Published by CCDCOE

2019

11th International Conference on Cyber Conflict: Silent Battle

T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)

CYCON

IEEE
Advancing Technology for Humanity

28 May – 31 May 2019, Tallinn, Estonia

2019
11TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: SILENT BATTLE

IEEE Catalog Number: CFP1926N-PRT
ISBN (print): 978-9949-9904-4-3
ISBN (pdf): 978-9949-9904-5-0

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

LEGAL NOTICE: This publication contains the opinions of the respective authors only. They do not necessarily reflect the policy or the opinion of NATO CCD COE, NATO, or any agency or any government. NATO CCD COE may not be held responsible for any loss or harm arising from the use of information contained in this book and is not responsible for the content of the external sources, including external websites referenced in this publication.

NATO COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE

The NATO Cooperative Cyber Defence Centre of Excellence (CCD COE) is a NATO-accredited knowledge hub in Tallinn, Estonia, offering a unique interdisciplinary approach to the most relevant issues in cyber defence. We conduct research, trainings and exercises in four core areas: technology, strategy, operations and law. The heart of the Centre is a diverse group of international experts from military, government, academia and industry, representing currently 21 member nations. Almost half as many nations are aspiring to become member in the years to come.

NATO CCD COE's mission is to support its member nations and NATO in the fields of cyber defence research, training and exercises. The Centre provides cyber defence expertise in the fields of technology, strategy, operations and law, often in an interdisciplinary manner. NATO CCD COE embodies and fosters the cooperation of like-minded nations in cyber defence. Our member nations are allies in NATO and like-minded partners beyond the Alliance.

Research Areas

Among other research areas, NATO CCD COE experts are currently involved in the analysis of autonomous features of cyber operations, digital forensics, protection of critical infrastructure, Cyber Command and Control, cyber deterrence, cyber effects in battlefield and attribution.

NATO CCD COE is home of the Tallinn Manual 2.0, the most comprehensive guide for policy advisors and legal experts on how International Law applies to cyber operations carried out between and against states and state actors. An invaluable analysis by an international group of renowned scholars published in 2017, continues to inspire both academic research and state practice.

Most of the Centre's publications and research papers are available online on the Centre's website www.ccdcoe.org, similarly to a database of national cyber security strategies, the International Cyber Developments Review and comprehensive overviews of national cyber security organizations.

Education and Training

NATO CCD COE promotes lifelong learning in cyber security. Our training courses are based on our latest research and cyber defence exercises. To best meet the training requirements of our Allies, Partners and NATO as a whole, we provide courses in different formats and locations, covering a broad range of topics in the technical, legal, strategic and operational cyber security domains.

NATO CCD COE is responsible for identifying and coordinating education and training solutions in cyber defence for all NATO bodies across the Alliance. NATO Allied Command Transformation has provided NATO CCD COE with an unconditional quality assurance accreditation for its contribution to high-quality NATO Education and Training.

CYCON 2019 SPONSORS

TECHNICAL SPONSOR

IEEE
Advancing Technology
for Humanity

DIAMOND SPONSORS

Microsoft

guardtime

SILVER SPONSORS

thinklogical®

A BELDEN BRAND

EclecticIQ

TABLE OF CONTENTS

INTRODUCTION

The annual International Conference on Cyber Conflict (CyCon) is entering its second decade. CyCon 2019 is the 11th iteration of the conference, organised by the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE) and taking place in Tallinn from 28 to 31 May 2019. Since 2009, CyCon has become a landmark conference addressing cyber conflict and security from the perspectives of technology, strategy, operations, law and policy. It brings together a trusted circle of decision-makers and experts representing government, the military, academia and the private sector. With the launch of the annual CyCon US conference series in cooperation with the US Army Cyber Institute in 2016, CyCon also serves as a transatlantic forum for the community of interest to discuss the most pressing issues of the cyber domain twice a year. The debates and presentations at CyCon stem both from original research papers submitted by the academic community and from insights offered by renowned experts in the field.

The core topic of CyCon 2019 is ‘Silent Battle’. The topic is broad enough to allow for diverse interpretations: a techie will think in terms of vulnerabilities, exploits and patches; a policy advisor could approach it as detection and attribution; a lawyer may interpret it through the lens of responsibility; and the military may approach it in terms of situational awareness. However, the underlying concern remains: the community of like-minded democracies is, more than ever before, being challenged by threats from cyberspace. How best can we cope with those challenges to our national security, from a strategic perspective? Where is the equilibrium in a silent battle and how can we cope with it? How can AI, machine learning and big data help us? How will international law develop in light of the serious effects of state-sponsored operations that may or may not be hard to attribute? These and many other questions will shape the interdisciplinary discussions of CyCon 2019.

The Call for Papers in June 2018 resulted in 111 abstracts submitted by October. After careful selection and peer review by the Academic Review Committee, 29 articles were selected for publication in this book and their authors were invited to present at the conference.

The papers for the strategic track of the conference are the most numerous, reaching a total of 12. **Martin C. Libicki** discusses collective defence in cyberspace and the idea of establishing a Baltic-area cyberspace alliance, considers what such an alliance would do, assesses its costs and benefits for its members, and considers its implications for NATO and for the United States. Using multiple case studies, **Keir Giles** and **Kim Hartmann** explain the recent shift towards a more transparent policy on cyber conflicts and its future implications for numerous nations and NATO. As the first step of their multiphase research, **Daniel Kapellmann** and **Rhyner Washburn** investigate various information-sharing platform designs for streamlining the exchange of knowledge, discussion and management of ICS vulnerabilities, a topic that possibly has not been sufficiently in focus to date. **Barış Egemen Özkan** and **Serol Bulkan** show how, besides hardware, commercial-off-the-shelf software obsolescence leads to major vulnerabilities for nations in cyberspace, especially with regard to critical infrastructure and military systems, and offer possible mitigations. **Bilyana Lilly**, **Quentin Hodgson**, **Lillian Ablon** and **Adam Moore** propose a high-level practical approach to the cyber Indications and Warnings (I&W) concept by examining a set of I&W frameworks to effectively anticipate and defend against cyber threats. **Erwin Orye** and **Olaf Maennel** consider how to predict and measure the outcome of cyber effects and recommend a set of best practices for enhancing cyber effects in modern warfare.

Jason Healey and **Neil Jenkins** outline a methodology and metrics for the recent counteroffensive cyber operations policy of the United States in terms of its deterrent impact. **Ji Young Kong**, **Kyoung Gon** and **Jong In Lim** describe the versatility of North Korea's many cyber operation state actors and suggest strategies to cope with them. **Max Smeets** provides an empirical analysis of the existing military cyber organisations of allied nations and offers solutions to address similar key organisational challenges among them. **Brad Bigelow** sets out a set of equivalent principles that could be applied to military cyberspace operations performed below the level of armed conflict, and assesses the functions to designate the role for the military. **Gil Baram** and **Udi Sommer** show why, and under which geopolitical circumstances, countries choose to give up the advantages of anonymity after experiencing cyber attacks. **James Pavur** and **Ivan Martinovic** present a strategic analysis of the impact of cyberspace on key stabilising factors and the threat posed to space's longstanding stability by cyber Anti-Satellite Weapons.

Three articles focus on the operational aspects of cyber defence. **Alicia Bargar**, **Janis Butkevics**, **Stephanie Pitts** and **Ian McCulloh** propose the use of social network analysis (SNA) to bolster the identification of false narratives used during information operations on social media. **Joe Burton** and **Simona R. Soare** explain the strategic implications of the weaponisation of AI for international security. **Robert Koch** highlights the potential risks to military operations coming from the Dark Web, and proposes ways to mitigate these risks.

There are five articles with a legal bent. **Kenneth Kraszewski** describes the SamSam ransomware attack on Atlanta in early 2018 and provides an analysis of the possible legal responses available to the United States. **Jeff Kosseff** analyses the United States' new operational concept to 'defend forward' and investigates the possible options available to the US within the limits imposed by existing international law. **Nikolas Ott** and **Anna-Maria Osula** examine the increasingly important role that regional organisations play in stabilising states' relationships in cyberspace and elaborate on their possible synergies with the UN efforts. **Przemysław Roguski** investigates the factors challenging the concept of sovereignty in cyberspace and proposes a different understanding of this foundational principle of international law, through a model of 'layered sovereignty'. **Barrie Sander** observes states' reluctance to agree on cyber-specific multilateral treaties and to publicly clarify the customary international rules applicable to hostile cyber operations, and suggests that the silence of states can be interpreted according to the different types of security threats they are facing.

Turning to the technical arena, **Giovanni Apruzzese**, **Michele Colajanni**, **Luca Ferretti** and **Mirco Marchetti** shed light on adversarial attacks that aim to affect the detection and prediction capabilities of machine-learning models. **Nicolas Känzig**, **Roland Meier**, **Luca Gambazzi**, **Vincent Lenders** and **Laurent Vanbever** present a system that quickly and reliably identifies command and control channels without prior network knowledge. **Roman Graf**, **Ross King** and **Aaron Kaplan** offer effective identifying and defeating methodologies for malware applications in Android smartphones. **Joonsoo Kim**, **Kyungho Kim** and **Moonsu Jang** discuss the design and construction of a universal cyber-physical platform through the final design choices. **Giuseppina Murino**, **Alessandro Armando** and **Armando Tacchella** seek a model-free, quantitative, and general-purpose evaluation methodology to extract resilience indexes from system logs and process data.

Pierre Dumont, **Roland Meier**, **David Gugelmann** and **Vincent Lenders** tackle the problem of detecting malicious shell sessions based on session logs, by analysing the sequence of commands that the shell users executed. **Martin Strohmeier**, **Matthias Schäfer**, **Marc Liechti**, **Markus Fuchs**, **Markus Engel** and **Vincent Lenders** analyse and discuss the challenges related to information gathering in the Dark Web for cyber security intelligence purposes. **Artūrs Lavrenovs** introduces a methodology for measuring different properties of individual devices participating in distributed denial-of-service (DDoS) attacks. Finally, **Robert Koch** and **Mario Golling** analyse the characteristics of silent battles and hidden cyber attacks and summarise the current and expected developments.

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

Academic Review Committee Members for CyCon 2019:

- Siim Alatalu, NATO CCD COE
- Geert Alberghs, Ministry of Defence, Belgium
- Maj Vincent Banse, NATO CCD COE
- Donara Barojan, NATO StratCom COE; DFR Lab
- Henrik Beckvard, NATO CCD COE
- Prof Col Daniel Bennett, Army Cyber Institute, United States
- Cdr Stefano Biondi, NATO CCD COE
- Bernhards Blumbergs, CERT Latvia, NATO CCD COE Ambassador
- Maj Pascal Brangetto, French Ministry of Defence
- Prof Thomas Chen, City, University of London, United Kingdom
- Prof Michele Colajanni, University of Modena and Reggio Emilia, Italy
- Torsten Corall, NATO CCD COE
- Samuele De Tomas Colatin, NATO CCD COE
- Dr Thibault Debatty, Royal Military Academy, Belgium
- Prof Dorothy E. Denning, Naval Postgraduate School, United States
- Dr Kenneth Geers, NATO CCD COE Ambassador;
  Atlantic Council Senior Fellow
- Keir Giles, Chatham House, Conflict Studies Research Centre,
  United Kingdom
- Rudi Gouweleeuw, Netherlands Organisation for Applied Scientific Research
- Prof Michael R. Grimaila, Air Force Institute of Technology, United States
- Dr Jonas Hallberg, Swedish Defence Research Agency
- Dr Jakub Harašta, Masaryk University, Czech Republic
- Prof David Hutchison, Lancaster University, United Kingdom
- Dr Ji-Jen Hwang, University of California, Berkeley
- Prof Gabriel Jakobson, CyberGem Consulting
- Cpt Raik Jakschis, Bundeswehr Cyber Security Centre
- Taťána Jančárková, National Cyber and Information Security Agency,
  Czech Republic
- Maj Harry Kantola, Finnish Defence Forces
- Kadri Kaska, NATO CCD COE
- Prof Sokratis K. Katsikas, Norwegian University of Science and Technology
- Dr Panagiotis Kikiras, European Defence Agency
- Markus Kont, NATO CCD COE
- Dr Csaba Krasznay, National University of Public Service, Hungary
- Clare Lain, NATO CCD COE
- LtCol Franz Lantenhammer, NATO CCD COE
- Dr Scott Lathrop, US Army
- Artūrs Lavrenovs, NATO CCD COE

