https://www.w3.org/TR/CSP/.

World Wide Web Consortium, "Good Practices for Capability URLs", W3C First Public Working Draft, 18 February 2014, https://www.w3.org/TR/capability-urls/.

World Wide Web Consortium, "Referrer Policy", W3C Candidate Recommendation, 26 January 2017, https://www.w3.org/TR/referrer-policy/.

World Wide Web Consortium, "Subresource Integrity", W3C Recommendation, 23 June 2016, https://www.w3.org/TR/SRI/.

Ming Ying and Shu Qin Li, "CSP adoption: current status and future prospects", in *Security and Communication Networks*, Vol 9, Issue 17, 25 November 2016, pages 4557-4573. DOI: https://doi.org/10.1002/sec.1649.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

On the Effectiveness of Machine and Deep Learning for Cyber Security

Giovanni Apruzzese
Department of Engineering
'Enzo Ferrari'
University of Modena
and Reggio Emilia
Modena, Italy
giovanni.apruzzese@unimore.it

Michele Colajanni
Department of Engineering
'Enzo Ferrari'
University of Modena
and Reggio Emilia
Modena, Italy
michele.colajanni@unimore.it

Luca Ferretti
Department of Engineering
'Enzo Ferrari'
University of Modena
and Reggio Emilia
Modena, Italy
luca.ferretti@unimore.it

Alessandro Guido
Department of Engineering
'Enzo Ferrari'
University of Modena
and Reggio Emilia
Modena, Italy
alessandro.guido@unimore.it

Mirco Marchetti
Department of Engineering
'Enzo Ferrari'
University of Modena
and Reggio Emilia
Modena, Italy
mirco.marchetti@unimore.it

Abstract: Machine learning is adopted in a wide range of domains where it shows its superiority over traditional rule-based algorithms. These methods are being integrated in cyber detection systems with the goal of supporting or even replacing the first level of security analysts. Although the complete automation of detection and analysis is an enticing goal, the efficacy of machine learning in cyber security must be evaluated

with the due diligence. We present an analysis, addressed to security specialists, of machine learning techniques applied to the detection of intrusion, malware, and spam. The goal is twofold: to assess the current maturity of these solutions and to identify their main limitations that prevent an immediate adoption of machine learning cyber detection schemes. Our conclusions are based on an extensive review of the literature as well as on experiments performed on real enterprise systems and network traffic.

Keywords: *machine learning, deep learning, cyber security, adversarial learning*

1. INTRODUCTION

The appeal and pervasiveness of machine learning (ML) is growing. Existing methods are being improved, and their ability to understand and answer real issues is highly appreciated. These achievements have led to the adoption of machine learning in several domains, such as computer vision, medical analysis, gaming and social media marketing [1]. In some scenarios, machine learning techniques represent the best choice over traditional rule-based algorithms and even human operators [2]. This trend is also affecting the cyber security field where some detection systems are being upgraded with ML components [3]. Although devising a completely automated cyber defence system is yet a distant objective, first level operators in Network and Security Operation Centres (NOC and SOC) may benefit from detection and analysis tools based on machine learning. This paper is specifically addressed to security operators and aims to assess the current maturity of these solutions, to identify their main limitations and to highlight some room for improvement.

Our study is based on an extensive review of the literature and on original experiments performed on real, large enterprises and network traffic. Other academic papers compare ML solutions for cyber security by considering one specific application (e.g.: [4], [3], [5]) and are typically oriented to Artificial Intelligence (AI) experts rather than to security operators. In the evaluation, we exclude the commercial products based on machine learning (or on the abused AI term) because vendors do not reveal their algorithms and tend to overlook issues and limitations. First, we present an original taxonomy of machine learning cyber security approaches. Then, we map the identified classes of algorithms to three problems where machine learning is currently applied: intrusion detection, malware analysis, spam and phishing detection. Finally, we analyse the main limitations of existing approaches. Our study highlights pros and cons of different methods, especially in terms of false positive or false negative alarms. Moreover, we point out a general underestimation of the complexity of managing ML architectures in cyber security caused by the lack of publicly available and labelled

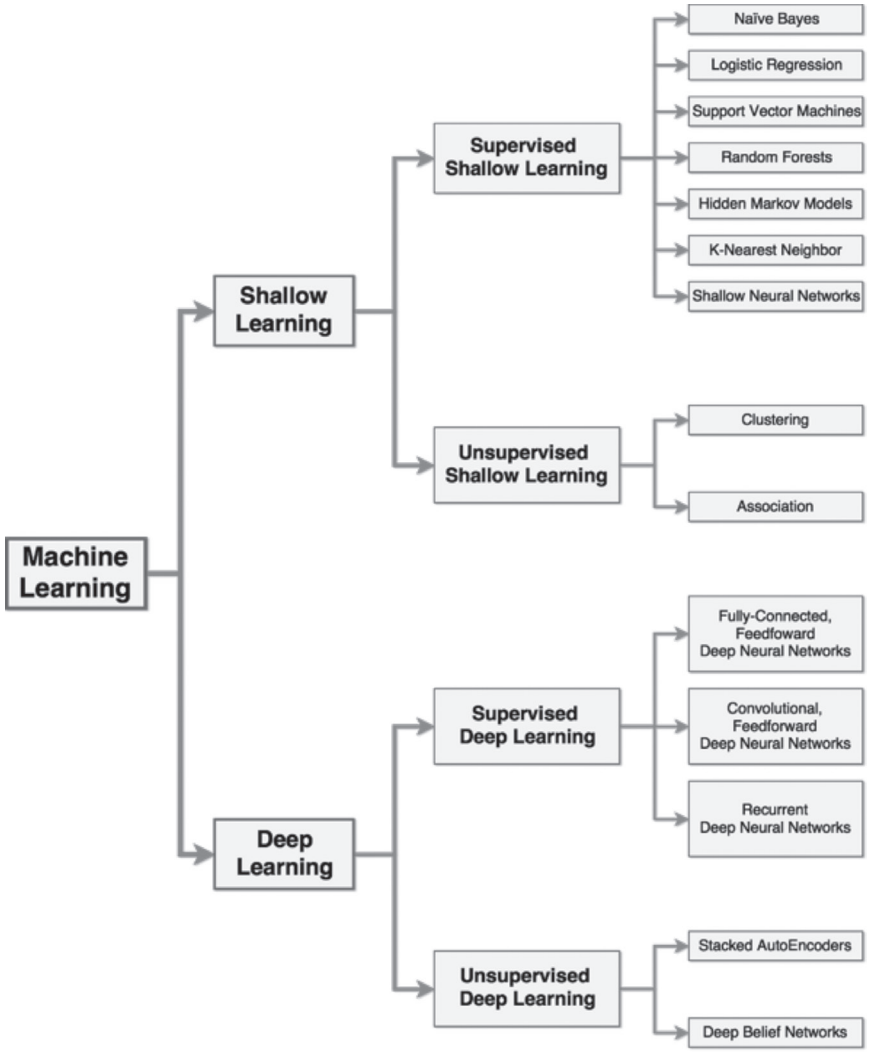
data for training, and by the time required for fine-tuning operations in a domain characterized by continuous change. We also consider recent results emphasizing the effectiveness of *adversarial attacks* [6] [5] in evading ML detectors. The evidenced drawbacks pave the way to future improvements that ML components require before being fully adopted in cyber defence platforms.

This paper is structured as follows. Section 2 proposes our original taxonomy of ML algorithms applied to cyber security. Section 3 outlines the three classes of cyber security problems considered in this paper and Section 4 compares and evaluates ML solutions for cyber security. Section 5 concludes the paper with some final remarks.

2. CLASSIFICATION OF MACHINE LEARNING ALGORITHMS FOR CYBER SECURITY

Machine learning includes a large variety of paradigms in continuous evolution, presenting weak boundaries and cross relationships. Furthermore, different views and applications may lead to different classifications. Hence, we cannot refer to one fully accepted taxonomy from literature, but we prefer to propose an original taxonomy able to capture the differences among the myriad of techniques that are being applied to cyber detection, as shown in Figure 1. This taxonomy is specifically oriented to security operators and avoids the ambitious goal of presenting the ultimate classification that can satisfy all AI experts and application cases. The first discriminant evidenced in Figure 1 is between the traditional ML algorithms, which today can be referred to as **Shallow Learning (SL)**, in opposition to the more recent **Deep Learning (DL)**. Shallow Learning requires a domain expert (that is, a *feature engineer*) who can perform the critical task of identifying the relevant data characteristics before executing the SL algorithm. Deep Learning relies on a multi-layered representation of the input data and can perform feature selection autonomously through a process defined *representation learning*.

FIGURE 1. CLASSIFICATION OF ML ALGORITHMS FOR CYBER SECURITY APPLICATIONS.



SL and DL approaches can be further characterized by distinguishing between *supervised* and *unsupervised* algorithms. The former techniques require a training process with a large and representative set of data that have been previously classified by a human expert or through other means. The latter approaches do not require a pre-labelled training dataset. In this section, we consider and compare the most popular categories of ML algorithms, which appear as the leaves of the classification tree in Figure 1. We remark that each category can include dozens of different techniques[1].

1 For a detailed list of existing ML algorithms, see: https://cran.r-project.org/web/views/MachineLearning.html

A. Shallow Learning

1) Supervised SL algorithms

- **Naïve Bayes (NB).** These algorithms are probabilistic classifiers which make the a-priori assumption that the features of the input dataset are independent from each other. They are scalable and do not require huge training datasets to produce appreciable results.
- **Logistic Regression (LR).** These are categorical classifiers that adopt a discriminative model. Like NB algorithms, LR methods make the a-priori independency assumption of the input features. Their performance is highly dependent on the size of the training data.
- **Support Vector Machines (SVM).** These are non-probabilistic classifiers that map data samples in a feature space with the goal of maximizing the distance between each category of samples. They do not make any assumption on the input features, but they perform poorly in multi-class classifications. Hence, they should be used as binary classifiers. Their limited scalability might lead to long processing times.
- **Random Forest (RF).** A random forest is a set of *decision trees*, and considers the output of each tree before providing a unified final response. Each decision tree is a conditional classifier: the tree is visited from the top and, at each node, a given condition is checked against one or more features of the analysed data. These methods are efficient for large datasets and excel at multiclass problems, but deeper trees might lead to overfitting.
- **Hidden Markov Models (HMM).** These model the system as a set of states producing outputs with different probabilities; the goal is to determine the sequence of states that produced the observed outputs. HMM are effective for understanding the temporal behaviour of the observations, and for calculating the likelihood of a given sequence of events. Although HMM can be trained on labelled or unlabelled datasets, in cyber security they have mostly been used with labelled datasets.
- **K-Nearest Neighbour (KNN).** KNN are used for classification and can be used for multi-class problems. However, both their training and test phase are computationally demanding as to classify each test sample, they compare it against all the training samples.
- **Shallow Neural Network (SNN).** These algorithms are based on neural networks, which consist in a set of processing elements (that is, *neurons*) organized in two or more communicating layers. SNN include all those types of neural networks with a limited number of neurons and layers. Despite the existence of unsupervised SNN, in cyber security they have mostly been used for classification tasks.

2) Unsupervised SL algorithms

- **Clustering.** These group data points that present similar characteristics. Well known approaches include k-means and *hierarchical* clustering. Clustering methods have a limited scalability, but they represent a flexible solution that is typically used as a preliminary phase before adopting a supervised algorithm or for anomaly detection purposes.
- **Association.** They aim to identify unknown patterns between data, making them suitable for prediction purposes. However, they tend to produce an excessive output of not necessarily valid rules, hence they must be combined with accurate inspections by a human expert.

B. Deep Learning

All DL algorithms are based on Deep Neural Networks (DNN), which are large neural networks organized in many layers capable of autonomous representation learning.

1) Supervised DL algorithms

- **Fully-connected Feedforward Deep Neural Networks (FNN).** They are a variant of DNN where every neuron is connected to all the neurons in the previous layer. FNN do not make any assumption on the input data and provide a flexible and general-purpose solution for classification, at the expense of high computational costs.
- **Convolutional Feedforward Deep Neural Networks (CNN).** They are a variant of DNN where each neuron receives its input only from a subset of neurons of the previous layer. This characteristic makes CNN effective at analysing spatial data, but their performance decreases when applied to non-spatial data. CNN have a lower computation cost than FNN.
- **Recurrent Deep Neural Networks (RNN).** A variant of DNN whose neurons can send their output also to previous layers; this design makes them harder to train than FNN. They excel as sequence generators, especially their recent variant, the *long short-term memory*.

2) Unsupervised DL algorithms

- **Deep Belief Networks (DBN).** They are modelled through a composition of *Restricted Boltzmann Machines* (RBM), a class of neural networks with no output layer. DBN can be successfully used for pre-training tasks because they excel in the function of feature extraction. They require a training phase, but with unlabelled datasets.

- **Stacked Autoencoders (SAE)**. They are composed by multiple *Autoencoders*, a class of neural networks where the number of input and output neurons is the same. SAE excel at pre-training tasks similarly to DBN, and achieve better results on small datasets.

3. APPLICATIONS OF MACHINE LEARNING ALGORITHMS TO CYBER SECURITY

We consider the three areas where most cyber ML algorithms are finding application: *intrusion detection, malware analysis*, and *spam detection*. An outline of each field is presented below.

Intrusion detection aims to discover illicit activities within a computer or a network through Intrusion Detection Systems (IDS). *Network* IDS are widely deployed in modern enterprise networks. These systems were traditionally based on patterns of known attacks, but modern deployments include other approaches for anomaly detection, threat detection [7] and classification based on machine learning. Within the broader intrusion detection area, two specific problems are relevant to our analysis: the detection of *botnets* and *of Domain Generation Algorithms* (DGA). A botnet is a network of infected machines controlled by attackers and misused to conduct multiple illicit activities. Botnet detection aims to identify communications between infected machines within the monitored network and the external command-and-control servers. Despite many research proposals and commercial tools that address this threat, several botnets still exist. DGA automatically generate domain names, and are often used by an infected machine to communicate with external server(s) by periodically generating new hostnames. They represent a real threat for organizations because, through DGA which relies on language processing techniques, it is possible to evade defences based on static blacklists of domain names. We consider DGA detection techniques based on ML.

Malware analysis is an extremely relevant problem because modern malware can automatically generate novel variants with the same malicious effects but appearing as completely different executable files. These polymorphic and metamorphic features defeat traditional rule-based malware identification approaches. ML techniques can be used to analyse malware variants and attributing them to the correct malware family.

Spam and phishing detection includes a large set of techniques aimed at reducing the waste of time and potential hazard caused by unwanted emails. Nowadays, unsolicited emails, namely *phishing*, represent the preferred way through which an attacker establishes a first foothold within an enterprise network. Phishing emails

include malware or links to compromised websites. Spam and phishing detection is increasingly difficult because of the advanced evasion strategies used by attackers to bypass traditional filters. ML approaches can improve the spam detection process.

TABLE 1. APPLICATION OF ML TO CYBER SECURITY PROBLEMS.

		Intrusion Detection			Malware Analysis	Spam Detection
		Network	Botnet	DGA		
Deep Learning	Supervised	RNN [8]	RNN [9]		FNN [10] CNN [11] RNN [12]	
	Unsupervised	DBN [13] SAE [14]			DBN [15] SAE [16]	DBN [17] SAE [18]
Shallow Learning	Supervised	RF [3] NB [3] SVM [3] LR [3] HMM [3] KNN [3] SNN [3]	RF [19] NB [19] SVM [19] LR [20] KNN [21] SNN [22]	RF [23] HMM [23]	RF [24] NB [24] SVM [24] LR [24] HMM [25] KNN [24] SNN [26]	RF [27] NB [28] SVM [28] LR [27] KNN [27] SNN [27]
	Unsupervised	Clustering [29] Association [30]	Clustering [5]	Clustering [31]	Clustering [24] Association [32]	Clustering [33] Association [34]

In Table 1 we report the main ML algorithms that have been proposed to address the previously identified cyber security problems. In this table, rows report the family of algorithms presented in Section 2, while columns denote cyber issues. Each cell indicates which ML algorithms are used for each problem; empty cells denote that, to the best of our knowledge, there is no proposal for that class of problems. From this table, it emerges that SL algorithms are applied to all considered problems. Supervised DL algorithms find wide application to malware analysis, less to intrusion detection; spam detection relies only on unsupervised DL algorithms. Despite its relatedness to natural language processing [2], no DL algorithm is applied to DGA detection. As expected, the overall number of algorithms based on DL is considerably smaller than those based on SL. Indeed, DL proposals based on huge neural networks are more recent than SL approaches. This gap opens many research opportunities.

Finally, we highlight a significant difference among supervised and unsupervised approaches: the former algorithms are used for classification purposes and can implement complete detectors; the latter techniques perform ancillary activities [35]. Unsupervised SL algorithms are often used for grouping data with similar characteristics independently of predefined classification criteria, and excel at identifying useful features whenever the data to be analysed present high dimensionality [16].

4. EVALUATION

In this section we present seven issues that must be considered before deciding whether to apply ML algorithms in NOC and SOC. We can anticipate that, at the current state-of-the-art, no algorithm can be considered fully autonomous with no human supervision. We substantiate each issue through experimental results from literature or original experiments performed on large enterprises. We begin by describing the testing environments of our experiments, and the metrics considered for evaluation. The experiments focus on DGA Detection and Network Intrusion Detection, and leverage two ML algorithms: Random Forest and Feedforward Fully Connected Deep Neural Network.

For **DGA Detection**, we compose two labelled training datasets containing both DGA and non-DGA domains. The former dataset contains DGA created through known techniques, while the latter contains DGA created using more recent approaches. Non-DGA domains are randomly chosen among the Cisco Umbrella top-1 million. We report the meaningful metrics of the training datasets in Table 2. Moreover, we build a testing dataset of 10,000 domains extracted evenly from each of the training datasets. We also rely on a real and unlabelled dataset composed of almost 20,000 domains contacted by a large organization. The features extracted for this dataset are: *n-gram* normality score [36]; meaningful characters ratio [36]; number-to-character ratio; vowel-to-consonant ratio; and domain length. These datasets are used to train and test a self-developed Random Forest classifier composed of 100 decision trees leveraging the CART (classification and regression tree) algorithm.

TABLE 2. TRAINING DATASETS FOR DGA DETECTION EXPERIMENTS.

Dataset	DGA technique	DGA count	non-DGA count
1	Well-known	21,355	20,227
2	Well-known and recent	37,673	8,120

For **Network Intrusion Detection**, we use three labelled real training datasets composed of benign and malicious network flows[2] collected in a large organization of nearly 10,000 hosts. The labels are created by flagging as malicious those flows that raised alerts by the enterprise network IDS and reviewed by a domain expert. Meaningful metrics of these training datasets are reported in Table 3. We also generate a testing dataset of 50,000 flows evenly extracted among the training datasets. The considered features for these datasets include: source/destination IP address, source/destination port, number of incoming/outgoing bytes and packets, TCP flags, protocol used, duration of the flow and list of alerts raised. These datasets are used to test and train two self-developed classifiers, one based on Random Forests and one on

2 *Cisco Netflow:* https://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netflow/index.html

Feedforward Fully-connected Deep Neural Network. Different topologies have been considered for each algorithm. The RF is composed by 100 decision trees leveraging the CART algorithm. For the FNN, the overall number of neurons ranges from 128 to 16,384, distributed between 2 to 16 layers; the hidden layers leverage the *ReLU* activation function, whereas the output layer uses a *sigmoid* activation function.

TABLE 3. TRAINING DATASETS FOR NETWORK INTRUSION DETECTION EXPERIMENTS.

Dataset	Malicious flows	Benign flows
1	1,000	100,000
2	2,500	250,000
3	5,000	500,000

The quality of each classifier is measured through common performance metrics, namely *Precision, Recall, F1-score*, which are computed as follows:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where *TP*, *FP*, and *FN* denote true positives, false positives, and false negatives, respectively. For completeness, we consider a true positive to be a correct detection of a malicious sample. Precision indicates how much a given approach is likely to provide a correct result. Recall is used to measure the detection rate. The F1-score combines Precision and Recall into a single value. We do not rely on Accuracy[3] because, in a real organization, the number of legitimate events is several orders of magnitude greater than illegitimate events. Hence, all the Accuracy values are close to 1 and these results prevent capturing the true effectiveness of a classifier. Finally, to reduce the possibility of biased results, each evaluation metric is computed after performing 10-fold cross validation.

A. Shallow vs Deep Learning

Deep Learning is known to outperform Shallow Learning in some applications, such as computer vision [2]. This is not always the case for cyber security where some well configured SL algorithms may prevail, even given the DL proposals are scarce with respect to SL techniques in this domain. Just to give an example, we experimentally compare the performance of the two self-developed classifiers for Network Intrusion Detection, one based on RF (Shallow Learning) and another based on FNN (Deep Learning). Both are trained with the third dataset described in Table 3 and tested on the network intrusion detection testing dataset. To obtain more refined results, we repeat the training and test phase of these classifiers multiple times using different topologies. In Table 4, we show the classification results achieved by each method; for the FNN we report the results obtained by the best topology consisting

3 $Accuracy = \frac{TP+FN}{TP+TN+FN+FP}$, where *TN* denotes true negatives.

in 1.024 neurons spread across 4 hidden layers. The RF classifier performed better than the FNN, with an F1-score of nearly 0.8, against the 0.6 obtained by the FNN. Our takeaway is that security administrators should not be charmed by the alluring neuronal multi-layer approach offered by Deep Learning, as some of these methods might still be immature for cyber security.

TABLE 4. COMPARISON BETWEEN DL AND SL CLASSIFIERS.

Classifier	F1-score	Precision	Recall
Random Forest (SL)	0.7985	0.8727	0.736
Fully-connected Feedforward Deep Neural Network (DL)	0.6085	0.7708	0.5027

B. General vs specific detectors

Products based on machine learning are often promoted by vendors as catch-all solutions to a broad array of cyberattacks. However, unbiased experimental results show that ML algorithms may provide superior performance when they focus on *specific* threats instead of trying to detect multiple threats at once. We devise multiple intrusion detection systems based on the self-developed RF classifiers for network intrusion detection, each focusing on a specific type of attack, such as buffer overflows, malware infection, DoS. The training dataset for each classifier is based on the third dataset presented in Table 3. We train and test each classifier, and then compare their classification results with the classifier described in the first row of Table 4 that is our baseline. Table 5 shows the Precision, Recall and F1-score of the six classifiers that obtained the best results, alongside the baseline reported in the bottom row. These attack-specific classifiers obtain promising results on real traffic data with F1-scores of over 0.95, while the 'general-purpose' classifier performs significantly poorly. We conclude that entrusting a single ML detector to identify malicious flows is an enticing but as yet unfeasible goal. On the other hand, by having multiple detectors, each focusing on one attack type, it is possible to produce a defensive scheme with superior detection capabilities.

TABLE 5. CLASSIFICATION RESULTS FOR ATTACK-SPECIFIC CLASSIFIERS AND THE GENERAL CLASSIFIER.

Attack Name	F1-score	Precision	Recall
DOS attempt	0.9953	0.9938	0.9969
Overflow attempt	0.9939	0.9933	0.9946
SSH Brute Force login	0.9916	0.9941	0.9892
Suspicious DNS query	0.9753	0.9953	0.9586
Cache Poisoning attempt	0.9676	0.9872	0.9506
Possible Malware infection	0.9587	0.9939	0.9337
General approach (baseline)	0.7985	0.8727	0.7360

C. Vulnerability to adversarial attacks

Competent adversaries use novel strategies to evade detectors based on machine learning algorithms [5]. These activities, namely *adversarial attacks*, may attack the integrity, the availability, or the privacy of the target system [6]. Integrity violations evade a classification or a clustering algorithm by producing attacks classified as licit activities. Availability violations produce a multitude of normal events that are classified as an attack thus causing detectors to raise a huge amount of false alarms. Privacy violations let the attacker acquire information on the target network by exploiting the defensive ML algorithm. Moreover, recent advances in Deep Learning led to the development of *generative adversarial networks* (GAN) [37], which are DNN capable of automatically producing adversarial samples against a target ML system.

TABLE 6. DETECTION RATES OF THE RF CLASSIFIER AGAINST DIFFERENT DGA TECHNIQUES [36].

DGA method	Recall
corebot	1
cryptolocker	1
dircrypt	0.99
kraken_v2	0.96
lockyv2	0.97
pykspa	0.85
qakbot	0.99
ramdo	0.99
ramnit	0.98
simda	0.96
DeepDGA GAN	**0.48**

To demonstrate the effectiveness of a GAN in evading classifiers we analyse the case study of DeepDGA [36]. The authors initially train an RF classifier to detect DGA using known datasets, and then show that this classifier identifies DGA with good detection rates. Then, they develop a GAN to generate domains that evade such classifier. Results are presented in Table 6, where the first ten rows show the detection rate against ten real DGA, while the last row denotes the detection rate against samples generated by the DeepDGA GAN. We observe that the performance of the classifier (always above 0.85, and above 0.96 for nine out of ten DGA) drops below 50% for GAN-generated samples.

TABLE 7. DETECTION RATES OF THE RF CLASSIFIER AGAINST DIFFERENT DGA BEFORE AND AFTER HARDENING [36].

DGA method	Baseline Recall	Hardened Recall
corebot	0.97	**0.97**
dircrypt	0.95	0.93
qakbot	0.94	**0.94**
ramnit	0.94	**0.94**
lockyv2	0.87	0.84
cryptolocker	0.87	**0.88**
simda	0.75	**0.79**
krakenv2	0.72	**0.76**
pykspa	0.67	**0.71**
ramdo	0.54	**0.54**

To counter adversarial attacks, novel proposals introduce the paradigm of *adversarial learning* [6], in which adversarial samples are included in the training dataset to harden the ML detector. As an example, authors in [36] demonstrate the advantages of adversarial learning by enriching the training set of the classifier with adversarial samples produced by the GAN. Table 7 compares the detection rates of the RF classifier before and after this hardening process. Cells with a grey background represent the DGA for which the detection rate improved after adversarial learning (it should be noted that the dataset used for this test is different than that used for the experiments reported in Table 6). Detection rates for 8 out of 10 DGA families improved, thus showing the validity of adversarial learning.

D. Selection of a machine learning algorithm

Unbiased comparison of the effectiveness of two ML algorithms requires that they are both trained on the *same* training dataset and tested on the *same* dataset [3]. Even though many cyber security proposals rely on few and old public datasets, their results are not comparable due to several causes: the two algorithms consider different features; one or both algorithms may implement pre-filtering operations that alter the training dataset; and they may use a different split between test and training dataset. For these reasons, meaningful comparisons between detection performance in literature are extremely difficult. For example, papers such as [4] and [5] discuss ML methods for two cyber security problems, but they do not consider the different training and testing environments of the analysed works. Hence, although some solutions achieve higher accuracy than others, it is possible that results change significantly under different training settings. Furthermore, there is no guarantee that a method performing best on a test dataset confirms its superiority on different datasets.

Security administrators should be aware of this issue, and should thoroughly question the evaluation methodology before accepting the performance results of different machine learning algorithms.

E. False positives and false negatives

The implicit cost of a misclassification in the cyber security domain is a serious problem. False positives in malware classification and intrusion detection annoy security operators and hinder remediation in case of actual infection. In phishing detection, they might cause important, legitimate messages to not be delivered to end users. In contrast, failing to detect malware, a network intrusion or a phishing email can compromise an entire organization. We explore this problem by considering the performance of ML solutions devoted to malware analysis and phishing detection [27], while we perform an original experiment for intrusion detection that is oriented to detect DGA in a real, large enterprise.

For malware analysis, we consider the approach in [24] that proposes an original and effective method for malware classification. This paper contains a detailed analysis and comparison of different ML techniques which were trained and tested on the same datasets, thus satisfying the requirements for valid comparison of different techniques. Hence, we deem this paper to be a good representation of the state-of-the-art of ML for determining the family to which a malware sample belongs. The evaluation is performed on the DREBIN dataset;[4] for large malware families the proposed approach, which outperforms all other baselines, obtains an F1-score of 0.95, whereas for small malware families it achieves an F1-score of 0.89.

For phishing detection, we report the results described in [27] that, to the best of our knowledge, is the only paper on phishing email detection which compares different ML algorithms against the same comprehensive dataset. Therefore, we consider this work as a valid overview of the efficacy of different ML methods. The authors created a custom dataset of ~3,000 phishing emails on which several ML classifiers were tested: the best results were obtained by RF (lowest false positives) and LR (lowest false negatives), obtaining an F1-score of 0.90 and 0.89, respectively.

The scenario for intrusion detection is different, as modern solutions can achieve higher Accuracy scores [3]. Although near-perfect Accuracy may seem an appreciable result, the massive amounts of events generated daily in a large enterprise account for hundreds to thousands of false positives that need to be manually triaged by security operators. We highlight this problem through an original experiment. We consider two DGA detectors based on the self-developed Random Forest classifiers trained on the first and second datasets of Table 2, respectively. We then validate them on the real domain dataset. Results are summarized in Table 8 which presents the number

4 DREBIN dataset: https://www.sec.cs.tu-bs.de/~danarp/drebin/

of domains that are flagged as DGA by both classifiers, alongside its percentage on the total amount of domains included in the dataset. We can observe that the two classifiers obtain comparable detection performances on real traffic data, as they both signal about 400 domains. However, manual inspection revealed that they were not DGA, hence all the domains flagged as DGAs are actually false positives. As anticipated, even a false positive rate of 2% can account to hundreds of false alarms in a real organization.

TABLE 8. PERFORMANCE OF THE DGA DETECTION CLASSIFIERS WHEN USED ON REAL DATA.

Classifier	Training Dataset	Domains classified as DGA
1	Well-known	431 (2.16%)
2	Well-known and recent	397 (1.99%)

Despite these apparently promising results which are well beyond acceptable levels in other fields such as image recognition, these approaches are affected by an excessive number of false positives and false negatives to be considered for cyber defences without human supervision.

F. Re-training issues

A well-known limitation of traditional detection approaches based on static detection rules is the need for frequent and continuous updates (e.g., daily updates of antivirus definitions). A similar issue also influences advanced ML approaches; reliance on outdated training datasets leads to poor detection performance. This is a critical problem for all supervised learning approaches requiring labelled training datasets; the manual creation of similar datasets is an expensive process because they need to be sufficiently large and comprehensive to allow the algorithm to learn the difference between the classes. Furthermore, these operations are error prone and may lead to incorrect classifications. Finally, most organizations are unwilling to share their internal network data. This scenario leads to an overall scarcity of publicly available and labelled data for cyber security, thus rendering periodic retraining extremely difficult or impossible.

To show the detrimental effects of obsolete training sets, we perform an experiment comparing the performance of two instances of the same self-developed RF classifier for DGA detection. The first and second instances are trained with the first and second datasets reported in Table 2. Both classifiers are tested against the same synthetic domain dataset described in Section 4. We report the results in Table 9, which shows the Precision, Recall and F1-score obtained by the two classifiers for DGA detection. As expected, the performance of the second classifier is significantly better because

it obtains an F1-score for DGAs of 0.89 against 0.33. These results demonstrate that classifier performances are extremely sensitive to the freshness of the training set.

TABLE 9. PERFORMANCE OF THE DGA DETECTION CLASSIFIERS WHEN TRAINED ON OUTDATED AND RECENT DATASETS.

Classifier	Training Dataset	F1-score	Precision	Recall
1	Well-known	0.3306	0.1984	0.9913
2	Well-known and recent	0.8999	0.9126	0.8875

G. Deployment process

Security solutions based on ML achieve appreciable detection rates only if the training dataset is appropriate and the parameters of the algorithms are finely tuned. In most scenarios, these operations are still executed empirically and represent a resource intensive task that presents several risks. If these steps are not performed rigorously and/or training is not based on the right datasets, the results are underwhelming. We highlight these issues through a set of ML experiments applied to network intrusion detection. The goal is to show the considerably different results achieved by the same ML algorithm in different environments where either the number of features or the training dataset is changed. To this purpose, we rely on the RF classifier for network intrusion detection. We train it using the third dataset reported in Table 3 by choosing 5, 7, 10 or 12 features, selected through a *feature agglomeration* process; the testing phase is performed on the test dataset. We report the Precision, Recall and F1-score for the five sets of features in Table 10, where we observe that the same classifier yields different results, especially with regards to its Recall, with values ranging from 0.57 to 0.74.

TABLE 10. PERFORMANCE OF THE INTRUSION DETECTION CLASSIFIER WHEN TRAINED WITH DIFFERENT FEATURES.

Features	F1-score	Precision	Recall
12	0.7985	0.8727	0.7361
10	0.7801	0.8684	0.7093
7	0.7476	0.8893	0.6448
5	0.6920	0.8724	0.5734

Then, we keep the number of features fixed at 12 and we repeat the training process two more times by using the first and then the second dataset reported in Table 3, and then test them on the same testing dataset. Table 11 reports the Precision, Recall and F1-score for the three training datasets. These results confirm that the Recall between the best and the worst case may differ by 10% or over.

TABLE 11. PERFORMANCE OF THE INTRUSION DETECTION CLASSIFIER WHEN TRAINED ON DIFFERENT DATASETS.

Training Dataset	F1-score	Precision	Recall
1	0.7306	0.8753	0.6270
2	0.7757	0.8703	0.6996
3	0.7985	0.8727	0.7361

5. CONCLUSIONS

Machine and deep learning approaches are increasingly employed for multiple applications and are being adopted also for cyber security, hence it is important to evaluate when and which category of algorithms can achieve adequate results. We analyse these techniques for three relevant cyber security problems: intrusion detection, malware analysis and spam detection. We initially propose an original taxonomy of the most popular categories of ML algorithms and show which of them are currently applied to which problem. Then we explore several issues that influence the application of ML to cyber security. Our results provide evidence that present machine learning techniques are still affected by several shortcomings that reduce their effectiveness for cyber security. All approaches are vulnerable to adversarial attacks and require continuous re-training and careful parameter tuning that cannot be automatized. Moreover, especially when the same classifier is applied to identify different threats, the detection performance is unacceptably low; a possible mitigation can be achieved by using different ML classifiers for detecting specific threats. Deep learning is still at an early stage and no final conclusion can be drawn. Significant improvements may be expected, especially considering the recent and promising development of adversarial learning. Our takeaway is that machine learning techniques can support the security operator activities and automate some tasks, but pros and cons must be known. The autonomous capabilities of ML algorithms must not be overestimated, because the absence of human supervision can further facilitate skilled attackers to infiltrate, steal data, and even sabotage an enterprise.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, 2015.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
[3] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, 2015.
[4] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, 2008.
[5] J. Gardiner and S. Nagaraja, "On the Security of Machine Learning in Malware C8C Detection," *ACM Computing Surveys*, 2016.

[6] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial Machine Learning," in *ACM workshop on Security and artificial intelligence*, 2011.
[7] F. Pierazzi, G. Apruzzese, M. Colajanni, A. Guido, and M. Marchetti, "Scalable architecture for online prioritization of cyber threats," in *International Conference on Cyber Conflict (CyCon)*, 2017.
[8] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," in *IEEE International Conference on Platform Technology and Service (PlatCon)*, 2016.
[9] P. Torres, C. Catania, S. Garcia, and C. G. Garino, "An analysis of Recurrent Neural Networks for Botnet detection behavior," in *IEEE Biennial Congress of Argentina (ARGENCON)*, 2016.
[10] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, "Large-scale malware classification using random projections and neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.
[11] G. D. Hill and X. J. Bellekens, "Deep Learning Based Cryptographic Primitive Classification," *arXiv preprint*, 2017.
[12] R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu, and A. Thomas, "Malware classification with recurrent networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
[13] M. Z. Alom, V. Bontupalli, and T. M. Taha, "Intrusion detection using deep belief networks," in *IEEE National Aerospace and Electronics Conference (NAECON)*, 2015.
[14] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, 2016.
[15] Y. Li, R. Ma, and R. Jiao, "A hybrid malicious code detection method based on deep learning," *International Journal of Security and Its Applications*, 2015.
[16] W. Hardy, L. Chen, S. Hou, Y. Ye, and X. Li, "DL4MD: A Deep Learning Framework for Intelligent Malware Detection," in *International Conference on Data Mining* (DMIN), 2016.
[17] G. Tzortzis and A. Likas, "Deep belief networks for spam filtering," in *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2007.
[18] G. Mi, Y. Gao, and Y. Tan, "Apply stacked auto-encoder to spam detection," in *International Conference in Swarm Intelligence*, 2015.
[19] M. Stevanovic and J. M. Pedersen, "An efficient flow-based botnet detection using supervised machine learning," in *IEEE International Conference on Computing, Networking and Communications (ICNC)*, 2014.
[20] S. Ranjan, *Machine learning based botnet detection using real-time extracted traffic features*, Google Patents, 2014.
[21] B. Rahbarinia, R. Perdisci, A. Lanzi, and K. Li, "Peerrush: mining for unwanted p2p traffic," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 2013.
[22] A. Feizollah and e. al, "A study of machine learning classifiers for anomaly-based mobile botnet detection," in *Malaysian Journal of Computer Science*, 2013.
[23] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From throw-away traffic to bots: detecting the rise of DGA-based malware," in *USENIX Security Symposium*, 2012.
[24] T. Chakraborty, F. Pierazzi, and V. Subrahmanian, "Ec2: Ensemble clustering and classification for predicting android malware families," *IEEE Transactions on Dependable and Secure Computing*, 2017.
[25] C. Annachhatre, T. H. Austin, and M. Stamp, "Hidden Markov models for malware classification," *Journal of Computer Virology and Hacking Techniques*, 2015.
[26] J. Demme, M. Maycock, J. Schmitz, A. Tang, A. Waksman, S. Sethumadhavan, and S. Stolfo, "On the feasibility of online malware detection with performance counters," in *ACM SIGARCH Computer Architecture News*, 2013.
[27] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *ACM Proceedings of the Anti-Phishing Working Groups*, 2007.
[28] G. Xiang, J. Hong, C. P. Rose and, L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Transactions on Information and System Security (TISSEC)*, 2011.
[29] G. Apruzzese, M. Marchetti, M. Colajanni, G. Gambigliani Zoccoli, and A. Guido, "Identifying malicious hosts involved in periodic communications," in *IEEE International Symposium on Network Computing and Applications (NCA)*, 2017.
[30] F. S. Tsai, "Network intrusion detection using association rules," *International Journal of Recent Trends in Engineering*, 2009.

[31] F. Bisio, S. Saeli, L. Pierangelo, D. Bernardi, A. Perotti, and D. Massa, "Real-time behavioral DGA detection through machine learning," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, 2017.
[32] Y. Ye, D. Wang, T. Li, D. Ye, and Q. Jiang, "An intelligent PE-malware detection system based on association mining," *Journal in computer virology*, 2008.
[33] W.-F. Hsiao and T.-M. Chang, "An incremental cluster-based approach to spam filtering," *Expert Systems with Applications*, 2008.
[34] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, 2014.
[35] K. Rieck, P. Trinius, C. Willems, and T. Holz, "Automatic Analysis of Malware Behavior Using Machine Learning," *Journal of Computer Security*, 2011.
[36] H. S. Anderson, J. Woodbridge, and B. Filar, "DeepDGA: Adversarially-Tuned Domain Generation and Detection," in *ACM Workshop on Artificial Intelligence and Security*, 2016.
[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.
[38] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, 2009.
[39] R. Zuech, T. M. Khoshgoftaar, and R. Wald, "Intrusion detection and big heterogeneous data: a survey," *Journal of Big Data*, 2015.
[40] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of advances in information technology*, 2010.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Screen Watermarking for Data Theft Investigation and Attribution

David Gugelmann[1]
ETH Zurich
Zurich, Switzerland
david.gugelmann@alumni.ethz.ch

David Sommer[1]
ETH Zurich
Zurich, Switzerland

Vincent Lenders
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Markus Happe
ETH Zurich
Zurich, Switzerland
markus.happe@alumni.ethz.ch

Laurent Vanbever
ETH Zurich
Zurich, Switzerland
lvanbever@ethz.ch

Abstract: Organizations not only need to defend their IT systems against external cyber attackers, but also from malicious insiders, that is, agents who have infiltrated an organization or malicious members stealing information for their own profit. In particular, malicious insiders can leak a document by simply opening it and taking pictures of the document displayed on the computer screen with a digital camera. Using a digital camera allows a perpetrator to easily avoid a log trail that results from using traditional communication channels, such as sending the document via email. This makes it difficult to identify and prove the identity of the perpetrator. Even a policy prohibiting the use of any device containing a camera cannot eliminate this threat since tiny cameras can be hidden almost everywhere.

To address this leakage vector, we propose a novel screen watermarking technique that embeds hidden information on computer screens displaying text documents. The watermark is imperceptible during regular use, but can be extracted from pictures of documents shown on the screen, which allows an organization to reconstruct the

1 equally contributing authors

place and time of the data leak from recovered leaked pictures. Our approach takes advantage of the fact that the human eye is less sensitive to small luminance changes than digital cameras. We devise a symbol shape that is invisible to the human eye, but still robust to the image artifacts introduced when taking pictures. We complement this symbol shape with an error correction coding scheme that can handle very high bit error rates and retrieve watermarks from cropped and compressed pictures. We show in an experimental user study that our screen watermarks are not perceivable by humans and analyze the robustness of our watermarks against image modifications.

Keywords: *data theft, investigation, attribution, screen watermarking, malicious insiders, infiltration*

1. INTRODUCTION

Organizations not only need to protect their proprietary information from external attackers but also from insiders [17], i.e., agents infiltrating the organization or malicious employees. To this end, data loss prevention (DLP) solutions are increasingly deployed. State-of-the-art DLP software can either be configured only to log or additionally to block users' actions, such as accessing the Internet, sending emails, printing, taking screenshots or accessing external media. Consequently, data leakage via these conventional communication channels can either be prevented or there is at least a log trail that shows a perpetrator's actions. This log trail can be used as evidence against the malicious insider in forensic investigations. However, DLP systems cannot prevent insiders from taking pictures of a computer screen with a digital camera. Any employee who is authorized to open a particular document on their computer screen can leak the contained information by taking a picture and sharing it with unauthorized parties. Using a camera allows a perpetrator to easily avoid a log trail, as DLP software cannot detect if a document is being photographed. This makes it difficult to identify and prove the identity of the perpetrator based on a recovered leaked picture. Smartphones with cameras have become ubiquitous and new technologies like digital glasses or lenses are gaining momentum, making this data leakage threat difficult to control [23]. Even a policy prohibiting the use of any device containing a camera cannot eliminate this threat since tiny cameras can be hidden almost everywhere.

We introduce a content-agnostic watermarking approach for textual information displayed on computer screens. The watermark is imperceptible during regular use but can be extracted *a posteriori* from pictures of documents shown on the screen.

This enables an organization to reconstruct the place and time of the data leak from recovered leaked pictures, which greatly facilitates the forensic investigation of data breaches involving leaked pictures of screens. Our contributions are:

- an analysis of the data leakage channel computer screen – digital camera (§3);
- a watermarking schema specifically developed and optimized for this leakage channel (§4); and
- a comprehensive evaluation of the suggested watermarking system – including a user study (§5) – and a discussion of attacks against our attribution approach (§6).

2. RELATED WORK

One can distinguish between watermarking solutions for multimedia files and approaches for text documents. Our scenario shows characteristics of both domains. Watermarks need to be imperceptible on screens showing textual contents and must be retained in pictures of the text.

Basic approaches for images simply place watermarks in the least significant bits of individual pixels of an image [20,2,13]. The resulting small color variations are imperceptible to humans, but most smart phone cameras also cannot capture color variations of individual screen pixels, as we found in preliminary experiments. Caronni [5] encodes the watermark by changing the brightness of multiple contiguous pixels, which is similar to our approach. However, his approach requires the original image for extraction of the watermark, while we do not require the original image. Most advanced multimedia watermarking methods operate in a transformed domain, such as an image's frequency spectrum [7,18,19]. This allows them to embed unnoticeable watermarks by introducing slight modifications in the frequency spectrum. This results in noise patterns in the spatial domain. This noise is not noticeable in colorful images but is usually well visible on text documents [12,1]. Therefore, image watermarking approaches operating in a transformed domain are not suitable for the task at hand.

Existing approaches for watermarking of text documents modify the text directly. Jalil et al. [9] distinguish between *image-based, syntactic*, and *semantic* approaches. Image-based approaches [4,3] adapt the typesetting of the text. Syntactic and semantic approaches modify the text itself. They fragment the text into blocks of words or letters, which are then moved or replaced. However, we have to assume that employees can edit documents. In this case, they will probably notice such text modifications. Furthermore, the integration of text-based watermarks is computational-expensive

and can hardly be embedded in real-time. Hence, text modifications are unsuitable for our scenario.

Piec et al. [16] develop a real-time screen watermarking approach for embedding watermarks into screenshots. Screenshots retain colors perfectly and no geometric distortions occur, which allows them to use standard QR codes with their build-in standard error correction for embedding the watermarks. In contrast, we use custom watermark symbols and error correction codes such that our approach not only works for screenshots, but also for pictures of computer screens, in which various image artifacts are present. Kuhn et al. [11] analyzed in their seminal work various approaches to tamper with as well as eavesdrop on information by modifying and analyzing electromagnetic radiation. However, their work analyzes skilled attackers who use hardware to process electromagnetic radiation, while we focus on an attacker using a commodity camera. Petitcolas et al. [14] present criteria for benchmarking watermark approaches and an overview of attacks against watermarks [15].

Printer stenography is related to our approach. For instance, color laser manufacturers encode the date and time a document was printed with tiny yellow dots on print-outs, which cannot be seen unless the print-out is magnified [24]. Recently, it was reported that printer identification code helped to identify the whistleblower Reality Winner in 2017 [25]. Unlike printer stenography, we encode our hidden information on computer screens.

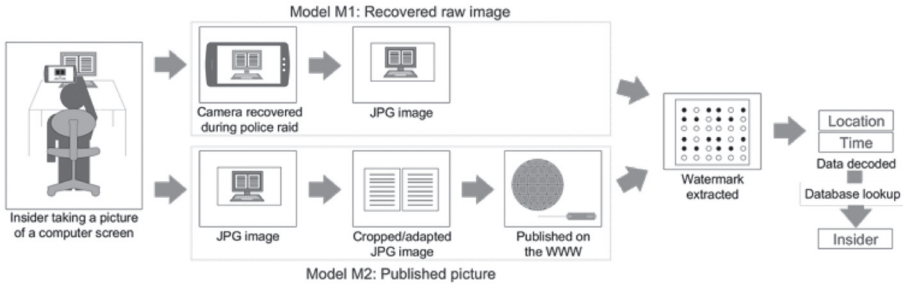
3. PROBLEM STATEMENT AND APPROACH

A. Problem Statement

State-of-the-art security measures cannot prevent insiders from breaching sensitive documents by taking pictures of their computer screens. Taking pictures leaves no log trail that identifies the perpetrator. As a result, it is very difficult to identify the perpetrator based on a recovered leaked picture. We approach this problem with respect to the two scenarios[2] depicted in Figure 1. Both scenarios have in common that: *(i)* an insider (attacker) takes a picture of sensitive information displayed on a screen and *(ii)* a forensic investigator can access the recovered picture and needs to identify the attacker based on the picture. In *scenario M1*, investigators get access to the original, unmodified picture of the camera, e.g., because it was found during a police raid. In *scenario M2*, investigators only see a modified version of the picture as it has been published.