- Prof Sean Lawson, University of Utah
- Dr Lauri Lindström, NATO CCD COE
- Prof Olaf Maennel, Tallinn University of Technology, Estonia
- Dr Kubo Mačák, University of Exeter, United Kingdom
- Merle Maigre, CybExer Technologies
- Dr Matti Mantere, Forcepoint
- Prof Evangelos Markatos, University of Crete, Institute of Computer Science, Greece
- Prof Paul Maxwell, Army Cyber Institute; United States Military Academy
- Maj Markus Maybaum, Bundeswehr Cyber Security Centre; NATO CCD COE Ambassador; Fraunhofer FKIE
- Prof Michael Meier, University of Bonn, Fraunhofer FKIE
- Andrea Melegari, Expert System
- Tomáš Minárik, NATO CCD COE
- Dr Anna Molnár, National University of Public Service, Hungary
- Maarja Naagel, NATO CCD COE
- Dr Jose Nazario, Censys
- Dr Lars Nicander, Swedish Defence University
- Maj Erwin Orye, NATO CCD COE
- Dr Anna-Maria Osula, Guardtime; Tallinn University of Technology, Estonia
- Nikolas Ott, Organisation for Security and Co-operation in Europe
- Dr Piroska Páll-Orosz, Ministry of Defence, Hungary
- 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, Bolt
- Prof Gabi Dreo Rodosek, Bundeswehr University Munich, Germany
- Henry Rõigas, Guardtime
- Prof Juha Röning, University of Oulu, Finland
- Sven Sakkov, International Centre for Defence and Security, Estonia
- Ragnhild Siedler, Norwegian Defence Research Establishment
- LtCol Dr Massimiliano Signoretti, NATO CCD COE
- Dr Max Smeets, Stanford University, Centre for International Security and Cooperation, 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
- Morta Strazdaitė, Paris School of International Affairs, France
- Dr Michail Sulmeyer, Harvard Kennedy School, United States
- Dr Thierry Tardy, NATO Defence College, Rome, Italy

- LtCol José Teixeira, NATO CCD COE
- Lt Ihsan Tolga, NATO CCD COE
- Dr Jens Tölle, Fraunhofer FKIE, Germany
- Dr Lauri Tuovinen, University of Oulu
- Prof Risto Vaarandi, Tallinn University of Technology, Estonia
- Ann Väljataga, NATO CCD COE
- Matthijs Veenendaal, Ministry of Defence (Defence Cyber Command), the Netherlands
- Dr Adrian Venables, Tallinn University of Technology, Estonia
- Prof Ari Visa, Tampere University of Technology, Finland
- Maj Gábor Visky, NATO CCD COE
- Prof Col David Wallace, United States Military Academy
- Prof Bruce W. Watson, Stellenbosch University, South Africa; IP Blox
- 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 for this publication. In addition, we highly appreciate the NATO CCD COE staff for their great organising skills and assistance during the publication process.

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

The CyCon 2019 Programme Committee

Siim Alatalu
Cdr Stefano Biondi
Tomáš Minárik
LtCol Dr Massimiliano Signoretti
Lt Ihsan Tolga
Maj Gábor Visky

NATO Cooperative Cyber Defence
Centre of Excellence
Tallinn, Estonia, May 2019

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

For a Baltic Cyberspace Alliance?

Martin Libicki [1]
Visiting Professor, Cyber Science Department
United States Naval Academy
Annapolis, MD
libicki@usna.edu

Abstract: In NATO, an attack on one is an attack on all. In recent years, this tenet has been extended to mean that a cyberattack on one is a cyberattack on all. But does what makes sense in the physical world also make sense if extended into cyberspace? And if there is virtue in collective cyberspace defense, is NATO necessarily the right grouping – in a world where, as far as the United States and the United Kingdom are concerned, more of what constitutes cyber defense circulates within the Five Eyes coalition rather than within NATO? To explore these issues, this essay moots the creation of a Baltic-area cyberspace alliance, considers what it would do, assesses its costs and benefits for its members, and concludes by considering whether such an alliance would be also be in the interest of the U.S. Keys to this discussion are (1) the distinction between what constitutes an "attack" in a medium where occupation may result and actions in media where occupation is (currently) meaningless and effects almost always reversible, (2) what collective defense should mean in cyberspace – and where responsibilities may be best discharged within the mix of hardness, pre-emption, and deterrence that constitute defense, (3) the relationship between cyberspace defense and information warfare defense, and (4) the relevance to alliance formation of the fact that while war is dull, dirty, and dangerous, cyber war is none of these three.

Keywords: *cyber defense, alliances, NATO*

[1] The opinions expressed in this paper are only those of the author, and do not represent the US Naval Academy, Department of Navy, or Department of Defense.

1. INTRODUCTION

Normally, countries do not benefit if their friends go off and form an alliance without them. But cyberspace may be different. The doctrines and arrangements that work in the physical world cannot be transported into the virtual world without asking whether the assumptions that hold in the physical world also apply to the virtual world. This helps to determine if enough tenets remain valid to justify adopting and adapting such arrangements for a new domain, or instead, starting over.

To this end, we moot an alliance of selected European states whose mission is mutual defense against cyberspace threats: operations that degrade, disrupt, corrupt, or destroy information systems, but might also include – largely because of interest in such things in the region – the use of cyber espionage to directly harm another party's interests: e.g., the DNC hack, or similar subsequent intrusions in France and Germany. Selected states, here, include countries that border the Baltic Sea (except, of course, Russia), perhaps with Norway and the Netherlands thrown in for good measure. In such an alliance, the key country would be Germany, but the inclusion of Sweden and Finland means that it would not be a subset of today's NATO. Throughout, we assume the continued existence of NATO as a traditional defense alliance and the continued working of the intelligence-sharing arrangements among the Five Eyes (the U.S., the UK, Canada, Australia, and New Zealand).

In laying out the case for a Baltic cyberspace alliance – a case that should promote both European and U.S. interests – this essay proceeds in several steps. First, we examine the elements of a cyberspace alliance: what tasks it fulfills and whether an explicit alliance is necessary or even helpful to carry out this or that task. Second, we adduce some benefits (again, from both the European and U.S. perspective) to the formation of such an alliance. Third, we discuss issues created by such an alliance.

2. COLLECTIVE DEFENSE IN CYBERSPACE

Central to NATO is the premise that an attack on one is an attack on all: each member is obliged to treat an attack on another member as if it were an attack on itself. Traditionally, the response to an attack was straightforward: a state of war is acknowledged; participating armies defend sovereign territory, attempt to disarm the other side, and have it sue for terms. In the Cold War, the collective strength of NATO's members was used to maintain a front against Soviet invaders, whilst the United States used the threat of a nuclear response to deter Soviet incursions into Europe (or to limit the depth and duration of such an incursion). Although the purpose of an alliance is trickier in domains where aggressors cannot occupy territory

– notably the seas and the air – broad notions such as strength from combining forces and the direct support that such forces can provide to the ground campaign make it straightforward to extend NATO into those two domains.

Conflict in cyberspace is of a different nature, but the nature depends on whether it is tactical or strategic. Tactical cyber war is what supports a kinetic military campaign; it may be a valuable niche capability, but it is the outcome on the ground that matters. A Baltic cyberspace alliance,[2] not being a kinetic military alliance (especially if it included non-NATO countries), would not play in that contest; only individual countries or NATO, collectively, would. If NATO were involved in a real shooting war, tactical and even strategic cyberspace operations could come under NATO's aegis.

Strategic cyber war,[3] by contrast, stands apart from kinetic conflict and may even take place in its absence.[4] One purpose could be to pressure other countries by imposing costs on them; another, conversely, may be retaliation to enable or reinforce deterrence. As a lesser included case, it may be used to enhance influence operations on countries, political groups, or individuals.

It is strategic cyber war for which a Baltic cyberspace alliance might prove useful.

3. NATO AS A CYBERSPACE ALLIANCE

The logic of international alliance is that bigger is usually better when defending against threats.[5] Although the size of the alliance and the need for consensus can complicate warfighting,[6] the concept of a common defense means that an adversary faces the combined militaries of multiple countries. In a world in which attackers are dissuaded by the prospect that united countries will interpose their forces between attackers and those being defended, the premise that more is better makes intuitive sense.

[2] The Council of Baltic Sea States, formed after the Soviet Union dissolved, is an intergovernmental organization that works on social, economic, legal, and environmental issues. More recently, Baltic states came together to address currency issues; see "Northern member states unite on euro-zone reform," December 8, 2018; https://www.economist.com/europe/2018/12/08/northern-member-states-unite-on-euro-zone-reform.

[3] As defined and used by the author in his *Cyberdeterrence and Cyberwar*, Santa Monica, CA (RAND), 2009 p. 117 -138. See also Tomas Rid, "Cyberwar Will Not Take Place," *Journal of Strategic Studies, 35:1* (2012), pp. 5-32.,

[4] Although a cyberattack *can* cause physical damage, it can be considered one more way that cyberattacks can impose costs on societies without necessarily being part of an armed conflict.

[5] "Usually" because adding members means getting drawn into more disputes, often in countries that are farther from the alliance's core and therefore harder to physically defend.

[6] Good examples can be drawn from the difficulties encountered in Operation Allied Force – in which NATO, incidentally, prevailed; see General Wesley Clark, *Waging Modern War: Bosnia, Kosovo, and the Future of Combat*, New York NY (Public Affairs), 2001.

But combat in cyberspace does not really benefit from economies of scale. Within a country, adding more offensive cyber warriors often means lowering the qualifications (at least initially, and perhaps in the long term) for what is, by nature, an inherently elite profession. This means not only that diminishing returns set in, but that the activities of the good-but-not-great can well tip off the other side. So tipped, the other side can improve its defenses in ways that are specific (e.g., vulnerabilities are patched after having been discovered) and general (i.e., a shift occurs in the tradeoff between security and cost/convenience). Correspondingly, the contribution of additional operators is limited. When these operators come from other countries, their contribution is further vitiated unless these countries operate seamlessly. Within an intelligence-sharing alliance (e.g., the Five Eyes), additional members do add heft (each country, for instance, can employ relationships that it has developed with communications companies around the world). Once seams intrude – and these seams are larger within NATO than within the Five Eyes – the level of coordination is less and the prospects for interference (e.g., two countries seeking access to the same target system) are greater.

When it comes to defense, the arguments that vitiate the benefits mass are different but similar. A large percentage of all cyber defense efforts requires looking after specific systems. Adding allies adds more systems to defend. This hardly helps defenders of existing systems.[7] Although there are defense activities where adding countries may help – e.g., intelligence fusion, collective learning, and forensics – none of these really requires a military alliance, and some of the contributors for these three efforts live in the private realm.

Alliances also express their weight through deterrence policies. In the Cold War, the United States deterred attacks by the Soviet Union on NATO allies with a nuclear threat: certainly, if the attacks involved nuclear weapons, and quite possibly if the attack involved overwhelming conventional force. In recent years the U.S. deterrence policy in cyberspace has been notionally extended to a NATO deterrence policy. Unfortunately, U.S. deterrence policy for attacks on the homeland is already an uncertain thing – and extending it adds further uncertainties. There are, for instance, serious questions about what constitutes a cyberattack serious enough to merit retaliation and what the form of retaliation would be; the United States has used sanctions in response to malign cyberspace activities, but there is little evidence that sanctions have deterred Russia, Iran, or North Korea.[8] NATO's retaliation capabilities – which are largely US retaliation capabilities (plus some UK capabilities) – are even less likely to be brought into play if the target of a cyberattack were European. In the

[7] A counter-argument is that larger alliances give foes more targets, forcing them to spread their efforts and thereby relieving defenders of existing systems. But that assumes that (1) all defense efforts counter those foes against whom the alliance is established, and (2) that new allies were not already under attack from such foes.

[8] In 2015, China agreed to cease its commercially-oriented cyber espionage under the *threat* of sanctions, but that was more like coercive diplomacy. The agreement unraveled by early 2017 as China perceived that it would be sanctioned over broader trade issues – and so it might as well spy.

five years prior to North Korea's hack of Sony – which the United States did respond to – South Korea suffered far worse depredations with no U.S. response.[9]

4. AND WHY IS THE NATO GROUPING NOT THE OBVIOUS ONE?

Yet Europeans lean on NATO; in large part, because it already exists and therefore does not need to be invented. But NATO is a military alliance, while cyberspace is essentially a conduit for information,[10] hence generally dominated by the community that deals with information *qua* intelligence. And the existence of the Five Eyes coalition only underlines this point: working relationships among that coalition in the information domain are tighter than they are in the information domain across the NATO alliance. And two of the Five Eyes are not even in NATO. In other words, the real coalition is not doing Europe (the UK excepted) in particular that much good. This matters, because the primary cyber war threat to Europe is from Russia, and the primary targets of Russian coercion are in European countries that face Russia. Two of the countries that face the gravest threats are not even in NATO.

Correspondingly, a Baltic cyberspace alliance would have a limited ambit: members would cooperate on defense and aver that a cyberattack on one is an attack on all. In practice, were such an arrangement made, the alliance would have its own definition of what constitutes an "attack"; it might include social media manipulation.

A cyberspace alliance, as such, would have three facets but lack one. The first facet is defense: each country would putatively participate more vigorously in those cyber defense activities that benefit from scale, as noted: threat intelligence, forensics, lessons learned. With very large cyberattacks, they could offer mutual aid for systems restoration. The second facet is defense by deterrence. It would require a consensus on what constitutes an actionable offense in cyberspace, notably the type and severity; what responses are appropriate (e.g., to an attack on the local power grid); and what kind of capability is required to retaliate in cyberspace. One rationale for responding in cyberspace is that less forceful options, such as alliance-wide sanctions, are likely to be even weaker than U.S. sanctions are. Conversely, threatening kinetic responses – given the lack of nuclear weapons among proposed alliance members – lacks credibility thanks to Russia's escalation dominance. The third facet would consist of offensive cyberspace operations used for coercive or, more likely, counter-coercive purposes or

[9] Iain Thomson, "South Korea faces $1bn bill after hackers raid national ID database," The Register, October 14, 2014, http://www.theregister.co.uk/2014/10/14/south_korea_national_identity_system_hacked/.

[10] Notwithstanding that some of this information (e.g., rogue machine instructions) may result in physical damage.

for retaliation against grave non-cyber offenses (e.g., the Skripal poisonings[11]). By contrast, cyberspace operations in support of kinetic operations (e.g., taking a SAM site offline while a NATO sortie flies overhead) would fall outside such an alliance because kinetic operations fall to NATO; if individual members carried out tactical cyberattacks, they would fall under NATO auspices (or their own fights).

5. CHARACTERISTICS AND ADVANTAGES OF A BALTIC CYBERSPACE ALLIANCE

What are the characteristics and advantages of such an alliance to its member countries?

First, the alliance, limited to cyberspace, would invariably focus on Russia, despite having to tend to other threats (e.g., from China's commercially-motivated cyberespionage) and despite the possibility that alliance members would probably be diplomatic in public about the alliance's purpose. Russia's cyberspace threats are malevolent, politically-directed, and often part of a larger campaign to sow disorder and facilitate coercion. NATO countries, as a whole, are not entirely focused on Russia, these days: those in North America pay as much attention to China;[12] those near the Mediterranean tend to look southward.

Second, such an alliance would include currently neutral countries, notably Sweden and Finland. Both countries punch above their weight, in cyberspace operations[13] and information operations[14] respectively. This raises the question: why don't such countries just enter NATO? To be sure, roughly half the citizens in both countries would like to – but the other half fear, justifiably, a neuralgic Russian reaction if they did (Finland's accession could put troops along miles of Russian borders). Although Russians would likely react badly to the formation of a Baltic cyberspace alliance, they would have a more difficult time summoning images of jackbooted soldiers while doing so. For Sweden and Finland, the cyberspace alliance could serve as a halfway house. If the Russian threat eases, their entry into NATO can be indefinitely postponed (in the unlikely event that the Russian cyberspace threat disappears, they can leave or the cyberspace alliance might wither away). If the Russian threat persists

[11] See, for instance, Larisa Brown, "Theresa May could order a cyber attack against Russia in retaliation for the nerve-agent strike as part of a secret package of measures to hurt Putin," March 12, 2018; http://www.dailymail.co.uk/news/article-5492835/Theresa-order-cyberattack-against-Russia.html.

[12] Just one small sample: Ryan Browne, "New acting secretary of defense tells Pentagon 'to remember China, China, China'," January 2, 2019;https://www.cnn.com/2019/01/02/politics/shanahan-pentagon-first-day-china/index.html.

[13] See, for instance, Hugh Eakin, "The Swedish Kings of Cyberwar," https://www.nybooks.com/articles/2017/01/19/the-swedish-kings-of-cyberwar/, January 13, 2017.

[14] See, for instance, Reid Standish, "Why Is Finland Able to Fend Off Putin's Information War? Helsinki has emerged as a resilient front against Kremlin spin. But can its successes be translated to the rest of Europe?" March 1, 2017; https://foreignpolicy.com/2017/03/01/why-is-finland-able-to-fend-off-putins-information-war/.

or worsens, these two countries will have had more practice interoperating with NATO countries, thereby easing their way into an alliance that spans the conventional domains of warfare.

Third, a cyberspace alliance would be a mechanism to get Germany to become more involved – or, more to the point, take leadership – in defending Europe against Russia. The current contribution of German military spending (1.2 percent of GDP) to the common defense of NATO is modest. Germany, nevertheless, remains Europe's largest economy, and would constitute roughly half of the weight of any Baltic cyberspace alliance. Germany has also stepped up smartly in developing its Cyber and Information Domain Service. Its manning, if plans hold,[15] would constitute 7.5 percent of Germany's total force level (13,500 out of 180,000);[16] its spending would be 6.3 percent of Germany's military (41.5 billion Euro) budget.[17] By way of contrast, USCYBERCOM's end-strength goal of 6,000 compares to 1.3 million military personnel in the overall U.S. military – less than a tenth as much concentration on cyberspace. Granted, this is not an apples-to-apples comparison: Germany's end-strength includes electronic warfare battalions; USCYBERCOM's end-strength does not. But, even after adjustments, Germany's commitment to fighting in cyberspace, relative to its overall military strength, looks more substantial than the U.S. commitment. Furthermore, a German focus on cyberspace (vis-à-vis kinetic elements of military power) is, again, less likely to engender a neuralgic reaction from Russia (no jackboots, etc.) but does put Russia on notice that its maneuvers in cyberspace have not gone unnoticed and will be resisted by those best placed to resist them.

The advantage of such an alliance to its members is that they put the power of all in service of each. This should give Russians second thoughts about their use of cyberspace for offensive purposes – although it may also initially goad them into carrying out operations against the non-NATO countries (Sweden and Finland) to inhibit their participation in such an alliance. Russia's doing so, conversely, could very well reinforce the value to today's neutral countries of having others to lean on when facing Russia. The countries in this alliance would be self-selected by virtue of their concern over Russian activities in cyberspace. By contrast, a unified and meaningful NATO response to Russian provocations has to surmount the objections of countries that reserve some sympathy for Russia (Hungary and Greece come to mind).

[15] "Defence Minister Ursula von der Leyen revealed plans to recruit up to 13,500 cyber soldiers in addition to around 500 civilian workers capable of defending the military's electronic intelligence as part of the new Cyber and Information Space Command, according to Germany's *The Local*." From Tom O'Connor, "German Military Battles Foreign Hacking with New Cyber Soldiers," April 5, 2017; https://www.newsweek.com/german-military-launches-new-cyber-division-amid-russian-hacking-claims-579573. In the interim, many of its employees could be reservists, though; "Germany struggles to step up cyberdefense," August 7, 2018; https://www.dw.com/en/germany-struggles-to-step-up-cyberdefense/a-44979677.

[16] The UK's formation of a 2,000-person strong *cyber* force, within an armed force of 160,000 total members, also suggests a higher percentage commitment to cyberspace than in the United States; see David Bond, "Britain Preparing to Launch New Cyber Warfare Unit," Sep 21, 2018, https://www.ft.com/content/eef717f2-bb6e-11e8-8274-55b72926558f.

[17] Sumi Somaskanda, "Cyberattacks Are 'Ticking Time Bombs' for Germany," June 4, 2018; https://www.theatlantic.com/international/archive/2018/06/germany-cyberattacks/561914/.

Fourth, this would give NATO competition in the alliance business. Arguably, this would weaken NATO – and is thereby a disadvantage. But competition can also be good: it persuades competitors to listen to their clients (customers, audience, etc.) and induces them to innovate in order to retain their standing. Otherwise, secure in the knowledge that their position is unassailable, they risk becoming sluggish and unresponsive – and when they fall or come apart, it is often "first slowly and then all at once".[18] Thus, when offering cyber security or countering cyberattacks, relevant countries can ask the institutions of NATO and also those of the Baltic cyberspace alliance what each of them can do – each knowing that they are competing both against Russia's malign influence and the other's benign influence. But competition can also raise problems: an institutionally aggressive cyberspace alliance may seek greater influence by stretching the definition of a cyberattack: e.g., to include electronic warfare, interference with space operations, and sabotage of or attacks on information infrastructures.

6. ADVANTAGES FOR THE UNITED STATES

The most basic advantage is that it makes Europeans more responsible for their own defense, albeit in just this one domain. In the 1980s, for example, three neutral countries – Switzerland, Sweden, and Finland – spent far higher proportions of their income on national defense than most European NATO allies did.[19] The "free rider" problem is, if anything, worse today. It may be that much more difficult to persuade European countries to arm themselves if, when such arms have to be used, it would be under a war effort led by the United States. The return of Russia as an aggressive power, since roughly late 2013, may not have been internalized by European countries, concerned as they are with internal fissures – many of which, ironically, were deliberately exacerbated by Russia's information warfare campaign. And the U.S. pivot to Asia, while more advertised than practiced, would necessarily mean a shift in U.S. resources that would otherwise be available for Europe.

But in cyberspace, countries in a Baltic cyberspace alliance would be pooling their resources under either their own individual command (as befits an activity so highly linked to intelligence) or, at least under the command of Europeans. And with the United States not in such an alliance, there is much less of a "free rider" problem (even if Germany would be roughly half the alliance, countries such as Sweden, Finland, and Estonia punch above their weight in this domain). The downside of the upside is akin to the owner of a hammer being persuaded that every problem is a nail: if given a choice between responding to hostile actions in the kinetic world and responding in cyberspace, the latter may be seen as particularly attractive because it

[18] The quote is from Ernest Hemingway's *The Sun Also Rises*.

[19] Each of the two NATO countries that *did* invest heavily in national defense – Greece and Turkey – largely did so to keep the other at bay.

relies on tools the alliance can wield themselves rather than tools largely wielded by the United States.