2 Our methodology could also be applied to other scenarios where information needs to be transported in pictures or screenshots.

FIGURE 1. USAGE SCENARIOS: AN INSIDER TAKES A PICTURE OF A COMPUTER SCREEN WHICH IS LATER RECOVERED, EITHER THE ORIGINAL PICTURE (MODEL M1) OR A MODIFIED VERSION OF IT (MODEL M2). THE WATERMARK IS THEN EXTRACTED TO DETERMINE WHEN AND WHERE THE PICTURE WAS TAKEN.



B. Approach

We approach this security threat by embedding hidden watermarks in computer screens. Our watermarks encode information such as the time and the workstation (location). A picture of a watermarked computer screen carries this information. If investigators get access to the (modified) photograph, they can decode this information and identify the perpetrator by verifying who was logged in at the workstation at the time. Watermarking text documents, images, and videos to trace their dissemination is a well-established technique (see §2), but the threat scenario of an insider taking pictures of sensitive data displayed on a screen poses several problems, which make established watermarking techniques unsuitable for this task. In particular:

(a) as the attacker can take a picture at any time, there is no controlled release process and the watermark must be present on any document displayed on the screen at any time;
(b) the watermark must be unnoticeable on text documents, but still be robust against the image artifacts introduced when taking photos of a computer screen; and
(c) the approach should allow for blind extraction, i.e., watermark extraction without the original document.[3]

While traditional text watermarking approaches encode data by modifying individual text passages, we embed information by overlaying a pattern of slightly brighter/darker areas to approach challenge *(a)*. The corresponding overlay mask is independent of the content displayed on the screen and can thus be pre-computed. This makes our watermarking process suitable for real-time embedding. To handle challenge *(b)*, we develop watermarking symbols that are based on the fact that the human eye, especially in light color areas [6], is insensitive to small continuous brightness gradients [2],

3 Alternatively, one would have to record all Desktop interactions resulting in major privacy issues.

while digital cameras capture small changes in brightness well. Further, we modify the design of a traditional convolutional coder and use evolutionary algorithms for deriving optimized generator polynomials in order to handle the high error rates caused by image artifacts. We use redundancy and split our watermarks into sub-watermarks to allow extractions of partly corrupted watermarks. Furthermore, we store cryptographic checksums in our watermarks to allow bit error corrections. To approach challenge *(c)*, we develop an algorithm for blind symbol extraction that is based on the observation that the background color is clearly dominating in typical text documents, allowing us to use local reference brightness values for the symbol decoding.

4. DESIGN

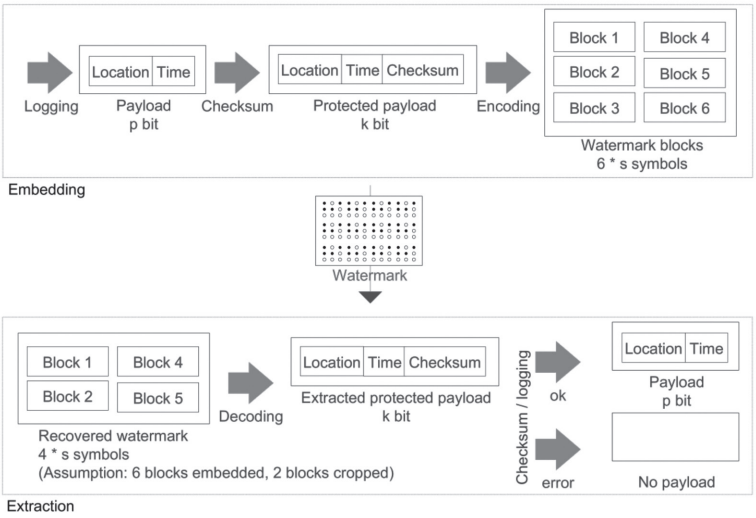
Figure 2 shows the workflow of the proposed watermarking system. The *embedding* process will interact with the graphics card on the watermarked end host (not implemented in the prototype used for this evaluation), while an investigator conducts the *extraction* using standalone software. We assume that a graphic card implementation of the embedding process does not lead to a noticeable increase in CPU usage or power consumption. Even if this assumption does not hold, the user has no baseline for these characteristics that allow them to identify that screen watermarks exist.

Watermark embedding involves the following steps:
(1) Logging: This module creates a bitstring that identifies the end host, user, and a point in time.
(2) Checksum (see §4.C): This module calculates a cryptographic checksum (incorporating a secret user key) for error detection and integrity checking. The checksum block is appended to the payload and the resulting protected payload is provided to the encoder.
(3) Encoding (see §4.B): The protected payload is encoded using an adapted convolutional encoder.
(4) Embedding (see §4.A): Watermark symbols representing the encoded data are generated and placed on the computer's screen.

The extraction of a watermark involves these modules:
(1) Extraction (see §4.A): The watermark symbols are extracted from the recovered picture.
(2) Decoding (see §4.B): The encoded data is decoded using the Viterbi algorithm [21] in order to extract the protected payload.
(3) Checksum/logging (see §4.C): The logging system stores which user was logged in

at the extracted time and location. The corresponding secret user key is retrieved from a database and the cryptographic checksum is verified. If the checksum is correct, the location and time are returned, otherwise the extraction fails with an error.

FIGURE 2. WATERMARKING OF A COMPUTER SCREEN (TOP) AND EXTRACTION FROM A PHOTOGRAPH (BOTTOM). THE PAYLOAD CONSISTS OF *P* BITS. THE CHECKSUM MODULE APPENDS A CHECKSUM TO THE PAYLOAD. THE ENCODING MODULE TRANSFORMS THE PROTECTED PAYLOAD INTO SIX WATERMARK BLOCKS. THIS PROCESS IS REVERSED DURING THE WATERMARK EXTRACTION.



A. Watermarking Symbols for Computer Screens

We introduce watermarking symbols that are a hybrid between traditional text and image watermarking symbols. We operate in the spatial domain, similar to existing text watermarking approaches. This way, the visible artifacts caused by embedding watermarks in a transformed domain are avoided. Still, we avoid the processing-intensive and thus slow text parsing by not changing or moving the text but by overlaying a pattern of slightly brighter and darker areas. Similarly to Caronni [5], we change the brightness of multiple contiguous pixels, which makes our symbols more robust against image artifacts and modifications. However, Caronni's symbol embedding does not allow for blind extraction. To solve this problem, we apply a form of pseudo-differential amplitude modulation. That is, instead of comparing the color values between the watermarked and the original image at the same position in the image, we compare, for each watermark symbol separately, the color within the watermark symbol to the color in the surrounding area. Further, we use circular patterns and soften their shapes by introducing white noise that causes as a smooth gradient

between the watermark center and the surrounding area to avoid sharp contrasts that can become visible on the homogeneous backgrounds in text documents.

The key steps for embedding and extracting watermarks are as follows:

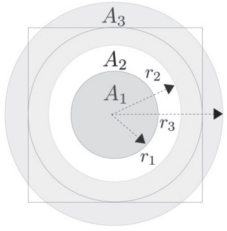
Symbol embedding. Embedding watermark symbols is a two-step process: *(i)* we calculate an overlay mask of slightly brighter/darker areas (symbols) and *(ii)* the watermarking system applies this mask to the screen output. We point out that *(i)* can be pre-computed, thus only *(ii)* is time critical.

Overlay mask. The symbol shape that we use for our approach is shown in Figure 3(a). Every symbol represents one bit. To embed a binary "0", we make the center of the symbol slightly brighter; and to embed a "1", we make the center slightly darker. A watermark consists of a matrix of these symbols. While the brightness of the inner-most circle of the symbol (r_1 in Figure 3(a)) is adapted, a smooth gradient and white noise are applied to the area A_2 to avoid any sharp brightness changes. The watermark decoder compares the background in A_1 to the background in A_3 to tell which binary value the symbol represents. To facilitate the manual extraction of watermarks, the software can further be configured to mark the corners of watermarks using small black markers, which look similar to pixel errors. A photograph of a resulting watermark for an intensity $I_{\max} = 2$ is show in Figure 3(b). Figure 3(c) highlights the watermarking symbols for illustration.

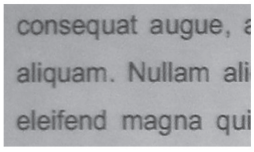
Applying the overlay mask. The application of the overlay mask is quite similar to applying a screen color profile. It requires only local brightness modifications, resulting in a very lightweight embedding process that can be parallelized on a GPU.

Symbol extraction. The extraction of a watermarking symbols from photographs takes place during forensic investigations and is the reverse of the symbol embedding. In contrast to the symbol embedding, this process is not time critical. To extract watermark symbols from a picture, the picture is de-skewed, the watermark symbols are located and the color values of the center of each individual symbol are compared to the surrounding area.

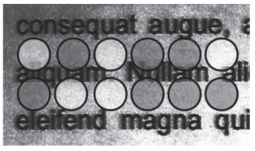
FIGURE 3. A SINGLE WATERMARK SYMBOL (A) AND PICTURES OF A WATERMARKED SCREEN: ORIGINAL (B) AND WITH HIGHLIGHTED SYMBOLS (C).



(a) A single watermark symbol.



(b) Unmodified picture of a watermarked screen.



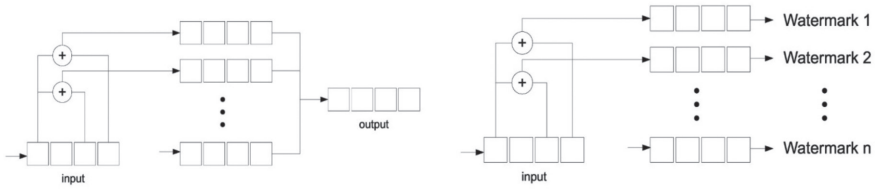
(c) The same picture with watermark symbols highlighted.

B. Encoding of Data in Watermarks

The proposed approach uses error-correcting convolutional codes to achieve a high robustness against incorrectly transmitted symbols. More than one thousand watermark symbols ("physical" bits) fit on a typical screen area of at least 1.024M pixels for our largest symbol size of 32x32 pixels, but we will only need to transport few data bits in a typical setup, therefore we can introduce a high degree of redundancy. Still, this large coding budget is required for a high robustness because error correcting codes for watermarks must be able to operate on short payloads and be robust against various errors [10]. In particular, one has to compensate for cropped images and a very high symbol error rate due to image artifacts. We achieve robustness against cropping by modifying the design of a traditional convolutional encoder and optimizing the error correcting polynomials for short payloads using evolutionary algorithms.

Instead of generating one large watermark, the output of the different generator polynomials used by the encoder is decoupled. This generates multiple smaller, independent sub-watermarks (blocks). Each block carries the complete payload (including a checksum), which allows the decoder to arbitrarily combine the blocks for extracting the payload. That is, on the one hand, one block with few bit errors can already be sufficient to reconstruct the payload. On the other hand, if multiple blocks are available, the decoder can arbitrarily combine these to an optimal combination to compensate for higher bit-error-rate (BER), which leads to very powerful error correcting capabilities. The difference between our and a traditional encoder is illustrated in Figure 4. The traditional encoder merges the output of all generator polynomials to one codeword. In contrast, our design partitions the outputs into smaller sub-watermarks, each with a coding rate of R = 0.5 (termination not included). Combining all sub-watermarks corresponds to the traditional decoder. For decoding the data, we use the common Viterbi algorithm [21].

FIGURE 4. INSTEAD OF MERGING THE OUTPUTS OF THE DIFFERENT GENERATORS (A), WE USE EACH OUTPUT FOR A SEPARATE WATERMARK (B).



(a) Traditional encoder (b) Our encoder

C. Cryptographic Checksum for Error Detection and Integrity Checking

As convolutional codes offer only limited capabilities in detecting errors, a Cyclic Redundancy Check (CRC) [22] is often included in communication protocols to detect decoding errors. We use a cryptographic checksum instead, which additionally allows us to verify the integrity of the extracted message. The checksum block is calculated on the concatenation of the payload and a randomly chosen secret key k_u. Every user u has its own secret key k_u assigned. This protects against accepting a maliciously or accidentally modified message.

5. EVALUATION

We use the following terms throughout the evaluation.

- ***Symbol:*** A symbol is a circular area on the screen that represents one raw bit.
- ***Block:*** As outlined in §4.B, we split a watermark into multiple self-contained blocks. A block is a collection of s symbols.
- ***Symbol size:*** The size of a single symbol (as shown in Figure 3(a)) in pixels.
- ***Watermark intensity:*** the intensity tells how much brighter or darker the symbols are than the surrounding background. We measure the intensity as tuple (Δr, Δg, Δb). The Δ-values are added to or subtracted from the red, green and blue color channel, respectively.

In general, the stronger the watermarks, the more reliable is the watermarking process. But stronger watermarks are also easier to perceive by humans and therefore more disturbing. Thus, we aim to find an operation point at which the watermarks are imperceptible to humans during regular use, but the watermarks can still be reliably extracted from photographs. We evaluate in the following the perceptibility, bit error rates, robustness to image transformation and overall performance of the watermarks.

A. Perceptibility of Embedded Watermarks

1) Setup I

We conduct a user study with 17 adult test subjects working in the defense industry. The aim of the study is to measure and elaborate the visibility of watermarks for different intensities. We embed watermarks of different intensities into a text document; some watermarks are placed in areas with text, while others are placed in a way such that they are not covered by any text. The document is displayed on a Samsung SynchMaster SA450 22 inch screen with a resolution of 1680 x 1050 pixels. The study participants were told that the study was on watermarks, but they did not know what the watermarks looked like. The subjects were asked to read the document. After reading the article, the subjects had to point out which watermarks they could see.

2) Results I

The results of the experiment are presented in Table I. The table distinguishes between watermarks placed on areas where there was no text (background) and watermarks placed in regions with text. All subjects recognized the control watermarks with intensity (20,20,20). But already half of all subjects did not recognize watermarks with intensity (10,10,10) if placed in areas with text. No test subject noted the watermarks of intensity (3,3,3) in text areas. On the other hand, in areas without text, 7 out of 17 subjects spotted watermarks of intensity (3,3,3). The watermarks of intensity (1,1,0) were never identified by any study participant. There is an additional interesting insight not shown in the table. We found that watermarks at the top of the screen were perceived significantly more often than their counterparts at the bottom of the screen. We inspected the screen that was used and found that color contrasts were stronger at the top of the screen than at the bottom.

In summary, we conclude from this study that *(i)* one can use considerably higher intensities for watermarks concealed by text and *(ii)* fine-tuning the intensity of watermarks for different screen regions can be beneficial in order to compensate for the inhomogeneous contrast representation of computer screens.

TABLE I. PERCEPTIBILITY FOR WATERMARKS OF DIFFERENT INTENSITIES. THE PERCEPTION RATE DENOTES THE RATIO OF TEST SUBJECTS IDENTIFYING THE CORRESPONDING WATERMARK.

on white background		in regions with text	
intensity	perception rate	intensity	perception rate
(1,1,0)	0/17	(3,3,3)	0/17
(1,2,1)	5/17	(5,5,5)	4/17
(2,2,2)	5/17	(10,10,10)	9/17
(3,3,3)	7/17	(20,20,20)	17/17

B. Bit error rates

1) Setup II

We measure the bit error rate (BER), i.e., the ratio of symbols that are incorrectly extracted, for different hardware devices. We focus on watermarks that are located in areas with text. We embed watermarks of intensity (2,2,2) in a text document with font size 10pt, display the text document on a Lenovo T430s such that the watermarks cover the whole screen, and we take pictures with different cameras. This laptop features a Twisted Nematic (TN) panel with a resolution of 1600 × 900 pixel. We use a different device for this experiment than for the user study. However, we compared the low contrast characteristics of the panels and found them to be very similar.[4] We measure the BER for three different symbol sizes and four smartphone cameras: Lumia920, SonySk17i, SamsungNexus, and MotorolaXT910. We place two (three for symbol size 20×20) randomly generated watermarks in the document such that they cover the whole screen and take five pictures with each configuration.

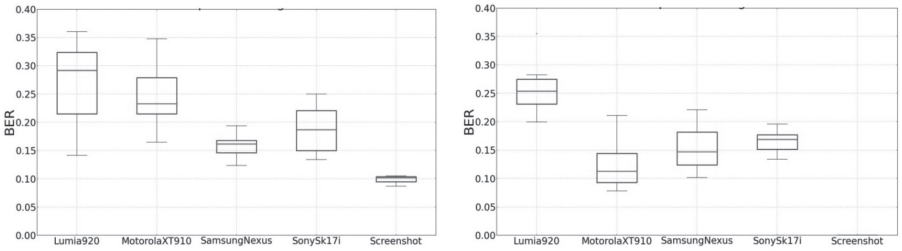
2) Results II

The results for the smallest and largest symbol sizes are shown in Figure 5. Each data point represents the BER for a single watermark. For reference, the right column in each figure shows the BER for symbols directly extracted from a screenshot. We use a screenshot for comparison to measure the influence of the image artifacts caused by taking photographs of the screen. There is a clear trend towards lower BER for larger symbol sizes. This is because pixels representing text are filtered during symbol extraction and more pixels remain after filtering for larger symbols, making the approach more robust. The screenshots also show some bit errors for the two smaller symbol sizes. We confirm this finding by measurements conducted on a watermarked document without any text (not shown in the Figure). For a blank document, the photographs of all symbol sizes achieve a BER of around 0.05 and the screenshots do not exhibit any errors. The user study already showed that contrasts were stronger on the top than on the bottom of the screen. We verified this finding by analyzing the topology of bit errors in Figure 6(a). Indeed, the BER is lower at the top of the screen than at the bottom.

In summary, we conclude that larger symbols are better for watermarks in text areas. For a symbol size of 32×32, we achieve a median BER between 0.12 and 0.25, the maximum BER is 0.28.

4 We tested three different TN panels and one PVA panel. The low contrast characteristics of all these devices were similar.

FIGURE 5. BIT ERROR RATES (BER) FOR WATERMARKS IN TEXT AREAS FOR THE SMALLEST AND LARGEST EVALUATED SYMBOL SIZES AND FOUR CAMERAS, AS WELL AS A REGULAR SCREENSHOT FOR COMPARISON. THE BER FOR A SCREENSHOT IS ZERO FOR 32 × 32 SYMBOLS. FIVE PHOTOS HAVE BEEN TAKEN FOR EACH CONFIGURATION.



(a) Symbol size 20 × 20.

(b) Symbol size 32 × 32.

C. Robustness to Image Transformations

We evaluate in the following the robustness of our approach to image transformations in regard to scaling and color adjustments.

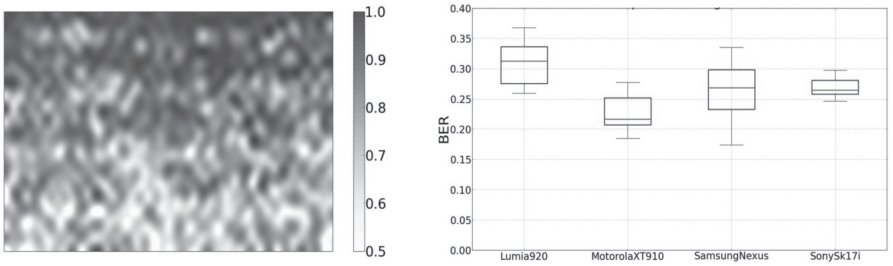
1) Setup III

To simulate a scenario in which images are compressed before being leaked, in this experiment we compress the pictures taken with the mobile phones by a factor of four. This means that the width and height of each image is halved. The resulting pixels are interpolated. A reduction by a factor of four can be considered as a worst-case scenario with respect to image compression for the pictures analyzed in this work, because further decreasing the resolution would make the text in the document very hard to read. Thus, it is unlikely that an attacker would further compress the images.

2) Results III

The resulting BER for a symbol size of 32 × 32 are shown in Figure 6(b). Resizing the images increases the BER by 10 to 15 percentage points compared to their original images, resulting in an average BER of approximately 25%.

FIGURE 6. INHOMOGENEOUS ERROR DISTRIBUTIONS DUE TO DISPLAY CHARACTERISTICS (LEFT) AND BER FOR RESIZED IMAGES (RIGHT).



(a) Ratio of correctly extracted symbols in text areas on a Lenovo T430s.

(b) BER for resized images of watermarked text. Symbol size 32 × 32.

D. Overall Performance

1) Setup V

We calculated the BER for different scenarios in the previous subsections of the evaluation. As the last step of the evaluation, we now relate the BER and the percentage of the available watermarked area to the probability that the transported data can be successfully extracted from a watermark. For this evaluation, we assume that a watermark capacity of $p = 40$ bit is required to encode a user identifier and a timestamp; a payload of $p = 40$ bits results in a protected payload of length $k = 72$ bits and $s = (k + m - 1) * n = 172$ symbols per block (see Figure 2). The parameter $m=15$ represents the length of the used shift register for the convolutional encoder and $n=2$ represents the number of output bits per input bit. We measure the performance of the applied convolutional coding by conducting a Monte Carlo Simulation with 6000 runs. Bit errors are modeled as i.i.d according to the given BER.

2) Results V

The results are shown in Figure 7(b). Every line in this Figure shows the performance of our approach for a different average BER. To give an example, the blue triangle in the upper center of the plot shows that for a BER of 0.25 and 3 recovered watermark blocks, the probability that the data can be successfully extracted from a watermark is around 85%.

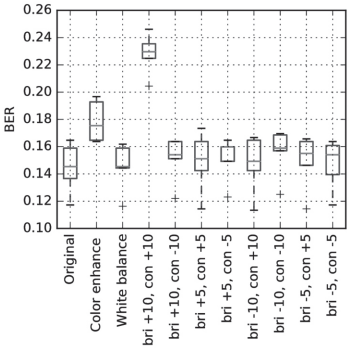
We first focus on pictures without color modifications (raw images). As shown in Figure 5(b), the BER for a symbol size of 32 x 32 is always below 20% for three out of the four mobile devices. Putting this number into Figure 7(b), we see that three out of six watermark blocks are sufficient to decode the data in this case. For the Lumia920, the average BER is 25%, thus we need four to six watermark blocks to

successfully extract the data with high likelihood. The maximum observed BER is 28%. The probability that the payload can be successfully extracted for this case is at least 98%, as Figure 7(b) shows.

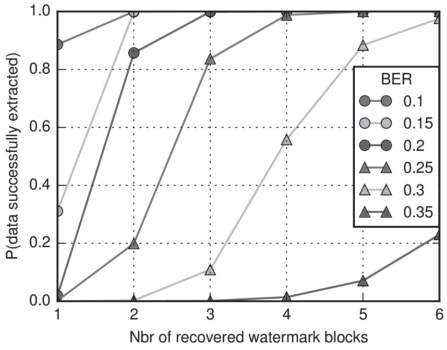
Resizing pictures results in a BER of around 30% (see Figure 6(b)). Figure 7(b) shows that the data can be extracted with a probability of 98% for a BER of 30%. Contrast and brightness changes and automatic color enhancements resulted in a BER below 25% (see Figure 7(a)). Already four out of six watermark blocks are sufficient to reconstruct the embedded data in 99% of cases.