Another advantage for the United States is that such an alliance may complicate Russia's cyber war efforts – largely by increasing the uncertainty that Russian efforts may be met with reprisals: the odds of retaliation from either the United States (as the premier cyberspace power of NATO) or from the Baltic cyberspace alliance will be higher than the odds of retaliation from each of them. This is particularly true for those cyberattacks that leave multiple victims: NotPetya, as an example, levied costs in the hundreds of millions of dollars from Merck and Federal Express (both U.S.-headquartered corporations) and from Maersk (headquartered in Denmark). The raised odds for a response may arise from meeting credibility thresholds (the United States may be wary and the Baltic cyberspace alliance less so or vice versa), attribution thresholds (the United States may have confidence and the Baltic cyberspace alliance may not or vice versa), and damage thresholds (the United States may recognize a higher threshold to warrant its retaliation if the effects of the cyberattack fall primarily on Europeans). Both the United States and the Baltic cyberspace alliance may retaliate but against different targets.[20]

An associated benefit is that if the *modus operandi* of whatever cyberspace operations ensue from NATO (that is, in practice, from the United States or the UK) and the Baltic cyberspace alliance are sufficiently similar, it may not be clear to Russia who struck back. This would complicate counter-retaliation targeting (and threats), in anticipation of which retaliation may be more likely, and the prospect thereof more credible. To be fair, attack-retaliation cycles in cyberspace remain loose: the closest example of a retaliatory cyberattack was the late 2012 DDOS campaign against U.S. banks by an Iran that had, two years earlier, discovered that its nuclear program had been set back by the Stuxnet worm. Attack-retaliation-counter-retaliation cycles are even more n^{th}-order relationships. Furthermore, Russia may have the SIGINT or HUMINT to make its own attribution – or it may not care and may conclude that the Baltic cyberspace alliance is an arm of NATO despite the former having neutral countries in it; indeed, it may see all opposing alliances as arms of the United States, facts notwithstanding.

Second-order considerations add complexities:

- Conceivably, neither NATO nor the Baltic cyberspace alliance retaliates in the belief that the other will – and that letting the other one do so may avoid counter-retaliation while gaining the benefit of deterrence (to wit, the "free rider" problem that affects NATO, writ large). Perhaps each side may interact with the other, but because cyberspace operations are so highly classified,

[20] Perhaps needless to add, if the United States has high confidence in attribution and high thresholds, and the Baltic cyberspace alliance has low confidence in attribution and low thresholds, neither may decide to retaliate.

each side may conclude that the failure to hear from their counterpart is no evidence that nothing is being planned – and so each assumes that the other is making plans.

- There could well be an exchange of intelligence between NATO and the Baltic cyberspace alliance that allows attribution evidence on Russian cyberattacks to strengthen the case over what each may conclude from its own efforts. That raises the question of why countries – the owners of intelligence services – do not simply cooperate within NATO to build that case. First, Sweden and Finland are not NATO members (although, as noted, Sweden shares information with the West). Second, a Baltic cyberspace alliance may well induce member countries to raise their cyber intelligence game (because they are helping themselves rather than others) giving them more to contribute.
- Just as having two independent sources of threats complicates Russia's calculus, it can also complicate the assurance component, wherein others are assured that small attacks will be treated less harshly than large attacks lest others lose any reason to moderate their bad behavior (colloquially: "in for a dime, in for a dollar"). An early version of this logic explains why Robert McNamara (the 1960s-era U.S. Secretary of Defense) was unhappy with France's nuclear capabilities and ambitions. If nuclear war ensued, he wanted the option of using nuclear weapons first against the nuclear systems of the USSR but not targeting cities specifically – and then threatening that if the USSR did strike Western cities, the United States, in turn, would target Soviet cities. But France's nuclear deterrent was too small to be used against Soviet nuclear installations exclusively – it was meant solely as a deterrent. Thus, if France used its nuclear weapons against Soviet cities, there would be little reason for the USSR to avoid hitting U.S. cities – even if the United States did not initially target Soviet cities. Again, the imprecision, loose coupling, and ambiguity associated with cyberspace operations may make this consideration notional for the time being.

7. THE RISKS OF ENTANGLEMENT

A classic problem of the politics of an alliance is the scope that it gives members, particularly the smaller ones, to get partners wrapped up in their fights. Take WWI: what was initially an Austrian-Serbian fight became a Russian-Austrian-Serbian fight, and then a German-Russian-Austrian-Serbian fight before morphing into a Franco-German-Russian-Austrian-Serbian fight. With two alliances, one specific to a particular domain, the complexities quickly mount (even ignoring a highly unlikely cyberattack by a NATO member outside the Baltic cyberspace alliance on a non-NATO member inside the alliance).

One entangled path may arise from a retaliatory cyberattack by a member of the Baltic cyberspace alliance which yields a kinetic retaliation. If this kinetic retaliation is considered an attack, and if the target state is a member of NATO, then an Article V issue arises; if NATO members agree, what was started by a non-member of NATO in cyberspace could descend into a kinetic conflict between NATO and Russia. Granted, NATO does not have to respond; it may argue that it played no role in the initial fracas – but a failure to invoke Article V under circumstances that would call for doing so would harm NATO's credibility as an alliance.

Another path is the conflation of information warfare with its subset,[21] cyber warfare – coupled with the latter's conflation with kinetic war. Somewhere on the spectrum between mischievous speech and Armageddon, every alliance needs to draw some line between acceptable and unacceptable practice. Otherwise, a country's (admittedly malign) attempts to manipulate social media messaging will start other countries down a slippery slope. Of this, several observations. First, a hard line on freedom of speech (and press) is baked into the U.S. Constitution. Europe is more apt to weigh such freedoms against community values (e.g., prohibiting hate speech and enhancing privacy and data protection). When mobilizing to "counterattack" unwanted expression, a U.S.-dominated NATO may be more reluctant than a Baltic cyberspace alliance (although the greater U.S. bent towards action could balance this out). Second, all this potential conflation suggests the need for any such alliance posture to make distinctions among levels of conflict – separating influence operations from cyberattack; cyberattack from kinetic attack; and conventional kinetic attack from nuclear attack. To proclaim that "an attack on one is an attack on all" without defining "attack" draws no such distinctions. Such levels may have to be crossed – the hypothetical cyberattack that kills thousands may outrank an exchange of naval gunfire at sea, for instance – but crossing should be a deliberate act; one, moreover, that reflects alliance consensus. There also needs to be room for some intra-war deterrence so that Russia does not heedlessly escalate from one level of hostility to another.

A third path leads from intelligence to operations. The two influence one another in all media, but the relationship is particularly close in cyberspace – where a penetration made for one purpose can be used for another and where a successful and persistent penetration is often the major part of any such operation. Problems may arise because friends spy on one another. In one infamous example, the NSA reportedly tapped the private phone used by Germany's Chancellor.[22] Although systems that are targets for cyber espionage are often implausible places to start a cyberattack from, exceptions

[21] As the author argues in "The Convergence of Information Warfare," *Strategic Studies Quarterly*, Spring 2017, 49-65.

[22] Melissa Eddy, "File Is Said to Confirm N.S.A. Spied on Merkel," July 1, 2015; https://www.nytimes.com/2015/07/02/world/europe/file-is-said-to-confirm-nsa-spied-on-merkel.html.

exist – and when malware is found, the target may be persuaded to overlook such distinctions to draw implausible conclusions. A Baltic cyberspace alliance may provide a mechanism for countries to develop a consensus on how aggressive cyber espionage could become without triggering Russia to retaliate out of fear of a pending cyberattack.

A fourth path arises from trying to distinguish between tactical cyberattacks carried out to support kinetic operations, and thus under NATO auspices, and strategic cyberattacks that could come under the auspices of a Baltic cyberspace alliance. Presumably, because the usefulness of tactical cyberattacks in the absence of kinetic conflict is minimal (because disruption, unlike destruction, can be reversed in short order), if there is no kinetic conflict at hand or on the horizon, there is no tactical cyberattack – everything else therefore is of a strategic nature and hence could come under the aegis of the Baltic cyberspace alliance. But the tactical-strategic divide is not at all a canonical one and, even if it were, there are tricky edge cases: e.g., implants into weapons systems at times of peace, cyberattacks against dual-use infrastructures (especially those European-wide), weaponized cyberespionage against European national security establishments and their members, and a heavier electronic jamming environment.

Finally, whatever alliance efforts (over and above national or private efforts) are made to secure dual-use critical infrastructures, they would have to be deconflicted so that NATO and the Baltic cyberspace alliance do not trip over each other being helpful. This is mostly a notional concern given the limited contribution that any outsider group (much less a foreign outside group) can make to defending specific networks.

8. ROADS NOT TAKEN

Among the objections to a Baltic cyberspace alliance is that there are other European institutions to take care of the matter, and that the countries best suited for such membership may not necessarily be Baltic at all.

One such institution is the EU. Because cyberattacks can influence economic and political well-being, there is a natural compatibility between the EU's mission and collective action to help promote cybersecurity. Certain critical infrastructures under threat from cyberattack, notably the electric grid, span the EU. Correspondingly, the EU is a vital participant in whole-of-infrastructure protection efforts. But cyber security is not just a matter of hardening networks and systems. It involves intelligence to understand how and why such systems may be attacked and it may involve active defenses to stymie imminent and ongoing cyberattacks.[23] There may also be circumstances where reprisals may be called for; even if some reprisals such as economic sanctions can be organized under EU auspices,[24] those that involve cyber operations are, again, incompatible with the EU's purpose. Intelligence, active defenses, and retaliatory cyberattacks are, instead, actions of national security communities.

The question of membership in the Baltic cyberspace alliance involves tradeoff: more members means more clout but also less focus and possibly less consensus. As noted, Norway and the Netherlands may be useful members of such an alliance even though neither abuts the Baltic. What about France? On the one hand, France's emphasis on cyberspace[25] looks much like Germany's, and the bilateral relationship between France and Germany can be understood as the cornerstone of Europe's stability. On the other hand, geography (e.g., distance from Russia) and history (e.g., former colonies) may lead France to different perspectives from Germany on the Russian threat from cyberspace. What about the UK? On the one hand, the UK government's skepticism regarding Russian intentions is well understood, and its GCHQ brings considerable assets to the fight in cyberspace. But the UK is part of the Five Eyes group; thus, any intelligence-sharing arrangement the UK has with Baltic states necessary means similar intelligence-sharing arrangements with all the other Five Eyes members (notably, the United States), who may be uncomfortable with such sharing. Furthermore, the advantage of ambiguity afforded by having two independent alliances taking on Russia in cyberspace would be vitiated if both alliances contained the same member.

9. CONCLUSION

A hypothesized Baltic cyberspace alliance, along the lines laid out above, would send a strong signal from Europe that it intends to oppose Russia's hybrid warfare activities in general and its information warfare campaign in particular. It would add complexities and uncertainties to Russia's aggressive campaigns, and should thereby slow them down and make them easier to counter.

[23] Reportedly, U.S. Cybercommand stymied 2018 Congressional election interference by blocking Internet access enjoyed by Russians' Internet Research Agency. Ellen Nakashima, "U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms," February 26, 2018; https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff3-22e9_story.html.

[24] "In joint conclusions after the EU summit, heads of state denounced aggressive cyber action but stopped short of signaling a move toward decisive EU deterrence against Russia." From Laurens Cerulus, "Russia dodges bullet of EU sanctions on cyber -- for now," October 18, 2018; https://www.politico.eu/article/russia-dodges-eu-sanction-on-cyber-for-now/.

[25] France plans a cyberspace force of 4,000 by 2025; see Arthur Laudrain, "France's New Offensive Cyber Doctrine," February 26, 2019; https://www.lawfareblog.com/frances-new-offensive-cyber-doctrine.

The most obvious alternative to such an alliance would be to strengthen NATO's cyberspace capabilities – which is already going on.[26] But the paper argues that a Baltic cyberspace alliance that operates above the tactical level (because it would not support kinetic operations) would offer several advantages. It would bring in friendly but neutral countries, allow Germany to exercise a leadership role in European defense, and have the countries in Europe most affected by Russian mischief cooperate in warding it off. Even some of the disadvantages – it might compete with NATO – can be advantages (competition is good). But most of all, it complicates Russian decision-making as regards cyberspace by making threats of retaliation more credible and harder to counter-deter.