We conclude that we can recover the watermarks from unmodified photographs for all tested smartphones. The Lumia920 introduces a bit error rate of 25%, which reduces the robustness to image modifications, such that 2/3 of the watermark blocks of cropped images are required. For the other three smartphones, we can scale down the image by a factor of four or increase the contrast and brightness by 10% and still extract the encoded data. Watermarked pictures taken with these smartphones are also very robust to cropping of the raw image, only 50% of the watermark blocks are required to extract the watermark.

FIGURE 7. ERROR RATES FOR MODIFIED PICTURES (LEFT) AND OVERALL PERFORMANCE (RIGHT).



(a) The first column shows the BER on the original pictures. The other columns show the resulting BER after applying GIMP's automatic color enhancement, GIMP's white balance function and various brightness (bri) and contrast (con) changes.



(b) Overall performance. The y-axis shows the probability that the data encoded in a watermark can be extracted depending on the number of available watermark blocks and the Bit Error Rate (BER).

6. DISCUSSION

An attacker who is aware of the fact that computer screens are watermarked could try to use our watermarking approach to hold another employee liable for a leaked picture. First, an attacker could attempt to create a fake watermark that contains the identifier of an employee E. However, the attacker also needs to generate the correct cryptographic checksum, which is based on a secret key k_u, otherwise the watermarking system rejects the watermark (see §4.C). An attacker does not know k_u, so they can only guess what the correct checksum is. The odds for guessing the correct checksum is in the order of one in one billion for a 32-bit checksum and six embedded watermark blocks.

Second, an insider could use an unlocked workstation to access the critical information or even access the information with stolen credentials. To detect such a case, one could combine our watermarking approach with biometric techniques that identify the employee currently using a workstation [8].

Third, an insider could take a picture of a document while another employee views the document on their screen. To investigate such and similar cases one would need to complement our approach with CCTV cameras monitoring the office environment. After extracting time and location from a watermark, an investigator could check the surveillance camera recordings of the corresponding office.

Finally, in order to frame an employee *E*, a skilled attacker could take a picture of *E*'s screen, extract the watermark from the picture, and embed it into a picture showing a document that *E* is not supposed to access. The watermark would show where and when the attacker took the picture. This information can be compared against the logs generated by our logging module (see §4), which would show that *E* never accessed the document. Further, CCTV cameras could identify the attacker.

7. CONCLUSION

In conclusion, our proposed watermarking scheme applies imperceptible low-intensity watermarks to the screen. The information embedded with our technique can later be retrieved from photographs or screenshots. We develop a coding scheme based on convolutional codes, which complements the watermarking technique and can cope with the particular challenges of screen watermarking, such as high error rates, inhomogeneous error distributions (caused by the underlying hardware) and partial pictures of screens. We conduct a user study showing that our watermarks are imperceptible during regular use and demonstrate in various experiments that our

watermarks are robust regarding resizing and basic image manipulations. In future work, we will investigate possible attacks against screen watermarks, e.g. by taking advantage of physical screen characteristics, and corresponding protection methods.

ACKNOWLEDGMENT

This work was partially supported by the Zurich Information Security Center. It represents the views of the authors.

REFERENCES

[1.] A. M. Alattar and O. M. Alattar. Watermarking electronic text documents containing justified paragraphs and irregular line spacing. *Proceedings of SPIE* - Volume 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, pages 685-695, Jan 2004.

[2.] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Syst. J.*, 35(3-4):313–336, Sept. 1996.

[3.] A. K. Bhattacharjya and H. Ancin. Data embedding in text for a copier system. In *Proc. Int. Conf. Image Processing (ICIP 99)*, volume 2, pages 245–249. IEEE, 1999.

[4.] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman. Electronic marking and identification techniques to discourage document copying. *IEEE J. on Selected Areas in Comm.*, 13(8):1495–1504, Oct 1995.

[5.] G. Caronni. Assuring ownership rights for digital images. In *Verlaessliche ITSysteme*, DUD-Fachbeitraege, pages 251–263. Vieweg+Teubner Verlag, 1995.

[6.] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2007.

[7.] I. J. Cox, J. Kilian, F. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. on Image Processing*, 6(12):1673–1687, Dec 1997.

[8.] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *NDSS*, 2015.

[9.] Z. Jalil and A. Mirza. A review of digital watermarking techniques for text documents. In *Proc. Int. Conf. on Information and Multimedia Technology (ICIMT)*, pages 230–234, Dec 2009.

[10.] S. Katzenbeisser and F. A. Petitcolas, editors. *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Inc., 2000.

[11.] M. G. Kuhn and R. J. Anderson. *Soft Tempest: Hidden Data Transmission Using Electromagnetic Emanations*, pages 124–142. Springer Berlin Heidelberg, 1998.

[12.] Y. Liu, J. Mant, E. Wong, and S. H. Low. Marking and detection of text documents using transform-domain techniques. *Proceedings of SPIE* - Volume 3657, Electronic Imaging Conference on Security and Watermarking of Multimedia Contents, pages 317-328, 1999.

[13.] N. Nikolaidis and I. Pitas. Robust image watermarking in the spatial domain. *Signal Processing*, 66(3):385 – 403, 1998.

[14.] F. A. Petitcolas. Watermarking schemes evaluation. *Signal Processing Magazine, IEEE*, 17(5):58–64, Sep 2000.

[15.] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn. Attacks on copyright marking systems. In *Proc. Int. Workshop on Information Hiding*, pages 218–238, London, UK, 1998. Springer.

[16.] M. Piec and A. Rauber. Real-time screen watermarking using overlaying layer. In *Proc. Ninth International Conference on Availability, Reliability and Security*, ARES '14, pages 561–570. IEEE Computer Society, 2014.

[17.] Ponemon Institute LLC. 2015 Cost of Data Breach Study: Global Analysis. http://www-03.ibm.com/security/data-breach/, May 2015.

[18.] C.-S. Shieh, H.-C. Huang, F.-H. Wang, and J.-S. Pan. Genetic watermarking based on transform-domain techniques. *Pattern Recognition*, 37(3):555 – 565, 2004.

[19.] T. K. Tsui, X.-P. Zhang, and D. Androutsos. Color image watermarking using multidimensional fourier transforms. *IEEE Trans. on Information Forensics and Security*, 3(1):16–28, March 2008.

[20.] R. Van Schyndel, A. Tirkel, and C. Osborne. A digital watermark. In *IEEE Int. Conf. on Image Processing (ICIP)*, volume 2, pages 86–90 vol.2, Nov 1994.
[21.] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260–269, 1967.
[22.] R. Wang, W. Zhao, and G. Giannakis. Crc-assisted error correction in a convolutionally coded system. *IEEE Trans. on Comm*., 56(11):1807–1815, 2008.
[23.] D. Lohrmann. Two New Insider Threats to Consider. *CSO Online* (2013-06-23), https://www.csoonline.com/article/2137207/infosec-staffing/two-new-insider-threats-to-consider.html.
[24.] S. Schoen. Secret Code in Color Printers Lets Government Track You. *Electronic Frontier Foundation Press Release* (2005-10-16). https://www.eff.org/de/press/archives/2005/10/16.
[25.] C. Baraniuk. Why Printers Add Secret Tracking Dots. *BBC, InDepth, Technology* (2017-06-07). http://www.bbc.com/future/story/20170607-why-printers-add-secret-tracking-dots.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Neural Network and Blockchain Based Technique for Cyber Threat Intelligence and Situational Awareness

Roman Graf
Austrian Institute of Technology GmbH
Vienna, Austria
roman.graf@ait.ac.at

Ross King
Austrian Institute of Technology GmbH
Vienna, Austria
ross.king@ait.ac.at

Abstract: Protecting Critical Infrastructure (CI) against increasing cyber threats has become as crucial as it is complicated. To be effective in identifying and defeating cyber attacks, cyber analysts require novel distributed detection and reaction methodologies based on information security techniques that can automatically analyse incident reports and securely share analysis results between Critical Infrastructure stakeholders. Our goal is to provide solutions in real-time that could replace human input for cyber incident analysis tasks (triage) to classify cyber incident reports, find related reports in a fast and scalable way, eliminate irrelevant information, and automate reporting life-cycle management. Our effective and fast incident management method is based on artificial intelligence and can support cyber analysts in establishing cyber situational awareness, and allow them to quickly adopt suitable countermeasures in the case of an attack. In this paper, we evaluate deep autoencoder neural network supported by Blockchain technology as a system for incident classification and management, and assess its accuracy and performance. This approach should reduce the number of manual operations and save storage space. We used a Blockchain smart contract technique to provide an automated trusted system for incident management workflow that allows automatic acquisition, classification and enrichment of incident data. We demonstrate how the presented techniques can be applied to support incident handling tasks performed by security operation centres.

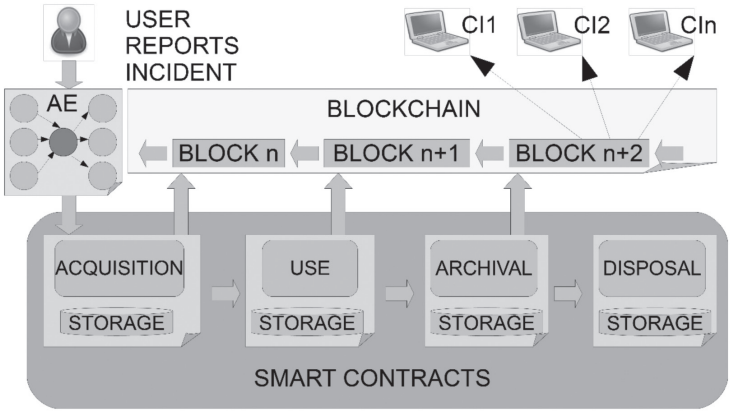
Keywords: *cyber threat intelligence, neural network, blockchain*

1. INTRODUCTION

Cyber Situational Awareness (SA) [1] is a perception of security and threat situations coupled with current and future impact assessment. In recent years, researchers in the SA field have created increasingly complex tools across many application domains. Speed of events, data overload, and meaning underload [2] make real-time SA of cyber operations very difficult to evaluate. Addressing data that is often vague and imprecise, we have to rely on imperfect information to detect real attacks and to prevent an attack from happening through appropriate risk management. Security Operation Centre (SOC) analysts receive a huge amount of daily threat reports. These analysts face challenges finding relevant information in large, complex data sets when exploring data to discover patterns and insights and following organisation business processes, such as proper acquisition, use, archiving and disposal of threat reports. For humans to be effective in identifying and defeating cyber attacks, novel tools that can fill the gap between cyber data and situation comprehension are highly desired. The research presented here is designed to aid in developing a system (see Figure 1) that will automatically support a cyber analyst by analysing and classifying incoming cyber incidents by searching similar high severity cyber incidents that could affect cyber SA, and by life-cycle management of the incident.

Analysis is triggered by a cyber incident report generated by one of the stakeholders in the CI network. The incident analysis can be performed for large amounts of data by using a solid knowledge base (KB), and employing one of the available incident analysis tools. A deep autoencoder (AE) method can be used to analyse existing KB or particular large dataset. The primary purpose of designing a deep autoencoder for SA is to increase the speed of sharing highly severe information and to enable fast and trustworthy cyber incident classification, without the need for substantial human involvement. In our study, we compare existing cyber threat intelligence tools and techniques, describe automatic cyber intelligence analysis approach using a deep autoencoder neural network, and present evaluation results. We leverage expertise collected in available cyber intelligence tools with the power of the neural networks approach.

FIGURE 1. THE OVERVIEW OF ESTABLISHING THE CYBER SITUATIONAL AWARENESS USING NEURAL NETWORKS (AE) AND SMART CONTRACTS FOR INFORMATION CLASSIFICATION AND LIFE-CYCLE MANAGEMENT.



The primary contribution of this work is a real-time solution that could replace human input for a huge number of cyber incident analysis tasks. Another is a methodology, developed to improve information organization and access in cyber security information systems based on automatic classification of cyber security documents according to their expected threat level. We hypothesise that the application of Smart Contracts based on the existing Blockchain technology Ethereum [3] can solve some SA problems. The main purpose of designing Smart Contracts for SA is to enable rapid and trusted cyber incident classification and management, without the need for a large centralised authority. We propose that Smart Contracts based on decentralised assets such as Ethereum can reduce effort for incident life-cycle management and manual analysis costs. Novel techniques that can automatically make predefined decisions obvious by using Smart Contracts can help identify and defeat cyber attacks.

In our context, a Smart Contract basically is a piece of software that fixes and verifies negotiated behaviour and cannot be manipulated because it is distributed and executed on multiple nodes on a Blockchain. Another value of using Smart Contracts is that once deployed, it can function automatically, without the need for human interaction. In our proposed threat intelligence analysis system, we describe the incident handling procedure and instructions using a Smart Contract programming language (Solidity) and upload this Smart Contract to a Blockchain instance (a private Ethereum network). The source code of the Smart Contract defines instructions and rules; for our system, we created 'Acquisition', 'Use', 'Archival' and 'Disposal' Smart Contracts (see Figure 1). The state of the Smart Contracts is stored on the Blockchain and is transparent and accessible to all registered community members. The Smart Contract code is executed in parallel by a network of miners under consensus regarding the outcome of the

execution. The execution of the Smart Contract results in an update of the contract's state (BLOCKn+2) on the Blockchain that is synchronised with every participating user (CI1-CIn) through standard peer-to-peer mechanisms and a Proof-of-Work-based consensus mechanism. An incident report produced by one of the users (security expert protecting CIs) goes through the Smart Contracts and is handled automatically, according to the programmed instructions.

The management system is aimed at the automatic management of threat reports provided by threat analysis tools such as CAESAIR,[1] IntelMQ[2], or MISP[3] and should provide effective decision support for a SOC operator. Compared to manual classification, automatic classification by threat level can significantly support and accelerate reaction time of an SOC analyst. For example, the Collaborative Analysis Engine for the Situational Awareness and Incident Response (CAESAIR) tool [4] supports various security information correlation techniques and provides customizable import capabilities from a multitude of security-relevant sources. These sources include a custom repository, open source intelligence (OSINT) feeds and IT-security bulletins, as well as a standardised vulnerability library (Common Vulnerabilities and Exposures – CVE). CVEs are especially important for Smart Contracts with regard to likelihood assessments based on game theory [5] that implements risk scoring [6]. Employing CAESAIR with CVE scoring [7] and extending it by automated tagging can provide valuable input for information classification and life-cycle management. Such a system can be implemented using Smart Contracts created for a particular organization. Each institution may have multiple classification profile definitions dependent on the network, CI and the role of the cyber analyst.

This paper is structured as follows. Section 2 gives an overview of related work and concepts. Section 3 explains the cyber incident classification workflow. The cyber incident life cycle issues are covered in Section 4, Section 5 presents the experimental setup, applied methods end evaluation and Section 6 concludes the paper.

2. RELATED WORK

Threat intelligence in the cyber security (CS) realm is provided by a number of cyber incident analysis tools. For example, the CAESAIR tool provides analytical support for security experts carrying out cyber incident handling tasks on national and international levels, and facilitates the identification of implicit relations between available pieces of information. IntelMQ is an open source tool collaboratively developed by Austrian CERT and other parties aiming at parsing and correlating cyber incidents. MISP, the Malware Information Sharing Platform is another open source tool that performs automatic data correlation by finding relationships between

1 http://caesair.ait.ac.at
2 https://github.com/certtools/intelmq
3 https://github.com/MISP/MISP

attributes and indicators from malware, attack campaigns, or analysis. It incorporates an indicator database to store technical and non-technical information about malware samples, incidents, attackers and intelligence; and a sharing functionality to facilitate data exchange using different models of distribution.

The autoencoder approach is widely used for different analytical tasks. A machine learning framework based on recursive autoencoders [8] can be used for sentence-level prediction of sentiment label distributions. A very deep autoencoder [9] is employed for content-based image retrieval. In our approach, we are using this method for similarity searches. The advantage of the autoencoder method is that it learns automatically from examples. The autoencoder makes use of neural networks which are already in use by latent semantic analysis for text categorization [10] to reduce dimensionality and to improve performance. Another application [11] employs an artificial neural network to improve text classifier scalability. Classification methods implemented in the previously mentioned threat intelligence tools suffer from large vector sizes and are less effective as the number of incidents rise. The main drawback of existing text classification methods, such as SVM [12], Word Embeddings Neural Networks or the Gensim tool is that they require a huge database for training to provide meaningful results, but expected SOCs datasets are not large enough for such semantic-based tasks. Another common disadvantage of these techniques is the lack of results transparency due to employing vectors containing real-valued numbers. These tools provide results, but it is difficult to explain how the results were calculated. In particular, the SVM approach is limited by the choice of the kernel. Another disadvantage is the inability to handle unknown words or words which were not included previously in the training vocabulary. Consequently, for the particular use case of threat incident classification task for SOCs, we suggest using the autoencoder solution that scales well because of the small vector size while maintaining a high level of accuracy.

Multiple researchers are developing an automated technology that will support an information classification system. An attempt to classify the relationships between documents and concepts [13] employs principles of ontology. To improve information organization and access in construction management, a methodology [14] was developed based on automatic hierarchical classification of construction project documents according to project components. A survey of various cyber attacks and their classification [15] attempted to develop an ontology for cyber security incidents. They classify by characteristics, and by purpose and motivations. Additionally, cyber attacks can be classified based on the severity of involvement, scope, or network types with multiple sub classification terms. Contrary to this approach, we classify only by threat level that can differ from organisation to organisation. Our goal is to focus human expert resources on the most urgent incidents important for a particular organisation. An information life-cycle model described in [16] is also applicable to

the CS domain. Cyber incident reports are acquired, analysed and become outdated. Effective automatic classification, retention and disposal policies can mitigate risks to data and make information management more effective. Classification of data enables a company or SOC to focus their resources toward the most valuable or urgent incidents and to handle less valuable incidents, automatically saving time and other costs.

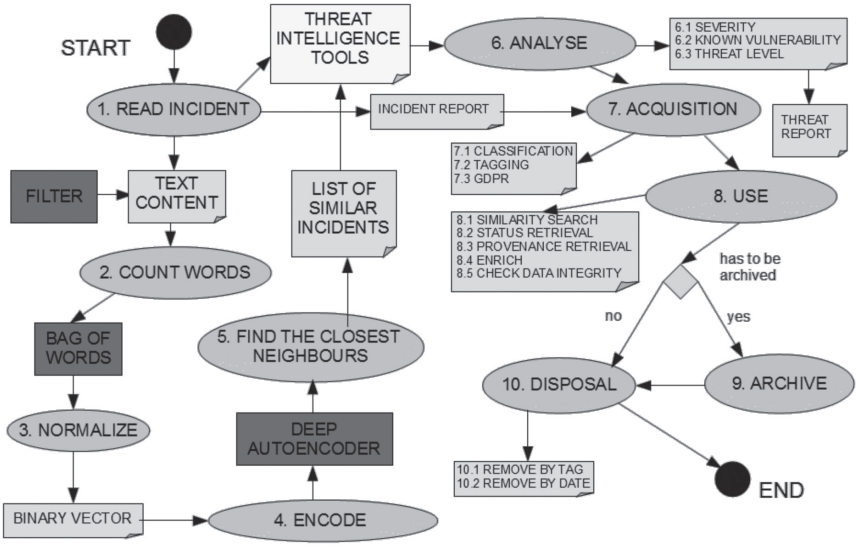
Because members of a CI network do not necessarily trust each other, do not have a central authority and have a need to store and share the life-cycle state of the incident, we suggest a Blockchain-based solution for life-cycle management. An overview of the Blockchain technology and its potential to facilitate money transactions, Smart Contracts design, automated banking ledgers and digital assets is provided in [17]. A Blockchain platform comparison [18] discusses five general-use Blockchain platforms and looks at how Blockchain technology can be used in applications outside of Bitcoin to build custom applications on top of it. This comparison suggests that Ethereum is currently the most suitable and well-established platform. Therefore, for cyber incident analysis we employ an Ethereum Blockchain (specifically, the Pyethereum implementation), which supports a focused Smart Contracts testing environment without the need of mining. In the proposed system, we intend to apply Smart Contracts for cyber incident classification and life-cycle management, which is unique for the given domain.

3. CYBER INCIDENT CLASSIFICATION USING AUTOENCODER

For our study, we assume that a cyber expert is responsible for a CI and detects suspicious behaviour in the system. The expert needs more information to select the correct mitigation strategy. She must collect and analyse all the available information related to ongoing and previous attacks for the particular use case, and transform it into actionable intelligence. Security information such as incident reports, vulnerability alerts, advisories, bulletins etc., usually come in the form of semi-structured text documents. Acquiring cyber threat intelligence from such documents requires manually reviewing and discerning what significant information they can find, and identifying implicit correlations among them in order to estimate their impact and outline possible mitigation strategies. To avoid this manual effort, the CIs expert can provide an incident report as an input to a deep autoencoder and receive a threat report back if it has sufficient severity. An automatic approach delivers a significant improvement in terms of personnel costs when compared to manual cyber incident handling. As a result, an analyst has the up-to-date SA status and we ensure fast and scalable information exchange and enrichment.

The idea behind applying the autoencoder approach is that we can map N-dimensional data onto the M orthogonal directions in which the data have the most variance and form a lower dimensional subspace. The acceptable drawback of this conversion is that in the remaining orthogonal directions we lose information about the original data point location.

FIGURE 2. THE WORKFLOW FOR CLASSIFICATION AND LIFE-CYCLE MANAGEMENT OF CYBER INCIDENT USING AUTOENCODER AND SMART CONTRACTS.



We employ a deep autoencoder that was trained as described in the workflow shown in Figure 2. The workflow execution starts with reading the incident report (1) and parsing the report content. Input data along with the expert profile settings, which are specific to the organisation, are converted to a binary vector using the 'bag of words' technique (2) and after the normalization step (3) passed to the autoencoder in encoded form (4). In this step, we compile the words most used in documents. The remaining vector is comprised of word counts irrespective of order. For simplicity, we use a binary count where we mark 1 if a word count is bigger than 0, and 0 if the given word is not present in an original document. Additionally, we ignore stop words (words with no discriminatory power, such as common articles and prepositions, that we do not need in analysis). To achieve reasonable performance and scalability, we reduce each vector to a much smaller vector that still comprises enough information about the content of the document. In the next step, we train the neural network to reproduce its input vector as its output. This forces it to compress as much information as possible into the 10 numbers in the central bottleneck. These 10 numbers are then a result of deep autoencoder training and a good way to compare documents (5) in a fast

and scalable way using the cosine similarity method. In the next step, we merge the detected related incidents with institutional settings and decide which priority level (see Equation 1) should be applied to the given incident. The compressed vectors are stored on the hidden level of neural network (see Table 1).

$$P = f(I_r, W_r, W_o, T_s, W_s) \qquad (1)$$

Equation 1 shows the incident priority level P that returns the value – either 0 that corresponds to 'Low' or 1 representing 'High'. Priority level is a function of aggregated incident evaluation metrics, which depend on basis indicators, such as 'number of related incidents' I_r, 'number of related words' W_r, 'number of original words' W_o, 'detected significant terms' T_s and 'vulnerability score' V_s.