[26] Not only has NATO declared cyberattacks an Article 5 issue, but in late August 2018, NATO established a military command center able to mount its own cyberattacks with capabilities offered by the United States, Britain, Estonia, and others. From Robin Emmott, "NATO cyber command to be fully operational in 2023," October 16, 2018; https://www.reuters.com/article/us-nato-cyber/nato-cyber-command-to-be-fully-operational-in-2023-idUSKCN1MQ1Z9.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

"Silent Battle" Goes Loud: Entering a New Era of State-Avowed Cyber Conflict

Keir Giles
Conflict Studies Research Centre
Northamptonshire, UK
keir.giles@conflictstudies.org.uk

Kim Hartmann
Conflict Studies Research Centre
Northamptonshire, UK
kim.hartmann@conflictstudies.org.uk

Abstract: The unprecedented transparency shown by the Netherlands intelligence services in exposing Russian GRU officers in October 2018 is indicative of a number of new trends in state handling of cyber conflict. US public indictments of foreign state intelligence officials, and the UK's deliberate provision of information allowing the global media to "dox" GRU officers implicated in the Salisbury poison attack in early 2018, set a precedent for revealing information that previously would have been confidential.

This is a major departure from previous practice where the details of state-sponsored cyber attacks would only be discovered through lengthy investigative journalism (as with Stuxnet) or through the efforts of cybersecurity corporations (as with Red October). This paper uses case studies to illustrate the nature of this departure and consider its impact, including potentially substantial implications for state handling of cyber conflict. The paper examines these implications, including:

- The effect of transparency on perception of conflict. Greater public knowledge of attacks will lead to greater public acceptance that countermeasures should be taken. This may extend to public preparedness to accept that a state of declared or undeclared war exists with a cyber aggressor.
- The resulting effect on legality. This adds a new element to the long-running debates on the legality of cyber attacks or counter-attacks, by affecting the point at which a state of conflict is politically and socially, even if not legally, judged to exist.

- The further resulting effect on permissions and authorities to conduct cyber attacks, in the form of adjustment to the glaring imbalance between the means and methods available to aggressors (especially those who believe themselves already to be in conflict) and defenders. Greater openness has already intensified public and political questioning of the restraint shown by NATO and EU nations in responding to Russian actions; this trend will continue.
- Consequences for deterrence, both specifically within cyber conflict and also more broadly deterring hostile actions.

In sum, the paper brings together the direct and immediate policy implications, for a range of nations and for NATO, of the new apparent policy of transparency.

Keywords: *cyber conflict, cyber policy, attribution, deterrence, transparency*

1. EMERGING PRACTICE

Coordinated disclosures by a number of Western powers of details of cyber attacks and other hostile actions appear to indicate a new multinational policy of state transparency regarding the handling of selected cyber incidents. Combined with the growing power of private citizens and non-governmental organisations engaging in open source intelligence collection and analysis, this may lead to a substantially new phase in the development of cyber conflict.[1]

State cyber activities have traditionally been deeply classified, for a range of reasons including not disclosing either capabilities or vulnerabilities. According to one analysis, “The entire phenomenon of cyber war is shrouded in such government secrecy that it makes the Cold War look like a time of openness and transparency”.[2] And yet, the unprecedented level of detail disclosed by the Netherlands intelligence services in exposing Russian GRU officers in October 2018 signalled a new departure in state handling of cyber conflict. US public indictments of foreign state intelligence officials, and the UK’s release of limited information which enabled third parties to independently identify the Salisbury attackers, set precedents for revealing information that previously would have been confidential, and confirmed a number of new trends in emerging practice.

[1] For an overview of the developmental phases of cyber conflict to date see Max Smeets and Jason Healey, “Cyber Conflict History”, Cyber Conflict Studies Association, 2017, http://static1.1.sqspcdn.com/static/f/956646/28023292/1541729131737/SotF+2017+CCSA+SIPA+History.pdf

[2] R.A. Clarke and R. Knake, *Cyber War: The Next Threat to National Security and What to Do About It* (2010), p. xi

In traditional state practice, a cyber incident would be subjected to a long and painstaking phase of incident analysis before any consideration was given to public attribution. This analysis would include technical evidence as well as supporting material from other sources (historical, geopolitical context, signals and human intelligence and more). The incident analysis would ordinarily be confidential and not available to the public, which might only learn details of the incident through the investigations of private sector cyber security corporations. The second phase, of public or diplomatic attribution by a body or representative of a state, would be considered based on foreign policy considerations as well as on objective evidence. Throughout 2018, however, a shift in practice has been observable as state victims of cyber incidents become increasingly transparent about the details of the investigative phase, whether before or after attribution to a perpetrator: there is increasing disclosure of codes, networks, names, locations, dates, procedures, methodologies, human relationships and relations to other cyber incidents.[3] If this process continues, cyber conflict will change from being a silent battle to one conducted at full volume in the same manner as other forms of state-on-state confrontation.

A general trend towards increased disclosure of cyber incidents in the corporate sector has been noted in the current decade.[4] However, disclosure of state-on-state confrontations increased significantly during 2018 in particular. The Centre for Strategic and International Studies (CSIS) reports on significant cyber incidents on a regular basis, with "significant" meaning attacks carried out "on government agencies, defense and high tech companies, or economic crimes with losses of more than a million dollars".[5] According to the CSIS "Significant Cyber Events List since 2006", during the year 2018, 112 significant cyber incidents were reported, and of these reports almost 45% were official government statements. In addition, these official statements were proactively offering deep insights into the incident detected, the measures taken to counter it, and specific details on the perpetrators.[6] By comparison, for the year 2017 CSIS logged 60 such incident reports, and only 38 in 2016 – and of the 2016 reports, only eight contained any detail over and above simple confirmation that an incident had occurred.

[3] This is partly facilitated by the investigative methods used in technical incidents, which generally include the immediate creation of a "forensic duplicate" of all items involved in the investigative phase. As this guarantees that no evidence from the system can be removed or altered, it allows earlier distribution of investigative results to a broader audience, even prior to the attribution of the incident to a perpetrator. Specifications for forensic duplicates may be found in "Leitfaden IT-Forensik' Version 1.0.1", Bundesamt für Sicherheit in der Informationstechnik (BSI), March 2011.

[4] Derryck Coleman, "Cyber Risk Disclosure On The Rise", Audit Analytics, 23 November 2016, https://www.auditanalytics.com/blog/cyber-risk-disclosure-on-the-rise/; Hilary, Gilles and Segal, Benjamin and Zhang, May H., Cyber-Risk Disclosure: Who Cares? (October 14, 2016). Georgetown McDonough School of Business Research Paper No. 2852519. Available at SSRN: https://ssrn.com/abstract=2852519 or http://dx.doi.org/10.2139/ssrn.2852519

[5] Centre for Strategic and International Studies (CSIS), Significant Cyber Incidents, 9 March 2019, https://www.csis.org/programs/cybersecurity-and-governance/technology-policy-program/other-projects-cybersecurity

[6] Centre for Strategic and International Studies (CSIS), Significant Cyber Incidents full report since 2006, 9 March 2018, https://csis-prod.s3.amazonaws.com/s3fs-public/190211_Significant_Cyber_Events_List.pdf

Emerging state practice also shows that in addition to occurring with higher frequency, transparency efforts are increasingly:

- **Collective:** increasingly, multiple states attribute cyber incidents jointly, and a nascent "transparent cyber alliance" is discernible.
- **Coordinated in policy:** there were at least two instances in 2018 when public release of details of a cyber incident was coordinated with other major political events (see case studies below). This pattern of coordination is reflected in the establishment of political tools and mechanisms, such as the EU Cyber Security Diplomatic Toolbox or NATO Mechanisms for Response.
- **Coordinated in time:** in early October 2018 the British, New Zealand and Australian governments published a list of GRU attacks described as "indiscriminate and reckless cyber attacks targeting political institutions, businesses, media and sport" around the world. Immediately afterwards, the Netherlands authorities released the details of the GRU attempt to hack into the headquarters of the Organisation for the Prohibition of Chemical Weapons in The Hague, detected and interdicted several months before in early April. Finally, on the same day, the US Department of Justice announced criminal charges against seven Russian military intelligence officers.
- **Independent of the scale, nature or impact of the event:** the disclosure of the attempted OPCW hack shows that states do not always consider only the scale and gravity of the operation as a rationale for public attribution, but also the target (as with the OPCW as an international organisation) and the context (the perpetrators involved being also involved in other major cyber incidents).

Key Western allies appear to have shifted to a "public engagement campaign" intended to disrupt and deter cyber attacks and other forms of hostile activity.[7] This is despite the absence of any official national or international statement on change of policy. Explicit policy changes appear limited to very specific types of attack, for instance disinformation attacks on the United States. At the July 2018 Aspen Security Forum, then US Deputy Attorney General Rod Rosenstein seconded a recommendation that the US Justice Department should, under certain circumstances, publicly disclose and attribute foreign influence operations, noting that: "Exposing schemes to the public is an important way to neutralize them" and that "attribution of foreign influence operations can help to counter and mitigate the harm caused by foreign-government-sponsored disinformation." In September of the same year, this became official policy,

[7] Alexander Smith, "Norway calling out Russia's jamming shows European policy shift", *NBC News*, 24 November 2018, https://www.nbcnews.com/news/world/norway-calling-out-russia-s-jamming-shows-european-policy-shift-n937886

as the US Justice Department included a section on “Disclosure of Foreign Influence Operations” as part of an update of the US Attorney’s Manual.[8]

Nevertheless, the move to wider public disclosure of the fine detail of cyber incidents is visible in the United States in particular. During late 2018, the pace of detailed US public indictments accelerated notably. In September, US officials indicted a North Korean man for his alleged role in the hack of Sony Pictures studios, almost four years after the attack. In October, seven Russian military intelligence officers were charged with “computer hacking, wire fraud, aggravated identity theft, and money laundering.” In early November, indictments were made public against more than a dozen Chinese men accused of hacking American aerospace firms for five years beginning in January 2010.[9] But, as the following case studies show, this trend is accompanied by substantial international cooperation to maximise the effect of transparency.

2. CASE STUDIES

Public attribution of a cyber incident by a state directly accusing another state is not in itself new, and case studies are available from before 2018. In May 2014, the US Department of Justice indicted five officers of China’s Unit 61398 for commercial theft in the US;[10] and in February 2015 Norway publicly accused China of commercial cyber espionage and use of the stolen data for the development of new military technology. But 2018 represented a watershed in the frequency, transparency, and method of delivery of public attribution. In addition to the instances already mentioned, in February NotPetya was publicly attributed to the Russian Federation by the UK, Denmark, the US, Canada, Australia and New Zealand, later supported by Estonia, Latvia, Lithuania, Finland and Sweden. In April Germany publicly accused Russia of a cyber attack on the IVBB government data network.[11] In mid-July the US charged 12 GRU officers with a range of offences connected with attacks on the 2016 presidential election.[12] And in October the UK Foreign Office issued a statement in which it jointly with Microsoft accused the Lazarus group, supported by the DPRK, of the WannaCry attack. This attribution was later supported by the US, Canada, New Zealand and Japan. Finally for the year, in late December the US announced a further

[8] Eliot Kim, “Summary: Justice Department Policy on ‘Disclosure of Foreign Influence Operations’”, *Lawfare*, 16 October 2018, https://www.lawfareblog.com/summary-justice-department-policy-disclosure-foreign-influence-operations

[9] Ben Watson, “Special Report: Is the US Ready to Escalate in Cyberspace?” *Defense One*, 21 November 2018, https://www.defenseone.com/ideas/2018/11/special-report-us-ready-escalate-cyberspace/153001/

[10] “U.S. Charges Five Chinese Military Hackers for Cyber Espionage Against U.S. Corporations and a Labor Organization for Commercial Advantage”, Department of Justice, 19 May 2014, https://www.justice.gov/opa/pr/us-charges-five-chinese-military-hackers-cyber-espionage-against-us-corporations-and-labor

[11] “Moscow likely behind hack on German govt, spy chief says”, Reuters, 11 April 2018, https://www.reuters.com/article/us-germany-security/moscow-likely-behind-hack-on-german-govt-spy-chief-says-idUSKBN1HI19D

[12] Indictment available at https://www.justice.gov/file/1080281/download

round of sanctions in retaliation for cyber attacks and "other malign activities,"[13] and the US and UK jointly accused China of a long-running campaign of intellectual property theft, in disclosures backed by Australia and New Zealand and seen as signalling "growing global coordination against the practice."[14]

Amid this accelerating pace of disclosures, late September and early October 2018 saw two instances which exemplified all the new features of the apparent internationally coordinated policy of transparency over hostile actions. In September the British government disclosed details of the two suspects in the poisoning of Sergey and Yuliya Skripal in Salisbury, UK. The next day saw a debate in the United Nations Security Council, initiated by the UK, which must have been preceded by a long period of painstaking multilateral diplomatic preparation. In prepared statements the leaders of the United States, France, Germany and Canada backed Britain's assessment,[15] while a round of statements from countries represented on the Security Council either condemned Russia or were cautiously equivocal, depending on how much each country had to lose from falling out with Moscow. More than 20 countries subsequently supported the UK in its allegations against Russia, expelling more than 100 Russian diplomats between them.