4. CYBER INCIDENT MANAGEMENT USING SMART CONTRACTS

We evaluate the application of Smart Contracts to classify and manage incident reports labelled by the autoencoder as a high priority threat. Smart Contracts can be used to estimate that the reported cyber incident is of high relevance, to remove it after some predefined time, to tag it by acquisition, to search by tag, to assign access rights (confidential, private, sensitive, public), to periodically check data integrity (preventing manual or hardware corruption), or to determine data provenance. Our goal is to save storage space, improve performance and to keep information up-to-date in a trustworthy way by leveraging the distributed nature of Blockchain technology. Once a Smart Contract is triggered, the analysis result is automatically propagated among all participants through inherent Blockchain mechanisms. One of the advantages of this approach is that Smart Contracts cannot be changed or compromised without being detected (through hashed transactions) and that the messages can be verified to originate from a trusted source (through public key encryption). After incident acquisition, a Smart Contract performs the classification of a report by threat level, stores the obtained threat level on a Blockchain and initiates the life-cycle management process for the given incident. In the next step, this report will be used, archived and disposed.

We employ four Smart Contracts for cyber incident processing, as depicted in Figure 2. The workflow execution after the classification steps performed by the autoencoder proceeds with the analysis of an incident report by reading and parsing the report content enriched with the classification results (6). Input data, along with organization-specific expert profile settings, are passed to the first Smart Contract 'acquisition' (7), which employs one of the threat intelligence tools. Classification occurs by employing

incident text, split by words or phrases, specific terms separated by low, middle and high threat relevance. We compute risk points, counting how many of terms are included in the incident report for each threat level. For threat level calculation, we either estimate threat level by applying thresholds for each level or we employ the weighted method from Formula 2, where we additionally multiply the calculated points on each threat level with a constant which represents the weight of the related threat level. The threat level scale ranges from 1 to 3, where 1 is 'low threat' and 3 is 'high threat'. Risk points RP is a sum of high risk points H_{rp} multiplied by high threat weight HT_w, middle risk points M_{rp} multiplied by middle threat weight MT_w and low risk points L_{rp} multiplied by low threat weight LT_w.

$$RP = H_{rp} * HT_w + M_{rp} * MT_w + L_{rp} * LT_w \quad (2) \qquad T_l = \begin{cases} 3(high) & if\ RP > HT_t, \\ 2(middle) & if\ RP \geq MT_t, \\ 3(low) & else\ RP < MT_t, \end{cases} \quad (3)$$

Where HT_w=3, MT_w=2, LT_w=1 and HT_t=10, MT_t=3. Threat level T_1 can be inferred using high threat HT_t and middle threat MT_t thresholds and weighted risk points RP from Formulas 2 and 3. The acquisition step (7) is split into different tasks. Automatic classification by threat level defines one of three threat levels: 'high' level requires fast reaction and mediation steps, triage process; 'medium' level assumes detection of 'Indicator of Corruption' (IoC) or metrics that indicate possible vulnerabilities, and requires SW update; and 'low' level addresses regular cyber security information and logs, and requires attention but should not necessary be a threat. Tagging means that specific tags can be assigned to a report to make it easier to find, shift or remove later. Removing personal information from the incident report to protect personal data may be required (by the European GDPR) before storing a normalised version of the incident. In the 'using' step (8), the workflow supports an automated similarity search, status and provenance retrieval, and enrichment with data and metadata periodic check for data integrity (using the hash of the incident report). Finally, depending on the threat level after some period of time, the incident can be archived (step 9) or removed e.g. by date or by tag (step 10).

We believe that this automatic smart-contracts-based approach would substantially support incident classification and management and could be used by analysts for the defence of CI. The suggested method would make SA analysis less cost-intensive and would perform with higher throughput. However, as is typical in this area, a human-based approach performs with higher accuracy.

5. EXPERIMENTAL EVALUATION

In the evaluation section, we measure how accurate our automated computations are and how long it takes for the deep autoencoder to make its calculations. Additionally, we report on measurements of the automated cyber incident classification and how long it takes for Smart Contracts to be executed and validated. We carried out measurements for several incident reports. The goal of this evaluation was to leverage the domain expert knowledge base for cyber incident classification and management as described in the workflow (see Figure 2), pointing out threat level relevant for SA.

A. Evaluation Data Set

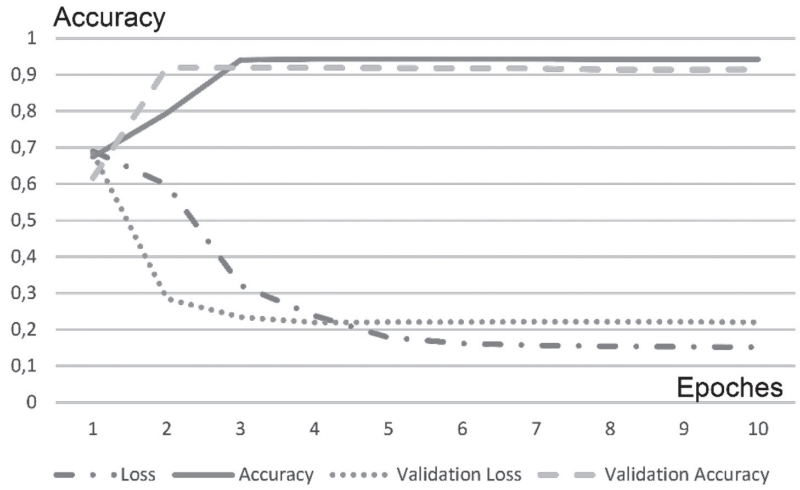
The cyber analyst's goal is to prioritise a detected cyber incident, either to mitigate it or to perform some other cyber incident response. For this test, we assumed that our CI is a financial organisation that employs MS Office products on Windows OS and using software products such as Internet Explorer, Firefox, Adobe, etc. The dataset used was aggregated from OSINT sources on the Internet. The dataset contained 5,850 training documents and 584 test documents. We evaluated cyber incident reports from the 'seclists' feed[4] from the last three years addressing four report categories. The 'fulldisclosure' category contained messages from the public, a vendor-neutral forum for detailed discussion of vulnerabilities and exploitation techniques, as well as tools, papers, news, and events of interest to the community. The 'bugtraq' category is a general security mailing list. The 'pen-test' category discloses techniques and strategies that would be useful to anyone with a practical interest in security and network auditing. The 'nmap-dev' category comprises an unmoderated technical development forum for debating ideas, patches, and suggestions regarding proposed changes to Nmap[5] and related projects. The specific cyber security terms were obtained from the CS glossary.[6] We anticipated that employing the described autoencoder and Smart Contracts approach should classify cyber incidents among a very large number of incident reports facilitating further cyber analysis and incident management.

B. Experimental Results and Interpretation

This evaluation took place on an Intel Core i7-3520M 2.66GHz computer using Python on Ubuntu OS. We performed a total of 10 training iterations (epochs) for the autoencoder. The autoencoder training and accuracy calculation process took about 262 seconds (see Figure 3). This figure shows that loss and validation loss decreased and accuracy and validation accuracy increased with each epoch. A final accuracy of 0.942 was achieved; this demonstrates how well input is reconstructed compared to the output.

4 http://seclists.org/
5 https://nmap.org/
6 https://scottschober.com/glossary-of-cybersecurity-terms/

FIGURE 3. ACCURACY AND LOSS CHARACTERISTICS BY AUTOENCODER TRAINING.



The neural network used a total of 502,000 parameters during the autoencoder training. The summary of the neural network training is presented in the Table 1. The neural network is composed of 1 input layer and 5 hidden layers. The number of neurons in these layers range from 10 to 2,000. Most layers use a rectified linear unit (ReLU) as an activation function. The last decoding layer employs a sigmoid activation function.

TABLE 1. SUMMARY OF THE DEEP AUTOENCODER TRAINING PROCESS.

Layer	Type	Activation Function	Neurons #	Parameters #
Input layer	InputLayer	ReLU	2,000	0
Hidden layer 1	Dense	ReLU	2,000	4,002,000
Hidden layer 2	Dense	ReLU	250	500,250
Hidden layer 3	Dense	ReLU	10	2,510
Hidden layer 4	Dense	ReLU	250	2,750
Hidden layer 5	Dense	Sigmoid	2,000	502,000

The autoencoder model simply maps an input to its reconstruction. To achieve this, we first train an autoencoder until it reaches the stable train/validation loss value. The deep autoencoder system starts the SA analysis with incident content retrieval, which is converted to an input vector by using word counts. This input vector then goes through encoding in multiple hidden layers and is reconstructed to an output layer after decoding in the final layers. Having trained the model, we were able to retrieve

the middle layer of the autoencoder model with the smallest number of neurons (10). Therefore, we retrieved trained 10-number-long IDs for each of the 584 test vectors and iterated this over all of the document vectors (10-numbers-long each) calculating a cosine similarity value for each document. For instance, the trained vector of the query incident report 'bugtraq-2017-Aug-1.txt' containing 10 numbers is [-8.73114914e-10, 1.01575899e+01, 2.12457962e-09, 1.29858088e+00, 2.67755240e-09, 9.32977295e+00, 4.54857439e-01, -5.82076609e-11, 8.55403137e+00, 5.52972779e-09]. This vector can be used for fast and scalable similarity search. Computation demonstrated that, for the given incident report, the first three most similar documents are: 'nmap-dev-2017-q2-8.txt, fulldisclosure-2017-Jan-68.txt, fulldisclosure-2015-Oct-71.txt'. During the correlation calculation using the deep autoencoder, there was a minor fluctuation of accuracy value in the last epochs (between 0.942 and 0.943). This is because the autoencoder employs a restricted Boltzmann machine (RBM), which treats the word counts as probabilities and makes use of random values in calculations. Therefore, it is possible that the highest level of accuracy can be achieved before all of the epochs are calculated (epoch 4 in our case).

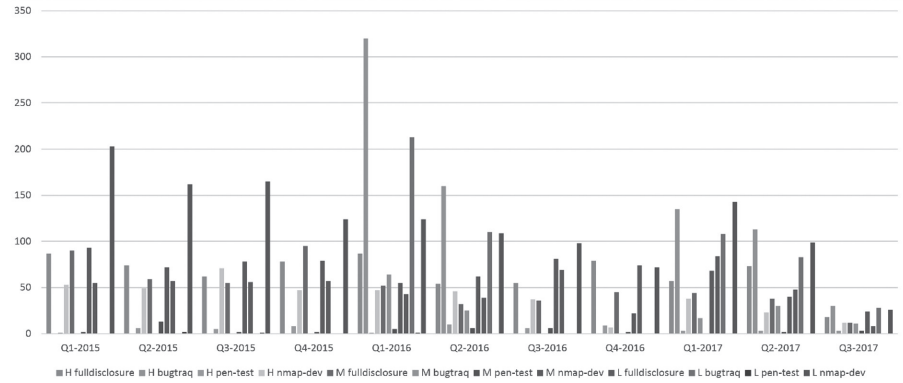
TABLE 2. EXCERPT OF CLASSIFICATION RESULTS FOR CYBER INCIDENT REPORTS BY THEIR ACQUISITION USING SMART CONTRACTS.

Incident ID	Related Incident ID	Similarity	Source	Block-chain ID	Time (sec)	Terms #	Threat Level
Fulldisclosure-2017-Jan-q1-75	fulldisclosure-2016-May-33.txt	108	Wolfgang feedyourhead at	68	0.371	5	3
Fulldisclosure-2015-Feb-q1-53	nmap-dev-2017-q2-8.txt	54	Scott Arciszewski	1,304	0.370	3	2
Fulldisclosure-2015-Feb-q1-90	fulldisclosure-2016-Aug-118.txt	83	Praveen D	1,314	0.451	1	1
Bugtraq-2017-Jan-q1-18	bugtraq-2016-Jan-146.txt	125	Vulnerability Lab	3,419	0.677	13	3
Bugtraq-2017-Jun-q2-56	fulldisclosure-2015-May-52.txt	130	SEC Consult Vulnerability Lab	3,829	0.532	13	3
Bugtraq-2016-Jan-q1-75	nmap-dev-2015-q2-40.txt	79	Slackware Security Team	4,009	0.332	2	1
Bugtraq-2017-Apr-q2-158	nmap-dev-2017-q2-8.txt	54	Salvatore Bonaccorso	4,215	0.432	1	1
Nmap-dev-2017-Mar-q1-226	bugtraq-2016-Apr-36.txt	65	Henri Doreau	4,831	0.533	3	2
Nmap-dev-2015-Nov-q4-107	fulldisclosure-2016-May-33.txt	108	Peter Houppermans	6,849	0.600	8	3
Nmap-dev-2015-Oct-q4-63	nmap-dev-2015-q4-276.txt	68	Mark Scrano	6,853	0.496	1	1
Pen-test-2017-Jul-q3-1	bugtraq-2017-Jul-8.txt	63	Hafez Kamal	7,994	0.252	3	2
Pen-test-2016-Feb-q1-2	nmap-dev-2017-q2-8.txt	54	Francisco Amato	8,071	0.357	2	1
Pen-test-2016-Dec-q4-0	bugtraq-2017-Mar-39.txt	115	ERPScan inc	8,072	0.521	9	3

In the test scenario, we investigated incident reports from 'seclists' CS feed to classify those by threat level and to automatically manage them from acquisition to disposal without involvement of human analyst (see Table II). Due to the large number of results in this table, we describe only selected classification results, which

demonstrate typical cases. Query incident ID in 'seclists' terms is presented in the first column. The second column shows the first of the detected related incident IDs. The similarity score for found related incidents for selected examples is nearly 1.0. In the third column, we show a number of detected common words between query and found incidents. Column 'Source' depicts an incident source that can be a person or an organisation. The next four columns are related to Smart Contracts and show assigned Blockchain ID, consumed time, number of significant terms and threat level. The experimental results are represented in Figure 4 and show the distribution of threat incident reports over the last three years, respective of high, middle, and low threat levels. Each incident category is flagged by an assigned colour. The Y axis is a range of the number of incidents and the X axis is a time scale split into quarters. The figure shows that the most productive category for high (up to 325) and low (up to 215) threats is a 'bugtraq' category, whereas 'nmap-dev' (93) and 'fulldisclosure' (97) are dominating middle threat reports. For a given period of time, most active phase for all levels is from 'Q4-2015' to 'Q3-2016'. Visualization of incident reports provides an analyst with a quick and descriptive SA picture. To focus on a particular area, the analyst can perform fine tuning, adjust the time scale or select a particular category or source.

FIGURE 4. PLOT FOR DISTRIBUTION OF THREAT INCIDENT REPORTS OVER LAST THREE YEARS FOR DIFFERENT THREAT LEVELS SHARED QUARTERLY.



As a use case scenario, assume that SOC has received an incident report from Vulnerability Lab in January 2017. On receiving this report, our Smart Contract triggers automatic analysis and classification of this incident report. According to Table 2, we see that this incident is assigned a Smart Contract identifier 3419 and the contract identifies 13 significant terms. Going through the contract logic we estimate both the regular and the weighted threat level as a 'high threat' (3). That means it should be handled soonest and with highest priority. The incident is automatically tagged and enriched with additional data from CS feeds and tools. Links to similar

incidents are established. All this facilitates the triage process for a cyber analyst and performs analysis steps that are usually done manually. According to the evaluated classification level, Smart Contract defines timestamps for automated archival and disposal of incident data. Therefore, a cyber analyst does not need to worry about the incident life-cycle and can focus their resources on triage for urgent cases.

The smallest duration for one Smart Contract operation was 0.252 seconds from Blockchain ID 7994 report and the longest operation time 0.677 report with ID 3419. This difference can be explained by the different report sizes (we calculate hash for report content) and different risk points numbers (3 for ID 7994 vs. 13 for ID 3419). This evaluation also gives a simple overview of detected significant terms, such as 'attack', 'hack', 'phishing' for high threat incidents, 'access', 'authentication', and 'encode' for middle threat incidents and 'key', 'capability', and 'investigation' for low level threats. Having a Smart Contract ID, the analyst is able to retrieve status data of a particular incident report from Blockchain using Smart Contract (e.g. by hash, provenance, time, tags, owner).

TABLE 3. OVERVIEW ABOUT AGGREGATED THREAT REPORTS FOR DIFFERENT THREAT CATEGORIES.

Threat Category	High Threat	Middle Threat	Low Threat	Total
Fulldisclosure	724	558	590	1,872
Bugtraq	758	147	542	1,447
Pen-test	55	43	4	102
Nmap-dev	430	674	1,325	2,429
Sum	1,967	1,422	2,461	5,850

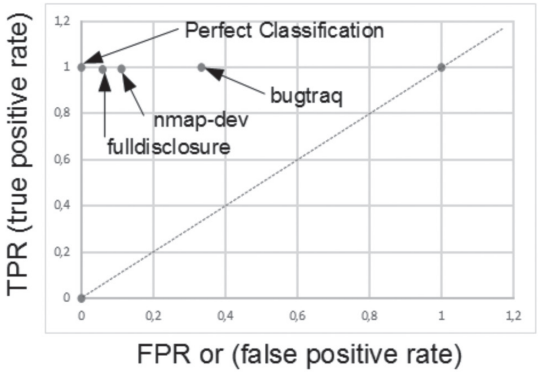
The category overview experimental results are presented in Table 3 which shows the distribution of high, middle and low threat level incidents for different incident categories. This table demonstrates that most incident reports (2,429) come from the 'nmap-dev' category, followed by 'fulldisclosure' (1,872) and 'bugtraq' (1,447). Most of incident reports belong to the low threat level (2,461) but the report number classified as high threat is also high (1,967). Most high threat level reports come from the 'fulldisclosure' (724) and 'bugtraq' (758) categories. That means that these categories should be addressed first by incident management.

C. Evaluation Effectiveness

We can see that, in general, the autoencoder training accuracy improves with every iteration (epoch) from 0.674 at the beginning to 0.942 at the end, which is sufficiently good; whereas training loss (error) of original information decreases from 0.691 to 0.152. This means that the decompressed outputs will be degraded compared to the

original inputs, but it is an acceptable rate. Similarly, validation accuracy is in the range between 0.616 and 0.915. Validation loss decreases from 0.684 to 0.220.

FIGURE 5. ROC SPACE PLOT.



The classification effectiveness for high priority incidents can be determined in terms of a Relative Operating Characteristic (ROC) using the labelled ground truth query dataset. SA analysis divided the provided incident reports into two groups: 'high' and 'low priority' by associated expert parameters and thresholds for each category; e.g. for the 'fulldisclosure' category the provided algorithm detected 229 true positive incidents, 14 true negative reports, one false positive incident and two false negative documents. The primary statistical performance metrics for ROC evaluation are sensitivity (0.991) or true positive rate and false positive rate (0.059). The associated ROC value is represented by the point (0.059, 0.991). The ROC space (see Figure 5) demonstrates that the calculated FPR and TPR values for the evaluated categories are located very close to the so called perfect classification point (0, 1). The calculation results demonstrate that the calculated similarity score values for the query documents are located very close to the labelled classification. These results demonstrate that an automatic approach for cyber incident classification of the method described is very effective and is a significant improvement on manual analysis. Therefore, an analysis method based on deep autoencoder techniques can be suggested as an effective method for incident classification, and as a supporting method to establish cyber SA. The results of the analysis confirm our hypothesis that an automated approach is able to reliably classify incidents, thus making analysis of a large number of cyber incidents a feasible and affordable process. However, further research is required to improve the decision and accuracy metrics of this method.

6. CONCLUSIONS

In this work, we have presented an automated approach to classify and manage incident reports for establishing cyber situational awareness using a deep autoencoder neural network for classification and a Smart Contracts technique provided by Blockchain technology for incident management. The developed system should assist cyber analysts by protecting Critical Infrastructures against increasing cyber threats. The main contribution of this work is a real-time solution that could replace human input for a large number of cyber incident analysis tasks in order to facilitate cyber incident classification, eliminate irrelevant information and focus on important information to promptly perform mitigation steps. Another contribution is the use of the Smart Contract techniques to provide an automated trusted system for an incident management life-cycle that allows automatic acquisition, classification, use, archiving, and disposal. An additional advantage of this approach is a reduction of human analysis costs. Ultimately, our research will lead to the creation of automated security assessment tools with more effective handling of cyber incidents.

REFERENCES

[1] P. Barford et al., 'Cyber SA: Situational Awareness for Cyber Defense,' in *Cyber Situational Awareness, Advances in Information Security*, vol 46, Springer, Boston, MA, 2010.

[2] A. Kott and C. Wang, *Cyber Defense and Situational Awareness*, Switzerland: Springer Int. Publ., volume 62, ISBN 978-3-319-11391-3, 2014.

[3] G. Wood, 'Ethereum: A Secure Decentralised Generalised Transaction Ledger,' in *EIP-150 REVISION*, http://gavwood.com/paper.pdf, 2014.

[4] G. Settanni, F. Skopik, R. Graf, M. Wurzenberger, and R. Fiedler, 'Correlating cyber incident information to establish situational awareness in Critical Infrastructures,' in *14th Annual Conference on Privacy, Security and Trust (PST)*, Auchland, New Zealand, pp. 78-81, 2016.

[5] L. Samarji, 'Coordination and Concurrency Aware Likelihood Assessment of Simultaneous Attacks,' in *Third International Conference on Security and Privacy in Communication Networks SecureComm*, vol. 152, pp 524-529, 2015.

[6] T. Reguly, 'Does Anybody Really Care About Vulnerability Scoring?,' in *International Conference on Computational Science and Engineering*, 2013.

[7] L. Maghrabi, E. Pfluegel, L. Al-Fagih, R. Graf, G. Settanni, and F. Skopik, 'Improved software vulnerability patching techniques using CVSS and game theory,' in *International Conference on Cyber Security And Protection Of Digital Services (Cyber Security)*, pp. 494-505, London, 2017.

[8] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, 'Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions,' in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, pp. 151-161, Edinburgh, Scotland, UK, 2011.

[9] A. Krizhevsky and G. E. Hinton, 'Using very deep autoencoders for content-based image retrieval.,' in Proceedings *ESANN*, Bruges, Belgium, 2011.

[10] Y. Bo, X. Zong-ben, and L. Cheng-hua, 'Latent semantic analysis for text categorization using neural network,' in *Knowledge-Based Systems*, volume 21, number 8, pp. 900-904, 2008.

[11] S. L. Y. Lam and D. L. Lee, 'Feature reduction for neural network based text categorization,' in *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*, pp. 195-202, Hsinchu, 1999.

[12] L. Auria, 'Support Vector Machines (SVM) as a Technique for Solvency Analysis,' in *DIW Berlin*, Paper 811, 2008.

[13] S.-S. Weng, 'Ontology construction for information classification,' in *Exp. Systems with Applications*, volume 31, number 1, pp. 1-12, 2006.
[14] C. H. Caldas and L. Soibelman, 'Automating hierarchical document classification for construction management information systems,' in *Automation in Construction*, volume 12, number 4, pp. 395-406, 2003.
[15] M. Uma and G. Padmavath, 'A Survey on Various Cyber Attacks and Their Classification,' in *International Journal of Network Security*, Coimbatore, volume 15, pp. 390-396, 2013.
[16] S. Harris and F. Maymi, *CISSP All-in-One Exam Guide*, book, New York: McGraw-Hill Education, 2016.
[17] G. W. Peters, *Understanding Modern Banking Ledgers Through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money*, Springer Int. Publishing, pp. 239-278, 2016.
[18] M. Macdonald, L. Liu-Thorrold, and R. Julien, 'The Blockchain: A Comparison of Platforms and Their Uses Beyond Bitcoin,' in *COMS4507 - Adv. Computer and Network Security*, Univ. of Queensland, 2017.

2018 10th International Conference on Cyber Conflict
CyCon X: Maximising Effects
T. Minárik, R. Jakschis, L. Lindström (Eds.)
2018 © NATO CCD COE Publications, Tallinn

Mission-Focused Cyber Situational Understanding via Graph Analytics

Steven Noel
Cyber Solutions Technical Center
The MITRE Corporation
McLean, Virginia, United States

Paul D. Rowe
Cyber Solutions Technical Center
The MITRE Corporation
McLean, Virginia, United States

Stephen Purdy
Software Engineering Technical Center
The MITRE Corporation
McLean, Virginia, United States

Michael Limiero
Cyber Solutions Technical Center
The MITRE Corporation
McLean, Virginia, United States

Travis Lu
Software Engineering Technical Center
The MITRE Corporation
McLean, Virginia, United States

Will Mathews
Cyber Solutions Technical Center
The MITRE Corporation
McLean, Virginia, United States

Abstract: This paper describes CyGraph, a prototype tool for improving network security posture, maintaining situational understanding in the face of cyberattacks, and focusing on protection of mission-critical assets. CyGraph captures complex relationships among entities in the cyber security domain, along with how mission elements depend on cyberspace assets. Pattern-matching queries traverse the graph of interrelations according to user-specified constraints, yielding focused clusters of high-risk activity from the swarm of complex interrelationships. Analytic queries are expressed in CyGraph Query Language (CyQL), a domain-specific language for expressing graph patterns of interest, which CyGraph translates to the backend native query language. CyGraph automatically infers the structure of its underlying graph model through analysis of the ingested data, which it presents to the user for generating queries in an intuitive way. CyGraph has been experimentally validated in both enterprise and tactical military environments.

Keywords: *common operating picture, situational understanding, mission assurance, graph analytics*

1. INTRODUCTION

Through centuries of experience and modern advances in technology, military commanders can rely on a fairly sophisticated common operating picture (COP) of the kinetic battlespace. However, significant challenges remain for extending the COP to include cyberspace as an operational domain [1]. Such an extended COP is needed for achieving appropriate levels of resilience to attack, maintaining situational awareness and understanding, and providing command and control of cyber (and joint cyber/kinetic) operations [2]. A cyber-extended COP needs to support the analysis of complex interactions among disparate data for decision making.

This paper describes CyGraph, a prototype tool for improving cyber resilience, maintaining situational awareness in the face of cyberattacks, and focusing on protection of mission-critical assets. CyGraph builds rich graph models from various network and host data sources, fusing isolated data and events into a unified model. From this, cyber operators can apply powerful graph queries that uncover multi-step graph reachability from threats to key cyber assets, as well as other patterns of cyber risk. In this way, the tool correlates and prioritizes alerts in the context of vulnerabilities and key assets. CyGraph analytics extract 'needle in haystack' patterns of cyber risk focused on mission assets, with interactive visualization of query results, giving a common operating picture of cyberspace.

Traditional graph formulations with entities (vertices) and relationships (edges) of a single homogeneous type lack the expressiveness required for representing the rich structures involved in analyzing cyber risk. CyGraph employs *property graphs*, i.e., attributed, multi-relational graphs with vertices and edges having arbitrary properties [3]. Property graphs have the power needed for expressing a range of heterogeneous vertex and edge types, which arise from combining data from a variety of sources into a coherent unified cyber security graph model.

Unlike previous graph-based tools that focus on specific analytic use cases against fixed data models, CyGraph employs a schema-free design with a property-graph data model. The specific security data model is defined implicitly, according to how source data are transformed to a property graph. To help analysts more easily work with such complex models, CyGraph automatically infers the underlying data model for a populated graph. It's domain-specific query language provides a simplifying layer of abstraction from the native query language of the graph database implementation.

CyGraph has been tested in military environments, including at the enterprise backbone and tactical command levels. In this paper, for sensitivity reasons, we

describe CyGraph analytics through simulated data; these datasets mimic patterns that we have observed in real data.

2. PREVIOUS WORK

There has been considerable previous work in graph-based approaches to cyber security. For example, a review in 2013 [4] describes hundreds of papers that employ various kinds of graph representations for security, with over 30 categories just for the specific case of modelling network attacks and defenses with acyclic graphs. A more recent study [5] examines over 50 proposed graph-based security models, each having key differences in representation. The state of practice has reached a level of maturity such that various off-the-shelf tools (both commercial and governmental) have emerged for graph analytics in operational environments [6] [7] [8] [9] [10] [11] [12] [13].

The wide range of proposed graph representations address the fundamental issue that classical graph algorithms alone are insufficient for solving analytic problems in cyberspace. Instead, specific data models are needed that capture the structure and semantics of the various kinds of entities and their relationships. But a significant limitation of the current generation of tools is that they have fixed data models, which limits their scope and ability to adapt to changes in operational environments and analytic requirements.

The idea of leveraging graph database technology for cyber security analysis is first explored in 2015 [14]. A proof-of-concept version of the CyGraph tool, which is implemented as a Java-based application running on a single host, is described [15] [16]. The proof-of-concept tool was applied for some security use cases, using simulated data or isolated examples of real operational data [17] [18]. A particular limitation of these preliminary examples is that mission functional dependency relationships are analyzed separately from cyberspace relationships.

Based on our initial success in proving the CyGraph concept, we have developed a more mature and capable CyGraph tool. This advanced prototype is a web-based (JavaScript) client-server application, distributed across three machines (user web browser as GUI, middle-tier intermediary service, and back-end database service). Leveraging this architecture, we have implemented multiple technologies for the CyGraph back-end graph database, including support for Apache Rya [19] within the Big Data Platform (BDP) [20] developed by the US Defense Information Systems Agency (DISA). The advanced CyGraph prototype also integrates with the Elastic Stack [21] (for Neo4j) or Accumulo [22](for DISA BDP) for scalable data ingest and

intermediate storage. A high-level overview of this tool architecture is described in [2], although no specific analytic results are given there.

The new web-based CyGraph tool has been validated using real data in operational network environments, at enterprise-level scale. The analytic examples that we describe in this paper are abstracted versions of the kinds of results we obtained for real data (abstracted here to protect the sensitive real data). This includes the development and validation of joint models for cyberspace and mission functions, e.g., for showing mission risk and/or impact as we describe. The present work also experimentally validates that CyGraph's loosely-coupled client-server architecture can support multiple back-end graph persistence technologies, while insulating the front-end functionality from the choice of back-end implementation. This in turn provides flexibility in matching the analytic architecture to the performance and scaling requirements for a given organization.

3. CYGRAPH MODEL

We begin by defining the formal structures that form the basis of an instance of a CyGraph model. A *graph* $G = (N, E)$ is a pair of sets of nodes and edges. The edges are, themselves, ordered pairs of nodes (n_1, n_2) from N. A *property graph* is a graph in which the nodes and edges come equipped with attributes, that is, arbitrary key/value pairs describing properties of the elements. We generally assume that nodes and edges have some minimal structure. Namely, nodes have attributes for a unique identifier and a type. Edges also have an attribute describing their type. They additionally have attributes identifying their source and destination nodes. Additional attributes may include such things as location information, mission criticality, or traffic packet counts.

A CyGraph model instance is defined by the properties attributed to the nodes and edges, as well as any constraints that may be in effect. Typically, particular property graphs conforming to a CyGraph model instance are progressively built from heterogeneous data sources with records containing information about the nodes and edges. Rather than requiring a fixed schema for the data sources, CyGraph applies data transformations that map elements of the source data to nodes, edges, and their properties. Thus, these data transformations implicitly define an instantiated CyGraph model.

To better understand how a property graph is built, consider the process of reading in a record r from a data source. Assume the graph built so far is $G = (N, E)$, and the transform T extracts information about two nodes, n_1 and n_2, and an edge e between them. The new graph is $G' = (N \cup \{n_1, n_2\}, E \cup \{e\})$, where the properties of n_1, n_2,

and e are defined by the transform T. If n_1 or n_2 was already in N then we simply update their properties according to any extra information contained in record r.

In general, any property (of nodes or edges) that has potential analytic value (in the sense of constraining graph queries) can be included as a node or edge property. The type for a node or edge can then be defined as an arbitrary function of its properties. Thus, the node and edge types depend on the source data, via the transformation to a CyGraph property graph.

For example, alerts from Host Based Security System (HBSS) [23] yield node types describing the category of the alert for the destination node, e.g., whether they are reconnaissance events (such as port scans) or represent actual host compromise. Network flow records yield the region in which the node is located (US, non-US, country of concern) or indicate that the node is key terrain (based on knowledge of services hosted there and mission dependencies). This transform prioritizes type information from HBSS alerts over the other two, so that if a host is in the US and is compromised, it is simply identified as compromised. One could easily define a different transform that extracts a type in the Cartesian product of the types defined by each data source. The choice of transform depends on the sort of questions one wants answered regarding the graph (i.e., the analytic queries).

Once CyGraph has constructed the property graph from its data sources, an analyst can explore the graph with *queries* expressed in CyGraph Query Language (CyQL), a domain-specific query language. An important aspect of graph structure pertains to reachability. CyQL allows for the specification of structural features of trajectories through a graph. When a query Q is applied to a graph G it results in a (possibly empty) subgraph $G' \subseteq G$. This matching subgraph is then displayed in the user interface.

A *directed trajectory* is an alternating sequence of nodes and edges (n_0, e_1, n_1,…, e_k, n_k) in which, for every $0 < i \leq k$, the source of e_i is n_{i-1} and the destination is n_i. An undirected trajectory is similar, except for any edge e_i, its source and destination may be n_i and n_{i-1} respectively. The length of a trajectory is the number of edges. The *graph of a trajectory* is ($\{n_0,\ldots,n_k\}$, $\{e_1,\ldots,e_k\}$) in which the sequence information has been forgotten. A trajectory t is a trajectory of graph $G = (N, E)$, if the trajectory's graph (N', E') is a subgraph of G (i.e. $N' \subseteq N$ and $E' \subseteq E$).

CyQL specifies trajectories by constraining the number of hops, and the types of the initial node, the end node, and the edges. Queries are built from the following clauses with their associated semantics:

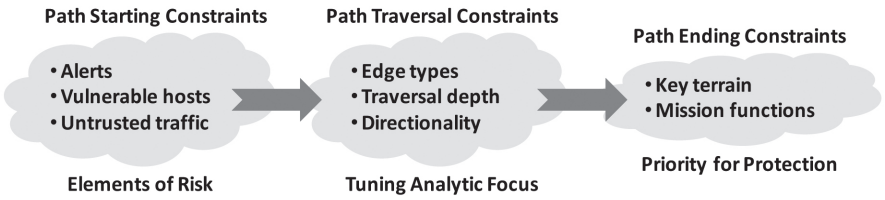
- **`hops($numHops)`**: A trajectory satisfies this clause if its length is **`$numHops`**.
- **`hops($minHops,$maxHops)`**: A trajectory satisfies this clause if its length is between **`$minHops`** and **`$maxHops`** (inclusive).
- **`startType($type)`**: Trajectory (n_0, e_1,…,n_k) satisfies this clause if n_0 is of type **`$type`**.
- **`endType($type)`**: Trajectory (n_0, e_1,…,n_k) satisfies this clause if n_k is of type **`$type`**.
- **`startId($id)`**: Trajectory (n_0, e_1,…,n_k) satisfies this clause if the unique node identifier $u(n_0)$ of node n_0 is equivalent to **`$id`**.
- **`endId($id)`**: Trajectory (n_0, e_1,…,n_k) satisfies this clause if the unique node identifier $u(n_k)$ of node n_k is equivalent to **`$id`**.
- **`edgeTypes($types)`**: A trajectory satisfies this clause if each edge is of one of the types in the comma separated list **`$types`**.
- **`undirected()`**: By default, satisfying trajectories must be directed. When this clause is used, undirected trajectories also satisfy the query.

A CyQL clause is a concatenated sequence of such clauses. A trajectory t satisfies a CyQL query Q (written $t \mid = Q$) if it satisfies all of the clauses. The result of applying Q to graph G is simply the union of all trajectories of G that satisfy Q. That is:

$$Q(G) = \{t \in \text{trajectories of } G\text{: } t \mid = Q\}.$$

CyQL provides a key aspect of risk analysis in CyGraph. In terms of the semantics of attack paths, query trajectory through the property graph corresponds to multi-step attack (or attack reachability) through the network. Conceptually, the aspects of CyQL can be organized as shown in Figure 1.

FIGURE 1. GRAPH TRAJECTORY PATH CONSTRAINTS IN CYGRAPH QUERY LANGUAGE (CYQL)



The left side of Figure 1 represents elements of risk within a network; i.e., things that we are protecting against. By specifying such risk elements as constraints on the starting points of a query traversal, trajectories represent ‘downstream’ relationships

emanating from risk points. Conversely, the right side represents high-valued assets within the environment; i.e., things that we are trying to protect. Defining those things as constraints on the traversal ending points cause paths to be focused on those assets as reachable from the risky elements. CyQL clauses that occur between these starting and ending extremes generally serve to constrain path trajectories in particular ways that help tune analytic focus; e.g., for managing the trade-off between comprehensiveness of query results and cognitive overload.

CyQL queries involve identifying trajectories that start from nodes representing risk elements, and end in nodes representing priorities for protection. The set of trajectories can be further refined by specifying additional traversal constraints regarding the edge types or total path length. This serves to focus an analyst's attention on the relationships that matter the most. By visualizing the results of CyQL queries, CyGraph allows users to quickly identify known risky patterns or anomalous structures that warrant further investigation.

For example, given the appropriate data sources, CyQL makes it straightforward to identify the set of hosts with vulnerabilities that reside within the same connected component as a key cyber asset. By limiting the query to vulnerable hosts within two hops of key cyber assets, one can more easily identify the vulnerable hosts that pose the greatest risks. Queries may also help to identify clusters within the graph that have interesting properties. A highly connected cluster of hosts with host-based alerts may be an indication of vigorous adversarial exploration and exploitation.

4. CYGRAPH ARCHITECTURE

CyGraph ingests data from various sources and normalizes it. It then transforms the elements of the normalized model into a graph model specific to the cyber security domain. Graph queries are issued from the client front end (translated from CyQL to native query language in a middle-tier service) and then executed on the backend database. The resulting query matches are then visualized in the web client (browser).

In this agile architecture, the graph model is defined by how the data sources are transformed into a property graph, rather than conforming to a predetermined schema. Model extensions are simply the creation of additional of nodes, relationships, and properties in the property graph model, and require no schema changes or other database renormalizing. CyGraph supports two options for backend data storage and query processing:

- Neo4j graph database [24] with normalized data in Elasticsearch [21].
- Apache Rya [19] RDF store with normalized data in Apache Accumulo [22].

Each of these options is available as open-source software, and (with the exception of Rya) have commercial support available. The second option (Rya+Accumulo) is available as part of DISA's Big Data Platform (BDP) [20].

In the CyGraph front-end analyst dashboard, graph pattern-matching queries are expressed in CyQL, which CyGraph compiles to Cypher [25] (for Neo4j) or SPARQL [26] (for Rya). This presents a simplifying layer of abstraction, designed specifically for the desired risk analysis, freeing the analyst from learning a complex general-purpose query language.

Typical inputs to CyGraph fall under four categories:

1. Network Infrastructure. This captures the configuration and policy aspects of the network environment.
2. Security Posture. Specification of network infrastructure is combined with vulnerability data to map potential attack paths through the network.
3. Cyber Threats. This captures events and indicators of actual cyberattacks, which are correlated with security posture to provide context for risk analysis and attack response.
4. Mission Dependencies. This captures how elements of enterprise missions depend on cyber assets.

CyGraph relies on other tools and data sources to build its cyber security graphs. For example, the TVA/Cauldron tool [6] [7] [8] [9] can build network attack graphs from host vulnerabilities, firewall rules and network topology. CyGraph can ingest data for both potential and actual threats, including Splunk [27], Wireshark [28], the National Vulnerability Database (NVD) [29], and Common Attack Pattern Enumeration and Classification (CAPEC) [30]. For capturing mission dependencies on cyber assets [17] [18], CyGraph ingests models developed through other tools [16], including Crown Jewels Analysis (CJA) [31] and Cyber Command System (CyCS) [32].

The CyGraph implementation is schema-free, so the model is decoupled from the storage implementation. The particular way in which the data is transformed to a property graph determines an instantiated CyGraph model. So, for example, not all of the data sources in the four categories listed above are necessarily needed for useful analysis – often only a single data source is ingested.

Data is continually streaming in that must be analysed for cyber risk correlation and prioritization by CyGraph. Leveraging the open source Elastic Stack, the Beats platform provides agents for gathering data, with Logstash for transformation and ingest into Elasticsearch. A CyGraph web service then creates a property graph model and imports it into the CyGraph graph data base (Neo4j).

There is a similar analytic flow for CyGraph deployment on BDP, in which data streams are processed by Apache Storm [33], stored in Accumulo [22] and queried in Rya [19]. In this analytic flow, the CyGraph data model is mapped to RDF. The result of a query is a combination of alerts, network flows and vulnerabilities represented as graph nodes and edges. The matching subgraphs for queries are typically orders of magnitude smaller than the full graph stored in Neo4j or Rya.

5. CYGRAPH OPERATION

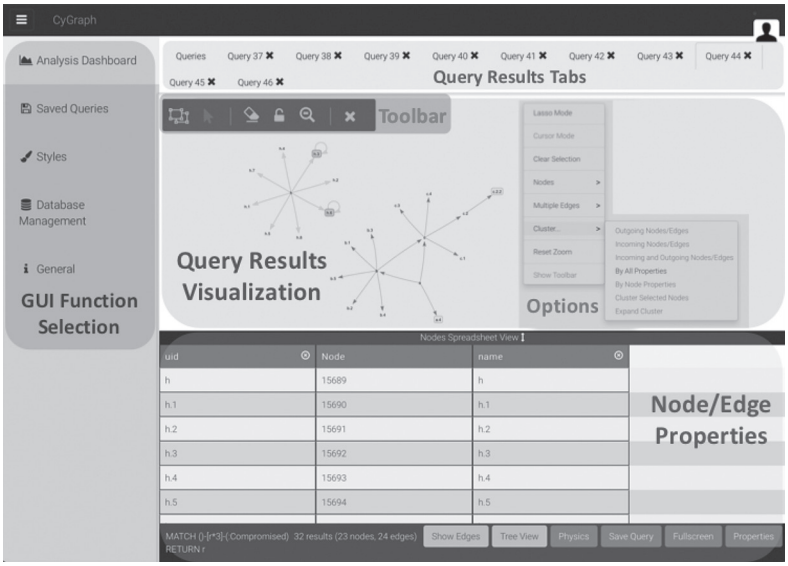
After ingesting data from various sources, CyGraph maps the data to a property graph stored in a graph database. It automatically infers the underlying graph model through inspection of the graph database. It then presents the model to the user in the browser user interface as an interactive graph visualization.

The analyst can interact with this graph model to generate queries in the domain-specific CyQL query language. In particular, user-selected combinations of edge types (diamonds) populate the CyQL **`edgeTypes($types)`** clause, which specifies edge types to be matched in a query. For example, edges of type **`IN`** define relationships between **`Machine`** nodes and **`Domain`** nodes, i.e., network machine membership in protection domains (e.g., subnets) [14].

Core clauses in CyQL define patterns of reachability through a graph, i.e., **`hops($numHops)`**, **`hops($minHops,$maxHops)`**, **`startType($type)`**, **`endType($type)`**, **`startId($id)`**, **`endId($id)`**, **`edgeTypes($type)`**, and **`undirected()`**. CyQL includes other features for matching patterns in the cyber security domain [15], including keywords for host names, IP addresses, subnet address ranges, arbitrary Boolean combinations of clauses and wildcards in parameter values. CyGraph queries are stored for sharing and reuse.

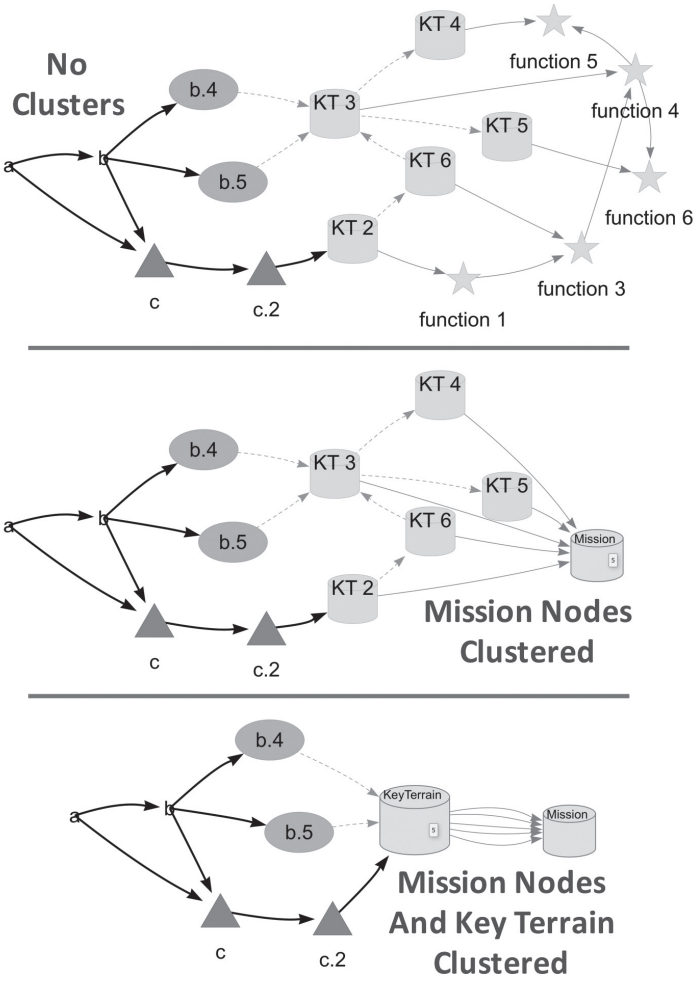
Once a query is executed, CyGraph displays the query results, as shown in Figure 2. Each query submission creates a new query pane, with tabs for selecting panes. The query results (matched subgraph) are visualized in a main panel. Optionally, the properties for selected nodes or edges are displayed below the graph visualization.

FIGURE 2. CYGRAPH WEB USER INTERFACE (QUERY RESULTS)



One of the user-interface options is to cluster elements of the visualized graph in particular ways, i.e., by user-selected nodes, incoming or outgoing edges for a node, or by node type. This is illustrated in Figure 3.