A month later, a similar degree of international coordination over disclosures was evident in the release by the Netherlands of highly detailed information on the interdiction of an attempted hack of the Organisation for the Prohibition of Chemical Weapons in The Hague in the previous April.[16] Near simultaneous announcements were made by the UK and US. A British government statement delivered by the UK Ambassador to the Netherlands promised further public action in close cooperation with allies "confronting, exposing and disrupting the GRU's activity."[17] And on the same day, the US charged seven GRU officers with hacking and other offences related to a report on Russia's systematic state-sponsored subversion of the sport drug-testing process. Four of the seven had travelled to The Hague to carry out the attempted cyber attack on the OPCW, and three had also been indicted in relation to attacks on the US presidential election. As in other instances, the indictment contained highly detailed

[13] "Treasury Targets Russian Operatives over Election Interference, World Anti-Doping Agency Hacking, and Other Malign Activities", U.S. Department of the Treasury, 19 December 2018, https://home.treasury.gov/news/press-releases/sm577

[14] "U.S., allies slam China for economic espionage, spies indicted", Reuters, 20 December 2018, https://www.reuters.com/article/us-china-cyber-usa/u-s-allies-slam-china-for-economic-espionage-spies-indicted-idUSKCN1OJ1VN

[15] Angela Dewan and Nada Bashir, "World leaders back UK's Novichok nerve agent allegations against Russia", *CNN*, 6 September 2018.

[16] "Netherlands Defence Intelligence and Security Service disrupts Russian cyber operation targeting OPCW", Government of the Netherlands, 4 October 2018, https://www.government.nl/latest/news/2018/10/04/netherlands-defence-intelligence-and-security-service-disrupts-russian-cyber-operation-targeting-opcw

[17] "Minister for Europe statement: attempted hacking of the OPCW by Russian military intelligence", UK Government, 4 October 2018, https://www.gov.uk/government/speeches/minister-for-europe-statement-attempted-hacking-of-the-opcw-by-russian-military-intelligence

descriptions of the activities of individual GRU officers, identifying fake accounts and domain names and the precise times and locations of specific online activities.[8]

The involvement of the US and the UK in both incidents reflects a shared perception of the Russian challenge in both governments. In the US this indicates recognition of the wide range of cyber threats emanating from Russia,[19] and in particular the broad range of hostile activities undertaken against the United States, including for example against key utilities and infrastructure.[20] And a new readiness by senior figures in the UK to publicly recognise and state the challenge of ongoing offensive cyber activity from Russia had been discernible from early 2018.[21] The heightened willingness of British intelligence agencies to respond firmly to Russia may account for later reports that a long-serving Russian spy in the Austrian armed forces was arrested on the basis of information provided by the UK.[22]

3. EFFECTS AND IMPLICATIONS

A policy of transparency has a range of implications beyond the possible immediate aim of deterring hostile cyber actors. Before considering deterrence itself, this section highlights potential second- and third-order effects of more open handling of cyber incidents.

A. Legality in Cyberspace

The result of greater publicity for cyber incidents is not only to turn up the volume on a previously silent battle. It also transforms cyber conflict from being invisible to being apparent and tangible. Details disclosed by states based on intelligence sharing/gathering or sophisticated investigations make cyber conflict comprehensible and real rather than an abstraction that publics find difficult to imagine and to relate to their own lives. This could add a new element to the long-running debates on the legality of cyber attacks or counter-attacks, by affecting the point at which a state of conflict is politically and socially, even if not legally, judged to exist.

[18] "U.S. Charges Russian GRU Officers with International Hacking and Related Influence and Disinformation Operations", US Department of Justice, 4 October 2018, https://www.justice.gov/opa/pr/us-charges-russian-gru-officers-international-hacking-and-related-influence-and

[19] Nicu Popescu and Stanislav Secrieru (eds.), "Hacks, Leaks and Disruptions: Russian Cyber Strategies", Chaillot Papers No. 148, October 2018, https://www.iss.europa.eu/sites/default/files/EUISSFiles/CP_148.pdf

[20] Rebecca Smith and Rob Barry, "America's Electric Grid Has a Vulnerable Back Door—and Russia Walked Through It", *The Wall Street Journal*, 10 January 2019, https://www.wsj.com/articles/americas-electric-grid-has-a-vulnerable-back-doorand-russia-walked-through-it-11547137112

[21] Lizzie Dearden, "Britain has entered 'new era of warfare' with Russian cyber attacks, Defence Secretary warns", *The Independent*, 15 February 2018, https://www.independent.co.uk/news/uk/home-news/russia-cyber-attacks-notpetya-gavin-williamson-defence-secretary-putin-hacking-ransomware-a8212801.html

[22] Michael Jungwirth, "Britischer Geheimdienst ließ Putins Spion in Österreich auffliegen", *Kleine Zeitung*, 11 November 2018, https://www.kleinezeitung.at/politik/aussenpolitik/5528189/Der-Tipp-kam-aus-London_Britischer-Geheimdienst-liess-Putins-Spion

Invocation of the principles of international humanitarian law in cases of cyber conflict remains rare. Only a few states have been explicitly clear about the application of international law in cyberspace: once again the UK,[23] the US and the Netherlands. And the division persists between the Western view of the applicability of international law in cyberspace, and that held by Russia, China and like-minded nations, despite a slowly evolving normative debate. Even where states have engaged in international forums on cyber norms (UN GGE, the Global Commission on the Stability of Cyberspace, regional organisations, the OSCE and more) there is an apparent reluctance to adopt an open position on what is lawful in cyberspace and what is not. This is partly due to considerations among states that consider themselves bound by the rule of law not to set a threshold below which an adversary can attack without fear of countermeasures.

But even if states do not explicitly invoke international law when publicly attributing or indicting individuals for cyber attacks, the rationale behind 'going loud' and the emerging State practice is to show that malicious cyber operations

- are not acceptable;
- will not remain secrets kept only by the respective intelligence communities; and
- will incur consequences (even if the eventual consequences or countermeasures if there is no prospect of prosecution of indicted individuals remain to be seen).

In general, open, transparent and public condemnation of incidents demonstrate states' understanding of legality in cyberspace, and their understanding of what constitutes unlawful behaviour. This assumption does not entirely work *a contrario*: if a state does not engage in naming and shaming, this does not mean that it perceives the cyber incident in question as legal, but perhaps it has not yet or fully determined its position on regulation in cyberspace – or indeed does not possess the capability to attribute clearly at any level. Nevertheless, overall a greater adoption of transparency must accelerate the development of international customary law, by forcing open and public consideration of specific documented instances rather than abstract and hypothetical studies.

B. Permissions and Authorities

The reluctance of states to commit to specific interpretations of legality in cyberspace leaves open the argument that cyber operations take place in a grey zone of legal ambiguity.[24]

[23] "Cyber and International Law in the 21st Century", UK Government website, 23 May 2018, https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century

[24] Kubo Mačák, "From Cyber Norms to Cyber Rules: Re-engaging States as Law-makers", *Leiden Journal of International Law* (2017), 30, pp. 877–899, doi:10.1017/S0922156517000358

At the same time, increased state transparency on cyber and other incidents will inevitably lead to greater public knowledge of attacks, and develop a broader consciousness of a state of ongoing conflict by highlighting instances of state-sponsored hostile action. This may start to redress the striking imbalance in public consciousness between aggressors and defenders. This is particularly marked in the case of countries such as Russia, whose state media has been promoting war rhetoric for almost a decade and whose population is constantly reminded that their country is in conflict with the West and that the internet presents a means through which the West can attack and subvert Russia.[25] By contrast, Western countries' publics are only dimly and intermittently aware that Russia wishes them harm.

In this case, public pressure for retaliatory measures may grow. In particular, public and political questioning of the restraint shown by NATO nations in responding to hostile actions by rogue states will intensify still further. This may in turn lead to adjustments to the restrictions on Western cyber and other agencies, whose permissions and authorities to take action generally presume a state of peace, and consequently are greatly more constrained than those of their adversaries. In short, if publics and policy-makers are more aware that war is being waged against them, whether declared or not, they are more likely to favour responses in kind.

Indicators of this kind of movement are already visible on the national and supranational levels. In the US, some of the restrictions governing the approval process for offensive cyber attacks against adversaries were lifted in September 2018,[26] accompanying a strategic reorientation in cyber described as "defend forward."[27] NATO declaratory policy, too, allows "responding in a coordinated manner" to attributed malicious cyber activity.[28]

C. Deterrence

These types of measures may in the medium term enhance the capability of Western nations to implement effective deterrence in cyberspace. For now, public identification of perpetrators, even if accompanied by indictments, is of limited effect if those perpetrators are unlikely ever to be present in a jurisdiction where they could be arrested and tried. Consequently the primary value of transparency at present is in combating the perceived anonymity and immunity of cyber operations;[29] in the US in

[25] Kim Hartmann and Keir Giles. "Net neutrality in the context of cyber warfare", *2018 10th International Conference on Cyber Conflict (CyCon)*. IEEE, 2018.

[26] Erica Borghard and Shawn Lonergan, "What Do the Trump Administration's Changes to PPD-20 Mean for U.S. Offensive Cyber Operations?" Council on Foreign Relations, 10 September 2018, https://www.cfr.org/blog/what-do-trump-administrations-changes-ppd-20-mean-us-offensive-cyber-operations

[27] Max Smeets and Herb Lin, "An Outcome-Based Analysis of U.S. Cyber Strategy of Persistence & Defend Forward", *Lawfare*, 28 November 2018, https://www.lawfareblog.com/outcome-based-analysis-us-cyber-strategy-persistence-defend-forward

[28] "Brussels Summit Declaration", NATO, 11 July 2018 https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2018_07/20180713_180711-summit-declaration-eng.pdf

[29] Jory Heckman, "WH cybersecurity coordinator seeks more 'naming and shaming' of hackers", *Federal News Network*, 29 January 2018, https://federalnewsnetwork.com/cybersecurity/2018/01/wh-cybersecurity-coordinator-seeks-more-naming-and-shaming-of-hackers/

particular, this follows recognition that the Obama administration's muted response to Russian attacks on the US democratic process during the 2016 presidential election was counterproductive, and encouraged Russia in the belief that it could carry out further attacks with little risk of adverse consequences. A secondary effect is to allow less complicated sharing of cyber intelligence; once the information is declassified and publicly available, there are no constraints on passing it on to third-party victim states, or to the media or private sector security corporations in order to assist their own investigations. Each of these actions will have its own deterrent effect.

But critics argue that there is little point in naming and shaming a perpetrator that feels no shame. Indeed in some cases Russia in particular may be appreciative of the publicity, since "just as with so many other aspects of Moscow's geopolitics, there is a theatrical aspect… as the country tries to assert an international status out of proportion with the size of its economy, its soft power and arguably even its effective military strength."[30] This suggests that the prospect of further and more substantive countermeasures may be required in order to deliver deterrence, and it is this consideration which probably lies behind public announcements that the UK had "war-gamed a massive cyber-strike to black out Moscow if Vladimir Putin launches a military attack on the West",[31] followed shortly by similar messaging from the US.[32]

In the US at least, the new policy of transparency has extended in at least one case to acknowledging countermeasures. Instances of operations in cyberspace that are combined with overt and public acknowledgement by the perpetrator are exceptional; ordinarily if there is any accompanying messaging it is kept strictly confidential, and in public responsibility is vehemently denied. The US is now tracing back and directly contacting individuals engaging in online disinformation operations on behalf of the Russian state, with the aim of overtly warning them they could be personally liable to public exposure, indictment, and sanctions from the US government.[33] This departure from anonymity constitutes a striking precedent, which if extended to other forms of cyber operation could substantially change how governments view the delivery of cyber effects.[34]

[30] Mark Galeotti, "Heroes of the Fatherland: Killing Here, Hacking There", *The Moscow Times*, 25 December 2018, https://themoscowtimes.com/articles/heroes-of-the-fatherland-killing-here-hacking-there-63901

[31] Caroline Wheeler, Tim Shipman and Mark Hookham, "UK war-games cyber attack on Moscow", *The Sunday Times*, 7 October 2018, https://www.thetimes.co.uk/article/uk-war-games-cyber-attack-on-moscow-dgxz8ppv0

[32] "The Pentagon has prepared a cyberattack against Russia", *Daily Beast*, 2 November 2018, https://www.thedailybeast.com/the-pentagon-has-prepared-a-cyber-attack-against-russia

[33] Sean Gallagher, "Russian trolls get DM from US Cyber Command: We know who you are. Stop it", *Ars Technica*, 23 October 2018, https://arstechnica.com/information-technology/2018/10/us-cyber-command-doxes-dms-warnings-to-russian-disinformation-trolls/

[34] Evan Perkoski and Michael Poznansky, "CyberCom Is Targeting Russia's Election Meddlers — and Changing How Governments Use Cyber", *Defense One*, 31 October 2018, https://www.defenseone.com/ideas/2018/10/cybercom-targeting-russias-election-meddlers-and-changing-how-governments-use-cyber/152455/

4. OUTLOOK AND CONCLUSIONS

In early 2019, the ongoing efforts of the Netherlands to name the perpetrators of state-sponsored hostilities appeared to be continuing. Importantly, this trend is not limited to cyber activities, but extends to other domains as well. In January, for instance, the Dutch government accused Iran of involvement in at least four assassination and bomb plots in Europe since 2015, and disclosed that investigations into two killings in the Netherlands had led to the expulsion of two Iranian diplomats in June 2018, a move that was not disclosed at the time.[35]

But the trend toward transparency in any domain should not be expected to proceed smoothly and without checks and reverses. One constraint on future application may be concern at the prospect of reprisals. One analysis of recent US moves holds that the response to Russia's information offensive has been deliberately restrained, "in large part to keep Moscow from escalating in response by taking down the power grid or conducting some other reprisal that could trigger a bigger clash between great powers."[36] Another significant risk is horizontal escalation, in particular when dealing with states that are willing to apply whole-of-government measures to attacking their adversaries. For instance, public attribution of cyber attacks that have been carried out by states with limited domestic application of the rule of law may lead to reprisals against private individuals. Both Russia and China have demonstrated willingness to retaliate against Western countries by targeting their citizens resident in or visiting those countries. In Russia, at the time of writing, joint US-British-Irish-Canadian citizen Paul Whelan was being held in apparent retaliation for the arrest in the United States of the Russian alleged agent of influence Maria Butina.[37] In China, larger numbers of Canadians have been detained following the arrest in Canada of Huawei Chief Financial Officer Meng Wanzhou.[38] US citizens are also affected by similar measures there. With effect from January 2018, US citizens travelling to China are advised to "exercise increased caution in China due to arbitrary enforcement of local laws," in particular the coercive use of "exit bans" to prohibit individuals from leaving China, sometimes keeping US citizens in China for years.[39]

[35] Adam Taylor, "Did Iran plot four attacks in Europe? The Dutch government thinks so", *The Washington Post*, 8 January 2019, https://www.washingtonpost.com/world/2019/01/08/did-iran-plot-attacks-europe-dutch-government-thinks-so/

[36] Julian Barnes, "U.S. Begins First Cyberoperation Against Russia Aimed at Protecting Elections", *The New York Times*, 23 October 2018, https://www.nytimes.com/2018/10/23/us/politics/russian-hacking-usa-cyber-command.html

[37] Catherine Philip and Tom Parfitt, "British citizen Paul Whelan held in Russia over 'spying for the West'", *The Time*s, 4 January 2019, https://www.thetimes.co.uk/edition/news/british-citizen-paul-whelan-held-in-russia-over-spying-for-the-west-ghglb88kw

[38] "Canada says 13 citizens detained in China since Huawei CFO arrest", Reuters, 4 January 2019, https://www.reuters.com/article/us-usa-china-huawei-tech-idUSKCN1OY05Q

[39] "China Travel Advisory", U.S. Department of State, 3 January 2019, https://travel.state.gov/content/travel/en/traveladvisories/traveladvisories/china-travel-advisory.html

States may choose to withhold public attribution even when confident in their findings and confident that the risk of reprisals can be avoided or mitigated. This means that selective application of transparency and disclosure should allow a calibrated response to cyber incidents. But in all cases, responses in an environment of greater public consciousness will require extremely close coordination between intelligence services, policy-makers, and the deliverers of cyber effects.[40]

State disclosures will not replace the role of non-state actors, whether information security corporations for cyber incidents, investigative journalism for hostile actions in other domains, or a mixture of the two and more. US indictments, and the release by the UK of limited information on the suspects in the Skripal attack, gave independent media and non-governmental investigators the leads required to develop a much clearer picture of the individuals and structures involved in hostile actions.[41] This harnessing of the power of the global media will serve an important function in bringing vulnerabilities to foreign attack to public notice in the victim state, while not compromising confidential sources or legal process by releasing classified information.

In addition, there will be second- and third-order effects of a new policy of open accusations of hostile acts by states that may as yet be imperfectly understood. One such example is in insurance against cyber attack and its consequences; if it is established that an incident was a state-on-state (and especially military) attack, rather than one carried out by criminals in the traditional sense, this will invalidate a whole range of insurance policies. The result could be substantial disruption to the business insurance market, as corporations look for insurance that does not exclude hostile cyber acts.[42]

Finally, and critically, the trend of greater public awareness is not limited to cyber activity or to disclosures by states. In December 2018, President Trump's inability to undertake a trip to Iraq in secret underscored the democratisation of detection of a wide range of formerly confidential government activity. Mass communications, crowdsourcing, and the widespread availability of open source intelligence analysis tools mean that "The era of spy versus spy—if it ever truly existed—has certainly been ended… Today it is spy versus tweeter, plane spotter, criminal, activist, journalist, bored teenage hacker, and who knows who else."[43] The result is that in

[40] As described in Max Smeets, "Integrating offensive cyber capabilities: meaning, dilemmas, and assessment", *Defence Studies*, Volume 18, 2018 - Issue 4, pp. 395-410, DOI: 10.1080/14702436.2018.1508349

[41] See for example "Investigative Report: On the Trail of the 12 Indicted Russian Intelligence Officers", *RFE/RL*, 19 July 2018, https://www.rferl.org/a/investigative-report-on-the-trail-of-the-12-indicted-russian-intelligence-officers/29376821.html

[42] Oliver Ralph and Robert Armstrong, "Mondelez sues Zurich in test for cyber hack insurance", *Financial Times*, 10 January 2019, https://www.ft.com/content/8db7251c-1411-11e9-a581-4ff78404524e

[43] James Ball, "Plane Enthusiasts Spy Air Force One, Reveal Trump's Secret Trip", *The Atlantic*, 28 December 2018, https://www.theatlantic.com/amp/article/579151/

those cases where governments determine that transparency is not the desired option and they wish to keep their enterprises silent, they will be forced to adopt an entirely new approach to measures to protect and disguise activities that otherwise will be conducted in public.[44] This also has implications for deterrence and its applicability to cyber activities. Previously it might have been possible to engage in deterrence by punishment, or simply assertive messaging, by undertaking a cyber operation that was comprehensible to the adversary but invisible to the general public, so the conspiracy of silence between the aggressor and victim would make it possible for the message to be received with no further escalatory retaliation.[45] Now, it may no longer be possible to message or punish privately and expect the incident to remain confidential for long. In short, in cyber operations, as in so many other areas of previously covert state activity, secrets will have a half-life.

Acknowledgement

The authors thank Berta Jarosova for her research contribution to this paper.

[44] Ric Cole, "Rethinking Camouflage", *Medium*, 15 October 2018, https://medium.com/@richard_iain_cole/rethinking-camouflage-74efadf14ff7

[45] Explored in detail by Austin Carson in *Secret Wars: Covert Conflict in International Politics*, Princeton University Press, 2018.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Call to Action: Mobilizing Community Discussion to Improve Information-Sharing About Vulnerabilities in Industrial Control Systems and Critical Infrastructure

Daniel Kapellmann
Cyber-Physical Threat Intelligence
Senior Analyst
FireEye
Reston, VA, USA
danielkapellmann.z@fireeye.com

Rhyner Washburn
National Consortium for the Study of
Terrorism and Responses to Terrorism
(START)
Research Affiliate
University of Maryland
College Park, MD, USA
rhynerwashburn@gmail.com

Abstract: Vulnerability management remains a significant challenge for organizations that handle critical infrastructure worldwide. Hallmark cyber-physical incidents with disruptive and destructive capabilities like Stuxnet (2010) and Triton (2017) have exploited known vulnerabilities in information technology (IT) and operational technology (OT) assets throughout the attack lifecycle. However, the global critical infrastructure security community is still nascent in the field of industrial control systems (ICS) vulnerability management, especially in information-sharing. While their counterparts in IT security have spent years elaborating multiple resources to track and disseminate information about known vulnerabilities, the ICS community lacks specialized mechanisms for knowledge-sharing. Multiple challenges exist when addressing this issue: a general lack of awareness about ICS cybersecurity, the need to consider multiple industry sectors and unique network architectures, and the need to find a balance between protecting and releasing sensitive information regarding critical infrastructure organizations or proprietary vendor knowledge.

[*] Opinions and findings from this paper are solely the authors' and do not necessarily reflect the views of their organizations.

Through a multiphase research initiative based on the user-centered design process, we intend to test and evaluate the feasibility and effectiveness of various information-sharing platform designs for streamlining the discussion of ICS vulnerabilities. In the first phase of this research, we surveyed ICS and critical infrastructure security stakeholders to gain insight into the range of cogent, shared, and divergent views of the community relating to the need for specialized resources to share information about ICS vulnerabilities. We then evaluated what these different perspectives imply for the adoption and success of certain information-sharing platform frameworks. Finally, utilizing these insights, we demonstrated possible alternative paths forward for addressing the challenge of sharing information about ICS vulnerabilities to keep critical infrastructure safe.

Keywords: *Vulnerability management, critical infrastructure, industrial control systems (ICS), norms and standards, cyber-physical, information-sharing*

1. INTRODUCTION

On December 2017, the US National Cybersecurity and Communications Integration Center (NCCIC) publicly released an in-depth analysis of the TRITON/HatMan malware framework [1]. For the first time, the industrial control systems (ICS) community learned about threat actors developing malware specifically to compromise safety instrumented systems (SIS) from critical infrastructure facilities, with potentially disruptive or even destructive implications. According to the report, two vulnerabilities in the Schneider Electric Triconex Tricon were exploited during the incident [1]. This was, however, not the first time that known vulnerabilities in ICS had been leveraged as tools during major cybersecurity incidents. In 2010, threat actors exploited vulnerabilities in Siemens S7 and WinCC during the Stuxnet attack lifecycle, resulting in the disruption of Iranian centrifuges [2]. In 2016, a denial-of-service (DoS) vulnerability in Siemens SPIROTEC products was exploited in Ukraine's power grid to render devices unresponsive and generate a power outage [3].