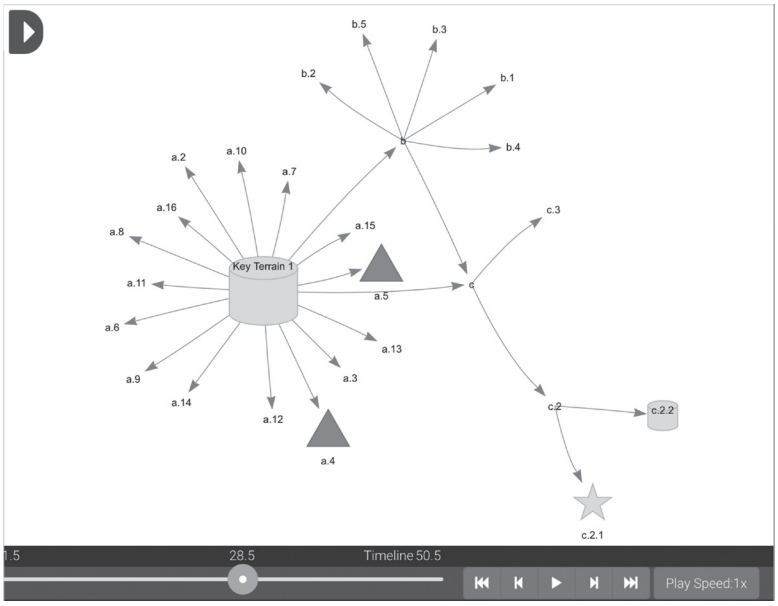
FIGURE 3. CLUSTERING NODES IN GRAPH QUERY VISUALIZATION



The top of Figure 3 is a query result, before clustering is applied. In the middle, clustering is applied via a node property denoting mission functions. At the bottom of the figure, additional clustering is applied, based on a node property denoting key terrain. Visually, such a clustering merges a set of nodes to a single one, with adjacent edges to other (non-clustered) nodes preserved. This kind of interactive visual clustering helps manage the complexity of graph analytics in CyGraph. For example, in Figure 3, clustering the mission and key terrain nodes helps focus attention on the alert destinations (triangles) and vulnerable hosts (ellipses) that are potential risks.

For time-varying models, CyGraph can dynamically visualize the evolving graph state. This is shown in Figure 4. This capability depends on time stamps being defined for edges during the ingestion process. Then, when a query result has a time defined for each edge, the user interface enables the timeline feature. This feature builds a time tick for each discrete event (unique value of time in the query result edge set). The timeline then provides video controls (e.g., play, single step forward/back, speed) for displaying the graph as edges appear over time.

FIGURE 4. INTERACTIVE TIMELINE FOR VISUALIZING GRAPH EVOLUTION OVER TIME

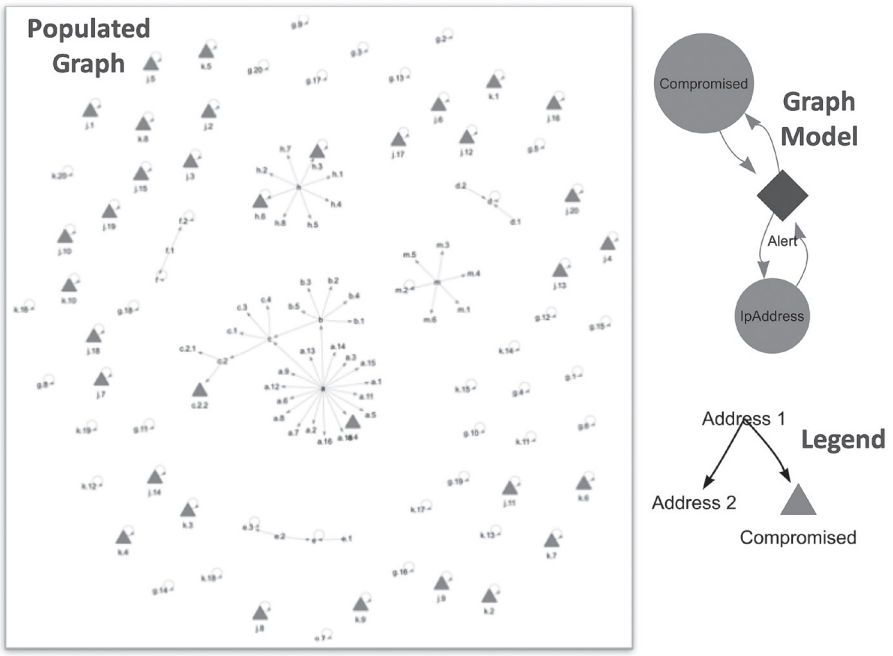


6. EXAMPLE CYGRAPH ANALYTICS

In this section, we describe a number of example applications of CyGraph for security analytics. These examples all use simulated data sets (thus avoiding sensitivity issues), which are designed to mimic patterns that we have observed in real datasets.

The first example (Figure 5) is based on intrusion detection alerts. CyGraph automatically infers the model (top left of the figure) from the populated graph. Nodes are typed as either **`Compromised`** (for destinations of alerts reporting compromise) or **`IpAddress`** (sources/destinations for other general kinds of alerts), rendered as in the legend. An edge is one or more alerts from source to destination.

FIGURE 5. GRAPH MODEL BASED ON INTRUSION ALERTS

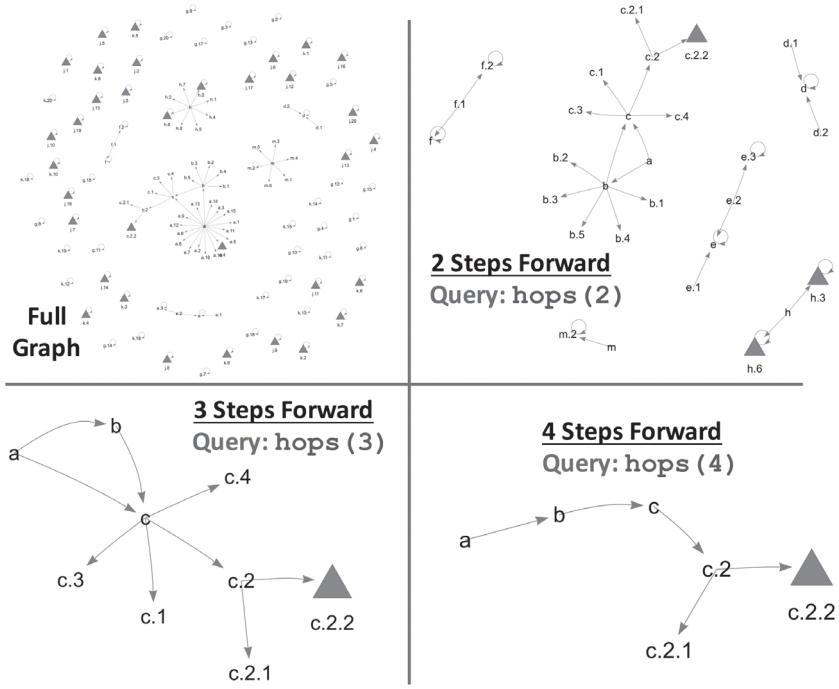


One use for risk analysis is identifying attack reachability in a particular direction, consistent with adversary lateral movement. This pattern is expressed in the CyQL query language via the **`hops($numHops)`** clause, as examined in Figure 6.

The upper left of Figure 6 shows the results for the full (unconstrained) query. The other portions of the figure show query results for **`hops(2)`** (upper right), **`hops(3)`** (lower left), and **`hops(4)`** (lower right). Queries with larger values of **`$numHops`** are more tightly constrained, in the sense of matching deeper traversal. Smaller (more loosely constrained) values of $numHops yield larger matching subgraphs.

Operationally, an analyst can adjust trajectory depth according to analytic need. One can begin with a larger value of **`$numHops`** to discover deep network infiltration as a higher-priority incident. Then, as deeper-level (and more rarely occurring) incidents are resolved, more shallow ones can be investigated. For example, an organization can set the trajectory depth such that there are available resources available to investigate the resulting graph query match.

FIGURE 6. GRAPH QUERY RESULTS FOR DIFFERENT TRAJECTORY DEPTHS



Clauses in CyGraph can be combined for further constraining query results. Semantically, this is a conjunction (Boolean AND), in the sense that conditions in all clauses must match in the query results. This is examined in Figure 7. Here, we combine the **`hops()`** clause with **`endType()`**, which constrains matching paths to end with a node of type **`Compromised`**.

As a use case for operational security, this example focuses on a more severe intrusion alert category as the locus of potential lateral movement by an adversary. Comparing the upper left of Figure 7 (no **`endType`** constraint) with the upper right of Figure 6 (with **`endType`** constraint), we see the result of constraining the **`endType`** (for trajectory depth 2). The query result is much smaller, with all trajectories ending at nodes of type **`Compromised`**. In terms of security analysis, this focuses on paths leading to (reportedly) compromised hosts, e.g., for investigating events leading up to those in question. We see the same kind of result for Figure 7 (upper right) versus Figure 6 (lower left), this time with a trajectory depth of 3. For alert response, this is tracing the investigation deeper into the potential attack.

We now consider a more complex CyGraph example, shown in Figure 8. Like real-world data, such an unconstrained graph visualization is difficult to understand in its entirety. This underscores the need for CyGraph to extract 'needle in haystack' patterns of cyber risk, focused on mission protection.

FIGURE 7. MULTIPLE CLAUSES IN QUERIES

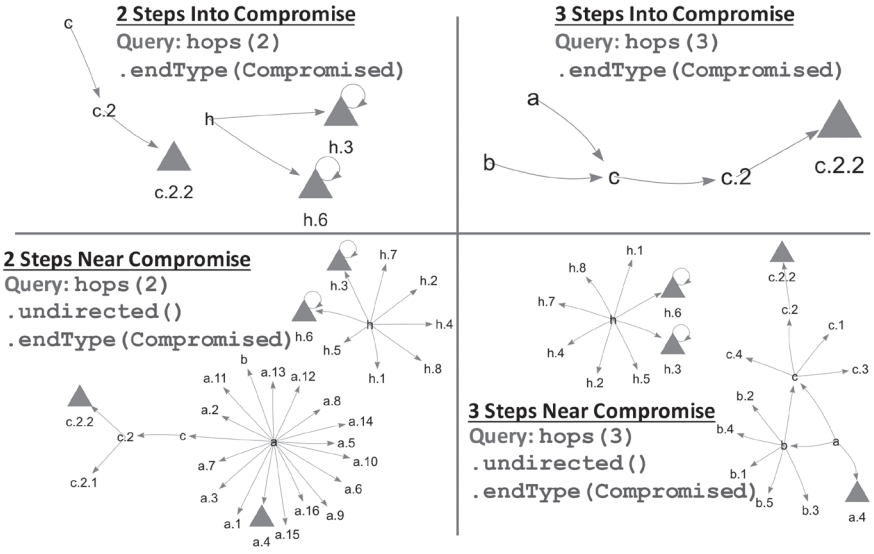
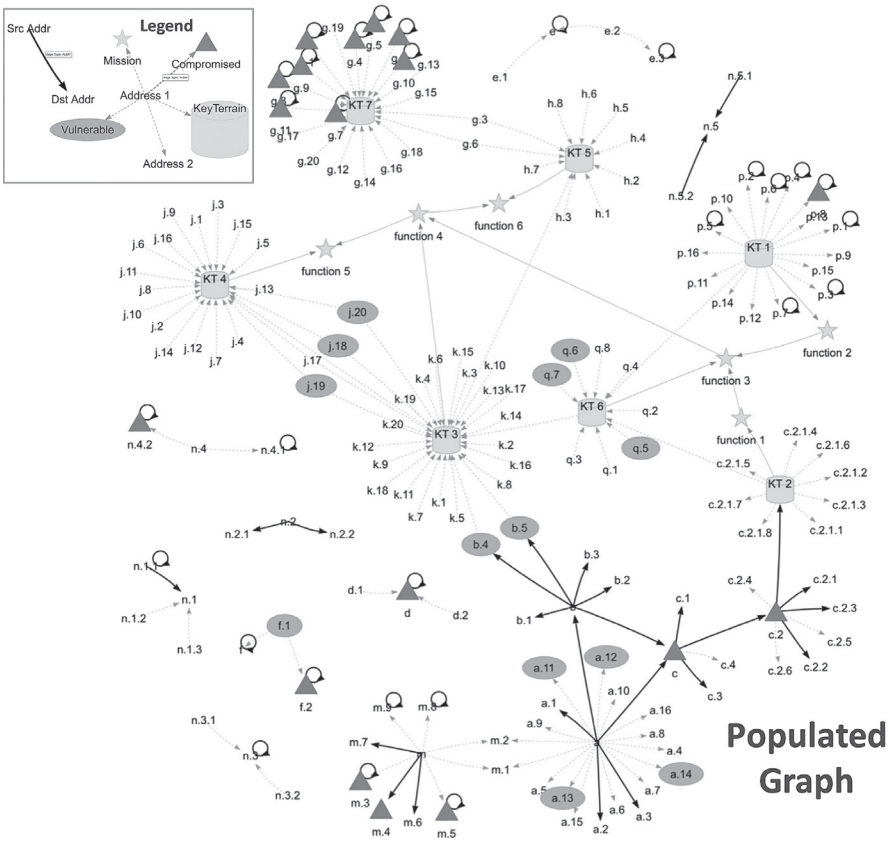
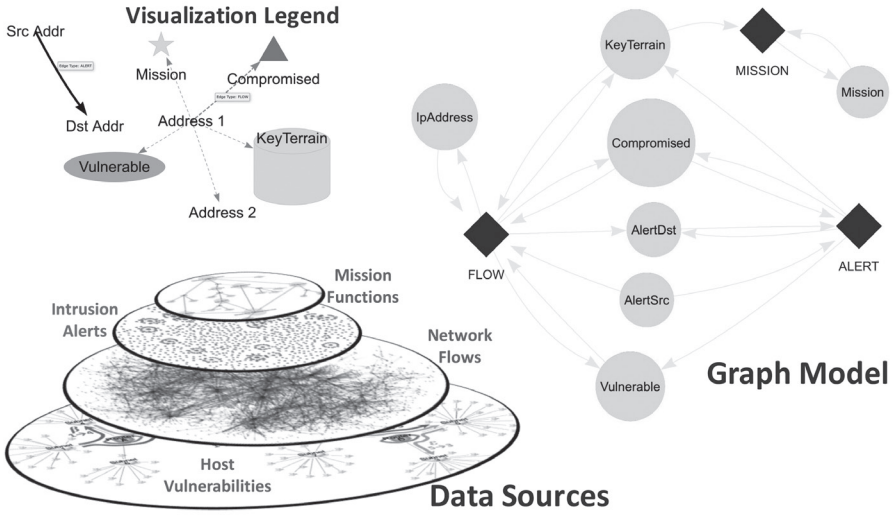


FIGURE 8. GRAPH POPULATED FROM MISSION FUNCTIONS, INTRUSION ALERTS, NETWORK FLOWS, AND HOST VULNERABILITIES



The graph in Figure 8 is populated via a process that transforms host vulnerabilities, network flows, intrusion alerts and mission functional dependencies (i.e., the data sources in Figure 9) to a property-graph model. CyGraph automatically infers the model from the populated graph database, which is the right side of Figure 9.

FIGURE 9. GRAPH VISUALIZATION LEGEND, DATA SOURCES, AND GRAPH MODEL FOR FIGURE 8



In this model, mission nodes are connected to one another (and to key cyber terrain) in terms of their dependencies (from 'provides' to 'needs'). Alert edges connect source and destination nodes of various types: key terrain, compromised (assumed vulnerable), vulnerable (not compromised), and other general alerts sources and destinations. General addresses observed in network flows which are not associated with alerts are connected to each other and to alert addresses via flow edges. In this way, network flows serve to fill in potential gaps from adversary activity not detected by intrusion detection (false negatives).

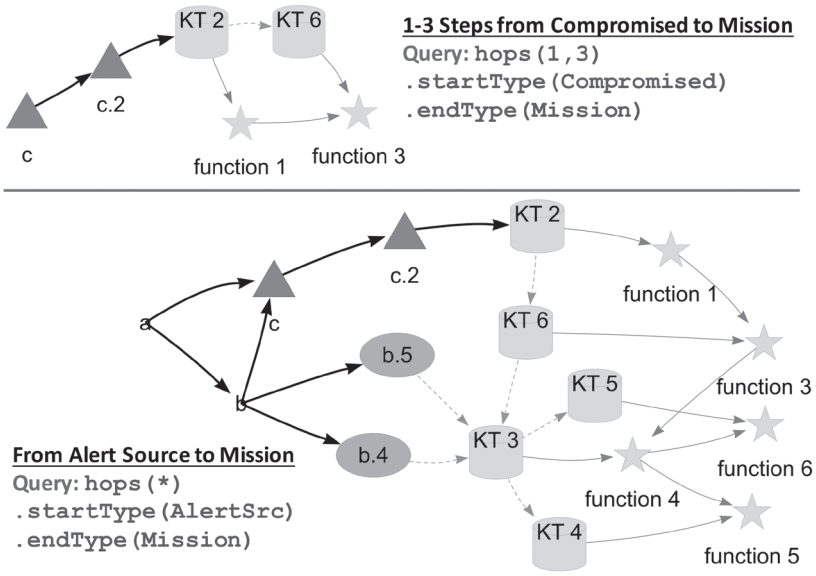
We now apply queries to the graph in Figure 8, in which various combinations of CyQL clauses match subgraphs of interest for analyzing this richer security model. These query clauses generally follow the pattern of constraining paths to start at risky elements and end at high-value mission elements, with intermediate constraints that tune analytic focus. Our examples here also demonstrate another kind of strategy for operational security – tightly constraining queries to initially focus on riskier patterns, then subsequently relaxing constraints to uncover new patterns of the next higher priority.

The top of Figure 10 is the query result for a significantly risky pattern – reported compromises that lead to mission functions within three steps. This query result shows that a compromised node is the source of another alert whose destination is key cyber terrain which supports a mission function. There is also traffic flow (dashed arrow) to

another key terrain node that supports a mission function. The traffic from *KT 2* to *KT 6* might warrant deeper inspection for potential missed detections (false negatives).

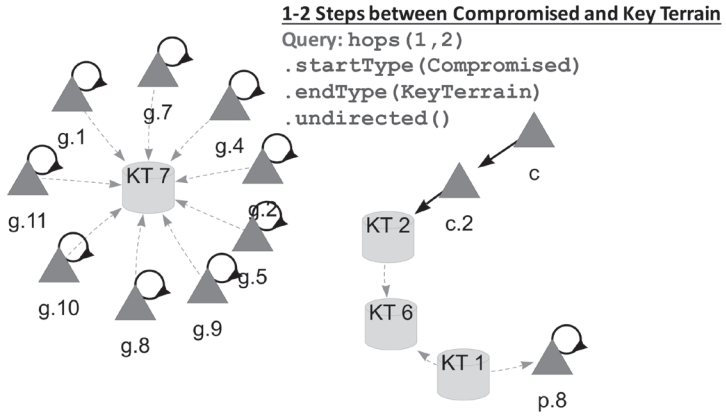
In the bottom of Figure 10, query constraints are relaxed somewhat to expand the analytic scope. In particular, the unlimited trajectory depth via **`hops(*)`** admits paths of any depth leading to mission nodes, and **`startType(AlertSrc)`** has paths starting at alert sources (any severity of alert) rather than compromised destinations. This query result shows additional alert trajectories (all starting from node a), including ones that end on vulnerable hosts, which have traffic to other key terrain supporting other mission functions.

FIGURE 10. RISKY PATHS TO MISSION FUNCTIONS



Next, we apply the **`undirected()`** clause of CyQL, which explores nearness by ignoring path directionality. This is shown in Figure 11. Here, we again apply **`startType(Compromised)`**, along with **`endType(KeyTerrain)`**, which stops at key terrain rather than going beyond to mission functions that depend on them. We also apply a more constrained **`hops(1,2)`** that admits only paths of depths one or two.

FIGURE 11. IGNORING DIRECTIONALITY IN QUERIES

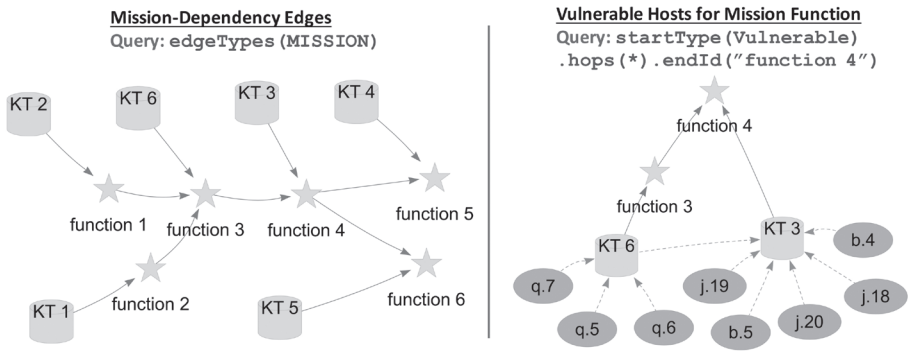


This query finds compromise-related paths in addition to those in Figure 10. This includes nine compromised nodes that communicate with key terrain *KT 7*, which we find by having the query end at key terrain rather than mission functions. As we show in Figure 12, key terrain *KT 7* does not have a known mission function that it supports, so this query identifies risk to such nodes. The query in Figure 11 also finds compromised node *p.8*, which communicates with *KT 1*. In this case, the network flow has *KT 1* as the source (e.g., the initiator of the flow). By ignoring directionality, this admits the possibility of general communication types, e.g., involving attacks against client-side vulnerabilities.

The left side of Figure 12 shows all mission dependencies in this graph model. Mission dependencies are represented as edges of type **`MISSION`**, between key terrain or mission functions, oriented from 'provides' to 'needs.' Thus, the CyQL clause **`edgeType(MISSION)`**, with no other query conditions, finds all such dependencies. Formally, because there is no **`hops`** clause, the query result is the union of edges rather than path trajectories.

The right side of Figure 12 finds all vulnerable hosts that are relevant to a particular mission function. The clause **`startType(Vulnerable)`** causes paths to start at vulnerable nodes. The clause **`endId('function 4')`** causes paths to end at *function 4*. The **`hops(*)`** allows paths of any depth.

FIGURE 12. ALL MISSION DEPENDENCIES (LEFT) AND VULNERABILITIES FOR A PARTICULAR MISSION FUNCTION (RIGHT)



7. SUMMARY AND CONCLUSIONS

Maintaining situational understanding and a common operating picture in cyberspace requires making sense of complex relationships among aspects as varied as security posture, cyber threats, security alerts, and mission dependencies on cyber assets. The volume and complexity of data needed for security operations are far too large for manual inspection or analysis. These challenges multiply with the need to go beyond considering isolated events, matching single-step rules, or generating summary statistics, which yield limited insight into complex adversary actions.

CyGraph creates a unified multi-relational graph model of cyber terrain, events, and mission dependencies. This rich repository of relationships among cyberspace and mission elements supports advanced analytic and visual capabilities. Through pattern-matching queries, CyGraph discovers clusters of high-risk activity from the swarm of complex interrelationships. This allows cyber operators to more easily understand evolving cyberattack situations, and to recommend best courses of action to commanders. By including mission dependencies on cyber assets, CyGraph shows how cyberspace activities influence mission success.

In CyGraph, domain-specific graph queries extract nuggets of important patterns from the swarm of data through query clauses that fine-tune graph path trajectories during query matching. These queries uncover multi-step graph reachability from vulnerabilities and threats to key cyber assets and mission functions. The domain-specific language provides a layer of abstraction that simplifies the operational burden. CyGraph also infers the underlying data model from a populated graph database, presenting that to the analyst to further aid in formulating queries.

CyGraph has a schema-free data model for flexibility in combining various types of relationships, aimed at addressing a wide variety of analytical questions. It is implemented as a 3-tier client-server web application with a graph database or triple store backend and interactive graphical interface in the browser. CyGraph employs a combination of powerful graph-based queries and advanced interactive visualization. It thus provides a significant capability to enable the storage and processing of diverse, mission-relevant cyber data at scale while making the technology readily accessible to cyber analysts. This in turn enables more accurate and rapid decision making for command and control.

CyGraph includes a number of custom capabilities for interactively visualizing and navigating graph query results. This includes clustering nodes according to various criteria, and dynamic rendering of time-varying graph evolution. Overall, these analytic and visual capabilities enable the discovery and understanding of 'needle in haystack' patterns of cyber risk focused on mission assets.

REFERENCES

[1] G. Conti, J. Nelson, and D. Raymond, 'Towards a Cyber Common Operating Picture,' in *5th NATO International Conference on Cyber Conflict*, Tallinn, Estonia, 2013.

[2] S. Noel, D. Bodeau, and R. McQuaid, 'Big-Data Graph Knowledge Bases for Cyber Resilience,' in *NATO IST-153 Workshop on Cyber Resilience*, Munich, Germany, 2017.

[3] M. Rodriguez and J. Shinavier, 'Exposing Multi-Relational Networks to Single-Relational Network Analysis Algorithms,' *Journal of Informetrics*, vol. 4, no. 1, pp. 29-41, 2009.

[4] B. Kordy, L. Piètre-Cambacédès, and P. Schweitzer, 'DAG-Based Attack and Defense Modeling: Don't Miss the Forest for the Attack Trees,' *Computer Science Review*, Vols. 13-14, 2014.

[5] H. S. Lallie, K. Debattista, and J. Bal, 'An Empirical Evaluation of the Effectiveness of Attack Graphs and Fault Trees in Cyber-Attack Perception,' *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, 2017.

[6] S. O'Hare, S. Noel, and K. Prole, 'A Graph-Theoretic Visualization Approach to Network Risk Analysis,' in *IEEE Workshop on Visualization for Computer Security*, Cambridge, MA, 2008.

[7] S. Jajodia, S. Noel, P. Kalapa, B. O'Berry, M. Jacobs, E. Robertson, and R. Weierbach, 'Network Attack Modeling, Analysis, and Response'. US Patent 7,904,962, 8 March 2011.

[8] S. Noel and S. Jajodia, 'Attack Graph Aggregation'. US Patent 7,904,962, 1 December 2009.

[9] S. Noel and S. Jajodia, 'Metrics Suite for Network Attack Graph Analytics,' in *9th Annual Cyber and Information Security Research Conference (CISRC)*, Oak Ridge National Laboratory, Tennessee, 2014.

[10] K. Ingols, R. Lippmann, and K. Piwowarski, 'Practical Attack Graph Generation for Network Defense,' in *Annual Computer Security Applications Conference*, 2006.

[11] RedSeal Networks, [Online]. Available: http://www.redsealnetworks.com/. [Accessed 18 February 2018].

[12] Skybox Security, [Online]. Available: http://www.skyboxsecurity.com/. [Accessed 18 February 2018].

[13] Sqrrl, [Online]. Available: https://sqrrl.com. [Accessed 18 February 2018].

[14] S. Noel, E. Harley, K. H. Tam, and G. Gyor, 'Big-Data Architecture for Cyber Attack Graphs: Representing Security Relationships in NoSQL Graph Databases,' in *IEEE Symposium on Technologies for Homeland Security (HST)*, Boston, Massachusetts, 2015.

[15] S. Noel, E. Harley, K. H. Tam, M. Limiero, and M. Share, 'CyGraph: Graph-Based Analytics and Visualization for Cybersecurity,' in *Cognitive Computing: Theory and Applications, Handbook of Statistics 35*, Elsevier, 2016.

[16] S. Noel and W. Heinbockel, 'An Overview of MITRE Cyber Situational Awareness Solutions,' in *NATO Cyber Defence Situational Awareness Solutions Conference*, Bucharest, Romania, 2015.

[17] W. Heinbockel, S. Noel, and J. Curbo, 'Mission Dependency Modeling for Cyber Situational Awareness,' in *NATO IST-148 Symposium on Cyber Defence Situation Awareness*, 2016.
[18] S. Noel, J. Ludwig, P. Jain, D. Johnson, R. K. Thomas, J. McFarland, B. King, S. Webster, and B. Tello, 'Analyzing Mission Impacts of Cyber Actions (AMICA),' in *NATO IST-128 Workshop on Cyber Attack Detection, Forensics and Attribution for Assessment of Mission Impact*, Istanbul, Turkey, 2015.
[19] R. Punnoose, A. Crainiceanu, and D. Rapp, 'Rya: A Scalable RDF Triple Store for the Clouds,' in *1st International Workshop on Cloud Intelligence*, Istanbul, Turkey, 2012.
[20] Defense Information Systems Agency (DISA), 'DISA's Big Data Platform and Analytics Capabilities,' [Online]. Available: http://www.disa.mil/newsandevents/2016/Big-Data-Platform. [Accessed 30 May 2017].
[21] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*, Sebastopol, CA: O'Reilley Media, 2015.
[22] The Apache Software Foundation, 'Apache Accumulo®,' [Online]. Available: https://accumulo.apache.org. [Accessed 30 May 2017].
[23] Wikipedia, 'Host Based Security System,' [Online]. Available: https://en.wikipedia.org/wiki/Host_Based_Security_System. [Accessed 31 May 2007].
[24] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*, Second ed., Sebastopol, CA: O'Reilly Media, 2015.
[25] E. Eifrem, 'Meet openCypher: The SQL for Graphs,' [Online]. Available: https://neo4j.com/blog/open-cypher-sql-for-graphs/. [Accessed 30 May 2017].
[26] W3C Recommendation, 'SPARQL 1.1 Query Language,' 21 March 2013. [Online]. Available: https://www.w3.org/TR/sparql11-query/. [Accessed 30 May 2017].
[27] 'What Is Splunk?,' [Online]. Available: https://www.splunk.com. [Accessed 31 May 2017].
[28] 'About Wireshark,' [Online]. Available: https://www.wireshark.org. [Accessed 31 May 2017].
[29] 'NVD – National Vulnerability Database,' [Online]. Available: https://nvd.nist.gov.
[30] S. Noel, 'Interactive Visualization and Text Mining for the CAPEC Cyber Attack Catalog,' in *ACM Interactive User Interfaces Workshop on Visual Text Analytics*, 2015.
[31] The MITRE Corporation, 'Crown Jewels Analysis,' [Online]. Available: http://www.mitre.org/publications/systems-engineering-guide/enterprise-engineering/systems-engineering-for-mission-assurance/crown-jewels-analysis. [Accessed 31 May 2017].
[32] The MITRE Corporation, 'Cyber Command System (CyCS),' [Online]. Available: http://www.mitre.org/research/technology-transfer/technology-licensing/cyber-command-system-cycs. [Accessed 31 May 2017].
[33] The Apache Software Foundation, 'Apache Storm,' [Online]. Available: http://storm.apache.org. [Accessed 1 June 2017].
[34] Defense Information Systems Agency, 'Assured Compliance Assessment Solution (ACAS),' [Online]. Available: http://www.disa.mil/cybersecurity/network-defense/acas. [Accessed 24 May 2017].
[35] S. McGillicudy, 'Flow Data is Top Source for Network Analysis,' [Online]. Available: https://www.kentik.com/flow-data-is-top-source-for-network-analysis/. [Accessed 31 May 2017].
[36] Sandia National Laboratories, 'Computer & Information Sciences (Labs Accomplishments May 2017),' [Online]. Available: http://www.sandia.gov/news/publications/lab_accomplishments/articles/2017/. [Accessed 11 July 2017].

BIOGRAPHIES

This section includes the biographies of the editors and co-editors and of those authors who presented their research at the conference.

Editors and Co-Editors

Dr. **Joe Burton** is Senior Lecturer at the New Zealand Institute for Security and Crime Science, University of Waikato. His research focuses on cyber security, NATO, and the impact of science and technology on international security. Joe holds a Doctorate in International Relations from the University of Otago and is the author of *NATO's Durability in a Post-Cold War World*. He has recently been a visiting researcher at the NATO Cooperative Cyber Defence Centre of Excellence (CCD COE) and is a recipient of the Taiwan Fellowship and US State Department's Study of the US Institutes (SUSI) fellowship.

Torsten Corall is a Senior Analyst in the Law Branch at the NATO CCD COE. He has professional legal expertise from his service in the Armed Forces of the Federal Republic of Germany (Bundeswehr) in the cyber field from 2010. Prior to his work at the NATO CCD COE, he worked in different legal positions of the Bundeswehr both at the levels of Higher Command and Joint Forces Operational Command. He has also extensive experience as a law scholar from his time at the Bundeswehr Signal School. As the military legal adviser to Force Commanders, he has taken part in international military operations.

Cpt **Raik Jakschis** is a member of the NATO CCD COE Technology Branch and currently focuses on cyber security research of ICS and SCADA systems. Prior to taking up his post at NATO CCD COE, Raik worked at the Bundeswehr Communication and Information Systems Service Centre (BwCISSC) to enhance and secure IT infrastructure of the German Armed Forces. He holds a Master of Science in Information Technology from Helmut-Schmidt-University, University of the Bundeswehr Hamburg.

Lauri Lindström has been a researcher at NATO CCD COE since May 2013. Prior to that, he worked in the Estonian Ministry of Foreign Affairs (2007-2012) as Director General of Policy Planning and held various positions in the Ministry of Defence (1995-2007) dealing mainly with issues related to international cooperation, Estonia's accession to NATO, defence planning and security policy. Lauri holds a Ph.D. from Tallinn University, Estonia.

Tomáš Minárik is a researcher in the NATO CCD COE's Law Branch. He holds a law degree from Charles University in Prague and has worked as a legal adviser at the International Law Department of the Czech Ministry of Defence, and at the National Cyber Security Centre of the Czech Republic. His current research focuses on the legal aspects of cyberspace operations, the right to privacy, anonymity networks, and the activities of international organisations in cyberspace.

Maarja Naagel is a researcher in the Law Branch of the NATO CCD COE. Her research area is public international law, including the law of armed conflict, and how it applies in cyberspace. In addition to her research efforts, she also participates in international cyber exercises and provides training in international law. Before joining the Centre, she served as a Legal Adviser at the Estonian Ministry of Defence. Maarja received her Master's degree from the University of Tartu, has participated in a number of specialised international law courses and studied at the Baltic Defence College Civil Servants' Course. Before entering the realm of national defence and international security, she worked in cultural administration (Tallinn Black Nights Film Festival) and legal translation (European Court of Justice).

Ann Väljataga works as a researcher at NATO CCD COE, where her areas of expertise cover national cyber security strategies and public international law. She holds a BA in Law from the University of Tartu and a Master's in Law and Technology from Tallinn University of Technology. Previously she has conducted legal research at the Estonian Human Rights Centre where she focused on privacy and data protection, and the European Union Agency for Fundamental Rights (FRA) where her work examined the fundamental rights implications of untargeted surveillance and biometric border control systems.

Authors

Brad Bigelow is Principal Technical Advisor to the Deputy Chief of Staff for Communications, Information Systems and Cyber Defence, SHAPE. He has over thirty years' experience in military communications, information security, space operations and project and programme management, including twenty-five years as an officer in the US Air Force. He served on the staff of the President's National Security Telecommunications Advisory Committee and on the Core Committee for the recent update of the *Standard for Program Management*. In his current position, he has been intimately involved in the development of the concept and structure for the proposed Cyberspace Operations Centre at SHAPE.

Dr. **Aaron F. Brantly** is Assistant Professor in the Department of Political Science and Hume Center for National Security and Technology Affiliated Faculty at Virginia Tech, and Cyber Policy Fellow at the U.S. Army Cyber Institute. He holds a Ph.D. in Political Science from the University of Georgia and a Master's of Public Policy from American University. His research focuses on national security policy issues in cyberspace including terrorism, intelligence, decision-making, and human rights. His books include: *The Decision to Attack: Military and Intelligence Cyber Decision-Making* and *US National Cybersecurity: International Politics, Concepts and Organization*.

Kenneth Geers (PhD, CISSP) is a COMODO senior research scientist based in Toronto, Canada and a non-resident senior fellow at the Atlantic Council's Cyber Statecraft Initiative. In addition, he is a NATO CCD COE ambassador, an affiliate with the Digital Society Institute-Berlin, and a visiting professor at Taras Shevchenko National University of Kyiv in Ukraine. Kenneth spent twenty years in the US Government, with time in the US Army, the National Security Agency (NSA), the Naval Criminal Investigative Service (NCIS), and NATO, and was a senior global threat analyst at FireEye. He is the author of *Strategic Cyber Security*, editor of *Cyber War in Perspective: Russian Aggression against Ukraine*, editor of *The Virtual Battlefield: Perspectives on Cyber Warfare*, technical expert to the *Tallinn Manual on the International Law Applicable to Cyber Warfare*, and author of more than twenty articles and chapters on cyber security. Follow him on Twitter: @KennethGeers.

Dr. **Roman Graf** is a research engineer at the Center for Digital Safety & Security in the Austrian Institute of Technology GmbH. He works on cyber security and data analytics, contributing to the development of several European research projects like Ecossian, Planets, Assets and SCAPE. He has published widely in the area of cyber security and risk management in digital preservation, being an active member of the Open Preservation Foundation (OPF). Roman also supported the development of the cyber threat intelligence solution CAESAIR, serving as one of the key developers, and he contributed a module to the MISP Open Source Threat Intelligence Platform.

Dr. **David Gugelmann** is a security analytics researcher and CEO of the ETH spin-off Exeon Analytics Ltd. Prior to founding Exeon Analytics in 2016, he was a postdoctoral researcher at ETH Zurich in the Networked Systems Group. His research interests are in big data analytics, digital forensics and machine learning for anomaly detection.

Kim Hartmann specialises in computer security and mathematical modelling, protocol security analysis, computer security risk assessment, and risk analysis of critical network infrastructures. As a member of the Department for Cyber and Information Security at the Conflict Studies Research Centre, Cambridge, UK, her

work focuses on secure network design principles, risk analysis, and assessment of networks, network components, and protocols. Kim is a regular contributor to research projects and conferences on cyber and network security. As an EU expert, she is regularly involved in the assessment of cyber and other security proposals within Horizon 2020 and related projects.

Quentin E. Hodgson is a senior researcher at the RAND Corporation. He came to RAND in 2017 from the MITRE Corporation, where he led projects supporting the Department of Homeland Security and the State Department on strategic planning, cyber security and capacity building. Prior to that, he spent 13 years at the Department of Defense in a variety of strategy and policy positions, including serving as the Director for Cyber Plans, Operations and Programs from 2011-2014. He was the principal author of Secretary of Defense Robert Gates's National Defense Strategy and led efforts to reform the Department's approach to force planning and analysis after the 2010 Quadrennial Defense Review. His primary research interests are cyber operations, risk management and decision-making. He holds Master's degrees from the Johns Hopkins University School of Advanced International Studies and the Industrial College of the Armed Forces, and was a Fulbright Scholar affiliated to the University of Potsdam, Germany.

Dr. **Krisztina Huszti-Orban** is a research fellow at the Human Rights Center of the University of Minnesota and a senior legal advisor to the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism. Her research focuses on the relationship of human rights to the emergence of stand-alone international security regimes regulating counter-terrorism. Krisztina has previously worked with the Office of the United Nations High Commissioner for Human Rights, the Human Rights, Big Data and Technology Project at the University of Essex, and the European Court of Human Rights. She holds a PhD in international law from the Graduate Institute of International and Development Studies, LLM degrees from the Geneva Academy of International Humanitarian Law and Human Rights and the Andrassy Gyula University, Budapest, and a degree in law from the Babes-Bolyai University in Cluj-Napoca.

Commander Dr. habil. **Robert Koch** is a General Staff Officer of the German Federal Armed Forces. He received a Diploma in Computer Science in 2002 and since then has had comprehensive operational and technical training in the German Navy and built up broad experience in the design, implementation and operation of high-security networks and systems while being Deputy Weapon Engineering and Weapon Engineering Officer onboard German frigates. Robert received his PhD in 2011 and his habilitation in 2017. He is now a senior research assistant and lecturer in Computer Science at the Universität der Bundeswehr and the University of Bonn.

His main areas of research are network and system security with a focus on intrusion detection in encrypted networks, security of COTS products, security visualization and the application of artificial intelligence. Currently, Robert is building up the new penetration testing capability at the Cyber-Security Centre of the Federal Armed Forces.

Jeff Kosseff is an Assistant Professor in the United States Naval Academy's Cyber Science department, where he teaches cybersecurity law and policy. He is the author of *Cybersecurity Law*, published in 2017, and *The Twenty-Six Words That Created the Internet, a history of Section 230 of the Communications Decency Act* to be published shortly. He clerked for Judge Milan D. Smith, Jr., of the U.S. Court of Appeals for the Ninth Circuit and for Judge Leonie M. Brinkema of the U.S. District Court for the Eastern District of Virginia. Jeff is a graduate of Georgetown University Law Center and the University of Michigan. Before becoming a lawyer, he was a journalist for *The Oregonian* and was a finalist for the Pulitzer Prize for national reporting.

Martin Libicki (Ph.D., U.C. Berkeley 1978) holds the Keyser Chair of Cybersecurity Studies at the U.S. Naval Academy. In addition to teaching, he carries out research into cyberwar and the general impact of information technology on domestic and national security. He is the author of a 2016 textbook on cyberwar, *Cyberspace in Peace and War*, and *Conquest in Cyberspace: National Security and Information Warfare* as well as various related RAND monographs. Prior employment includes twelve years at the National Defense University, three years on the Navy Staff (logistics) and three years with the US General Accounting Office.

Dr. **Asaf Lubin** is a Lecturer at Yale College, a resident fellow at the Yale Law School Information Society Project, a Post-Doctoral Fellow at the Fletcher School of Law and Diplomacy's Cyber Security and Policy Program, and a visiting scholar at the Hebrew University of Jerusalem's Cyber Security Research Center. Asaf's research focuses on the international regulation of espionage and draws on his experiences as a former intelligence analyst and Sergeant Major (Res.) with the Israeli intelligence community, as well as his vast practical training in national security law and foreign policy, including through his extensive experience with the Israeli Ministry of Foreign Affairs and the Turkel Public Commission of Inquiry. His work also reflects his time spent as a Robert L. Bernstein International Human Rights Fellow with Privacy International in London. He received his JSD and LLM degrees from Yale Law School and his joint LLB/BA in law and international relations from Hebrew University of Jerusalem. You can follow him on twitter at @AsafLubin.

Mirco Marchetti is a researcher at the University of Modena and Reggio Emilia (Italy). He obtained his PhD in 2009 with a dissertation on parallel and distributed

system for cooperative network intrusion detection. His research interests include cybersecurity for automotive, IoT and industry 4.0, as well as the design of secure Cloud services. He is an experienced security professional and teaches cybersecurity courses at postgraduate Master's degree in Digital Forensics and Cyber Technologies (Telecommunication School of the Armed Forces), at the advanced training course "Cyber Academy", and at the University of Modena and Reggio Emilia.

Roland Meier is a second year PhD student at the department of electrical engineering and information technology at ETH Zürich. His research focuses on the security of computer networks. In particular, he works on solutions which leverage recent advances in network programmability to make networks able to detect and mitigate attacks in the data plane and to provide more security and privacy. Roland received his Master's degree in electrical engineering and information technology from ETH Zürich in 2015.

Daniel Moore is a PhD researcher at King's College London's Department of War Studies where he focuses on the application of network operations to military doctrine and strategy. Daniel has experience in both the public and private sectors, having previously served as a lieutenant in the Israel Defense Forces and later working in companies such as IBM. Alongside his academic research, Daniel works as a security principle in Accenture Security's threat intelligence unit, iDefense, where he develops new collection and analysis capabilities.

Dr. **Steven Noel** is a researcher in MITRE's Cyber Security Technical Center. For nearly 20 years, he has led multi-disciplinary teams conducting advanced research in cyber security, particularly in the areas of vulnerability path analysis, optimal network hardening, cyber situational understanding, attack response, and mission impact analysis. He has over 70 publications, with 3000+ citations (h-index 27), holds 10 patents, and led the development of the Cauldron and CyGraph tools for cybersecurity graph analytics/visualization. He earned his PhD in Computer Science from the University of Louisiana at Lafayette in 2001.

Kārlis Podiņš is a Senior Threat Analyst with the Latvian Computer Emergency Response Team (CERT.LV), and the first Latvian government representative to the NATO CCD COE.

Francisco Jesús Rubio Melón has over 20 years' experience in software engineering and 7 years' experience in web security. He was a NATO CCD COE researcher from October 2015 till January 2018, where he carried out research into different web security aspects. He is currently a Spanish Joint Cyber Command member as web

applications specialist. In CyCon 2018 he will be presenting part of his research from the CCD COE, as well as conducting a workshop on web application security.

Cedric Sabbah is Director for International Cybersecurity and IT Law Affairs at the Office of the Deputy Attorney General (International Law) in Israel's Ministry of Justice. He advises Israel's National Cyber Directorate and other government departments on questions of cyber security, internet governance and information technology as they relate to international law. He was involved with the Israeli government's activities and positions with respect to the 2015 UN GGE and the Tallinn Manual 2.0. He has served as a foreign clerk at Israel's Supreme Court, and then as legislative counsel in Canada's Department of Justice (2002-2006). He worked as an M&A and hi-tech lawyer in the law firm Ephraim Abramson & Co in Jerusalem (2007-2011). He is a guest lecturer on cyber security and international law at the Interdisciplinary Center of Herzlyia. He holds an LLB and LLM from the University of Montreal.

Max Smeets is a cybersecurity postdoctoral fellow at Stanford University Center for International Security and Cooperation (CISAC). He is also a non-resident cybersecurity policy fellow at New America. Max was awarded the annual 2018 Amos Perlmutter Prize of the Journal of Strategic Studies for the most outstanding manuscript submitted for publication by a junior faculty member. He was previously a College Lecturer in Politics at Keble College, University of Oxford, and has held research positions at Oxford Cyber Studies Programme, Columbia University SIPA, and Sciences Po CERI. He holds an undergraduate degree from University College Roosevelt, Utrecht University, and an M.Phil (Brasenose College) and a DPhil (St. John's College) in International Relations from the University of Oxford.

Peter Stockburger is a senior managing associate with Dentons, the world's largest law firm. He focuses his practice on cybersecurity, data protection, and complex commercial litigation. A frequent author and speaker, Peter regularly advises clients on a broad range of cutting-edge legal issues, including cybersecurity attribution, privacy litigation, and foreign sovereign immunity. He has published in the area of cyber normative development, including on cyber attribution, and has been recognized as a "Rising Star" by Super Lawyers Magazine since 2015. Peter also serves as an adjunct professor at the University of San Diego School of Law where he teaches in the area of public international law and oral advocacy.

Martin Strohmeier is a post-doctoral researcher in systems security in the Department of Computer Science at the University of Oxford. His main research interests are in the area of wireless security and critical infrastructure protection. During his PhD at Oxford, he has extensively analysed the security and privacy of wireless aviation technologies and his work focuses on developing cyber-physical approaches which

can improve the security of air traffic control quickly and efficiently, for which he has received several awards from the aviation and computer security communities. He is also a co-founder of the aviation research network, OpenSky. Before coming to Oxford in 2012, he received his MSc degree from TU Kaiserslautern, Germany and joined Lancaster University's InfoLab21 and Lufthansa AG as a visiting researcher.

Dr. **Christopher Whyte** is an Assistant Professor in the program on Homeland Security & Emergency Preparedness at the L. Douglas Wilder School of Government & Public Affairs, Virginia Commonwealth University. He teaches coursework on cyber security policy, conflict, and law, and has broadly taught coursework on international security topics, political risk analysis, and strategic planning. His research interests include a range of international security topics related to information warfare and political communication, cybersecurity doctrine and policy, and artificial intelligence.