Industrial control systems are used to monitor and control physical processes for industrial production. They are a key component of critical infrastructure organizations, which are characterized for their importance to the national economic security, public health, and safety of a country [4]. Compromises of ICS are usually not the product of the exploitation of single vulnerabilities: they require threat actors to combine multiple techniques, tactics and procedures (TTPs) to move laterally across networks, and normally involve multilevel exploits at different points of an organization's network architecture [5]. However, single ICS vulnerability exploitation can also result in harm to critical infrastructure or industrial environments. This is mainly true in the case of internet-connected ICS that contain off-the-shelf embedded software. Multiple open source tools such as the Industrial Exploitation Framework (ISF) and Immunity Canvas Gleg Packs have been released to exploit vulnerabilities in ICS components. [6, 7] Following this premise, vulnerability management represents a key component of a defense-in-depth security approach as it enables organizations to address known weaknesses in key operational technology (OT) assets. Asset managers are challenged to perform timely vulnerability assessments and implement patches, updates or compensating controls to address vulnerabilities that are publicly disclosed (even to threat actors) in multiple open source repositories.

Despite the increase in the complexity of adversaries targeting ICS in critical infrastructure, the community continues to struggle to enforce standards that enable efficient information-sharing, which can help organizations implement vulnerability management programs. Most current mechanisms are based on solutions designed to address the needs of the information technology (IT) community, which responds to different priorities. In the IT domain, the cybersecurity priorities are the confidentiality, integrity, and availability of data. In contrast, critical infrastructure organizations prioritize the safety of people and equipment, and the reliability of physical processes [8]. Additionally, the identification and mitigation of vulnerabilities in IT systems is normally achieved leveraging automated tools and scanners [9]. In the case of ICS, organizations require thorough planning to establish vulnerability assessment methodologies, because failed attempts to mitigate weaknesses can cause instability, performance issues, or even a system crash [10]. Strategies to patch vulnerabilities in ICS are highly complex, due to the need to consider factors such as system architecture, configurations, costs and benefits of downtime, bandwidth limitations of legacy devices, equipment that is insecure by design, and vendor interoperability. As a result, the ICS cybersecurity community requires solutions that are tailored to address their specific information needs for ICS vulnerability management.

This paper is the foundation for a multiphase project. We apply the user-centered design process to test and evaluate the feasibility and effectiveness of different information-sharing platform designs for streamlining access to data about ICS vulnerabilities. In the first phase of this research, we distributed a survey to ICS security stakeholders to gain insight into the range of cogent, shared, and divergent views of the community relating to the need for specialized resources to share information about ICS vulnerabilities. We then evaluated what these different perspectives implied for the adoption and success of certain information-sharing platform frameworks. Finally, utilizing these insights, we demonstrated possible alternative paths forward. We highlight that, to the authors' knowledge, there is no pre-existing literature addressing the challenge of information-sharing for vulnerabilities from the ICS perspective.

2. INFORMATION-SHARING PLATFORMS

In 2013, Luc Dandurand and Oscar Serrano discussed the need of the cybersecurity community to develop tools to facilitate information-sharing and automation, in order to efficiently handle information about vulnerabilities, threats, and incidents. The authors identified that at the time most information-sharing mechanisms lacked interoperable standards, data quality validation, and mechanisms to govern and control the use of sensitive information. To address these challenges, they defined the Cyber Security Data Exchange and Collaboration Infrastructure (CDXI) concept, with the objectives of facilitating information-sharing, enabling automation, and fostering interorganizational collaboration [11]. The paper was focused on IT vulnerabilities, and preceded a series of improvements over the years for cybersecurity information-sharing. However, it did not evaluate sources pertaining to ICS vulnerabilities present in critical infrastructure.

The International Association of Crime Analysts (IACA) defines an information platform as a

> centralized computer system that allows authenticated users to collect, manage, share, and discover structured and unstructured datasets from a variety of sources. It is designed to facilitate two-way communication between users … serve as a channel for official and unofficial communication to facilitate top-down, bottom-up, and lateral communication.

The design of information-sharing platforms is based on multiple considerations, which include but are not limited to the types of entities sharing information, membership diversity, the types of exchanged information, the models used to access information, and the users' needs [12, 13, 14, 15].

Information-sharing platforms are intended to provide people or organizations from specific communities with the ability to access historic information, generate knowledge, and define future insights [12]. According to the European Network and Information Security Agency (ENISA), the main incentives for information exchange are economic benefits stemming from cost savings, and benefits from the quality, value, and use of shared data. Information-sharing mechanisms are economically valuable for organizations to streamline decision-making processes and define resource allocation. However, a key challenge to information-sharing is addressing misaligned economic incentives, given the reputational risks it poses for companies disclosing information [16].

Multi-stakeholder collaboration promotes the creation of quality data by concentrating multiple sources of information. However, high quality data requires the fulfillment of certain conditions, including timeliness, specificity, relevance to address the participants' concerns, and a suitable level of granularity [16]. Further research identifies quality and trustworthiness of data as key requirements for inter-organizational information-sharing. The author suggests four main considerations for trustworthiness: the perceived competence of other parties sharing information, openness, trust issues between parties, and reliability/consistency with which information is released [17]. In the next section, we present the landscape of information-sharing, specifically in the case of ICS vulnerabilities.

3. EVOLUTION OF ICS VULNERABILITIES INFORMATION-SHARING

Information-sharing is currently a controversial topic for ICS stakeholders. The community traditionally relied on a model known as "security by obscurity", where industrial networks relied on proprietary assets and were isolated from business networks [18]. Information about systems architecture and characteristics of ICS assets was exposed only to small groups of people to hide vulnerabilities from adversaries. However, "security by obscurity" is no longer appropriate for ICS, given the increasing integration between corporate IT and modern control system architectures [19]. The ICS community is divided between those who believe information about threats and vulnerabilities should not be shared, and those who believe that greater communication between organizations would improve preparedness against adversaries. Other considerations concern whether information-sharing would divert efforts from other more essential security controls, or whether the quality of shared contents and misinterpretations might generate adverse impacts [20].

Interest in ICS cybersecurity began to proliferate in 2010, parallel to the publication of "Protecting Industrial Control Systems from Electronic Threats" by Joe Weiss. Among other topics, the author elaborated on the lack of significant data to demonstrate ICS cybersecurity cases to executives, given the unwillingness of organizations to share information about incidents. He suggested the need for an ICS Computer Emergency Response Team (CERT) to centralize information from multiple stakeholders, process it and share insights with the community [5]. In 2011, the Stuxnet incident targeting Iranian critical infrastructure was publicly recognized. This caught the attention of the international cybersecurity community and drove a significant increase in ICS-specific vulnerability disclosures [21]. The incident highlighted the relevance of ICS cybersecurity as a key component of national security.

While information about threats and incidents against ICS is still handled discreetly, data related to vulnerabilities in assets is already commonly shared by public and private organizations in different platforms. However, private organizations have highlighted the low quality and integrity of public advisories [22]. Some of the most common platforms are vulnerability repositories, Information-Sharing and Analysis Centers/Information-Sharing Analysis Organizations (ISACs/ISAOs), and ICS vendor advisories. Other sources that are not further discussed in this paper include researcher websites and private industry services. Specialized online forums, such as the SANS ICS community, provide a platform for discussions among ICS cybersecurity practitioners, although none of these forums specifically addresses vulnerabilities. Furthermore, international regulation, such as the European Network and Information Security Directive (NIS), currently stresses the need for information-sharing about threats, incidents and vulnerabilities between different stakeholders [23].

A. Vulnerability Repositories

Online repositories are the most common information-sharing platforms for vulnerabilities. Information from the Vulnerability Database Catalog of the Forum of Incident Response and Security Teams (FIRST) indicates there were at least 22 officially recognized vulnerability databases by March 2016 [24]. Data about weaknesses in electronic components is abundant, as reflected by the United States National Vulnerability Database (NVD) which disclosed more than 15,000 vulnerabilities in 2018; however, the number of repositories releasing specialized information about ICS vulnerabilities is very low [25]. The most recognized repository for ICS vulnerabilities is ICS-CERT, which was created in 2009 and placed under the command of the US NCCIC in 2018 [26]. ICS-CERT not only releases information about ICS vulnerabilities, but also collaborates with vendors and researchers to coordinate the process of responsible disclosure. While ICS-CERT advisories are tailored for the ICS community and provide a higher granularity of data than other repositories, the platform still faces significant challenges.

Three main challenges are: concentrating information about ICS vulnerabilities from multiple sources using different data structures; elaborating practical mitigation recommendations that satisfy the needs of the ICS community; and organizing information in accessible and consumable formats [27, 28]. Other recognized repositories that contained information about ICS vulnerabilities were owned by Critical Intelligence [29] and the Open Source Vulnerability Database (OSVDB) [30]. Both databases disappeared between 2015 and 2016 due to the intense manual input required to concentrate the information, and low returns on investment. More recently, the Zero Day Initiative was launched by a private sector organization to reward researchers for vulnerability disclosure. While it does not contain only ICS-tailored information, it has encouraged collaboration with researchers for the disclosure of vulnerabilities.

B. ISACs and ISAOs

ISACs are mechanisms formed by critical infrastructure owners and operators to gather, analyze, sanitize and disseminate information between public and private stakeholders. These organizations are crucial for public-private collaboration in sharing information about vulnerabilities, threats, intrusions and anomalies, mostly in critical infrastructure sectors [31]. The value of ISACs depends on the collective consensus of the members and their willingness to share information. Some examples of ISACs from different sectors are: Electricity (E-ISAC), Oil and Natural Gas (ONG-ISAC), Mining and Metals (MM-ISAC), Maritime (Maritime-ISAC), and the Industrial Control Systems (ICS-ISAC) [29]. In 2015, the Obama administration issued an Executive Order introducing ISAOs as an alternative to address some of the information-sharing limitations of ISACs. These organizations seek to "encourage the formation of communities that share information across a region or in response to a specific emerging cyber threat." [32] Information shared within the ISACs is only communicated among members, limiting their value to the external community.

C. ICS Vendor Advisories

The disclosure of ICS vulnerabilities is highly reliant on the collaboration of commercial product vendors and service providers. While it is not in the scope of this paper to discuss the process of coordinated and responsible disclosure, ICS vendor advisories remain one of the most in-depth sources of information about vulnerabilities. Some of the main vendors of ICS products have invested in developing specialized platforms for sharing information. For example, both Schneider Electric's Cybersecurity Support Portal, and Siemens ProductCERT release regular vulnerability advisories [33, 34].

D. New Media

In 2016, a report from FireEye defining critical lessons from 15 years of ICS vulnerabilities indicated that "media coverage of significant events in ICS security, either attacks or research, will likely continue to fuel the vulnerability disclosure rate." [21] While there is no formal research published about the role of media in sharing information about ICS vulnerabilities, some specialized news outlets regularly share this information. An example is Security Week, which regularly releases notes expanding on the information released in vendor advisories and publications from vulnerability repositories [35]. Social media has also been a tool used by reputable

ICS organizations and experts: for example, ICS-CERT releases regular advisory notifications [36].

4. RESEARCH DESIGN AND METHODOLOGY

Despite the variety of information-sharing platforms available, it remains unclear to what extent they meet the needs of the ICS security community. To address this lack of assessment on information sources supporting ICS vulnerability management and ascertaining what information the ICS community values, we elected to design a subject matter: expert elicitation. Our primary tool for elicitation was a web-based survey, which we distributed among ICS stakeholders in the private, public, academic, and non-profit sectors. The survey was mainly shared on recognized community forums and remained open for one month. It consisted of 22 questions focused on participant background, access to ICS vulnerability data, information needs, and ideal methods for collecting or sharing such information. The seventh question filtered respondents who did not access information about known ICS vulnerabilities. The full questionnaire is available in Appendix A.

For this survey, we attempted to recruit across multiple professional domains and industries. To this end, instead of individually identifying participants, we sent the survey to specialized ICS forums including SANS-ICS community, the Industrial Control Systems Joint Working Group (ICSJWG), and the International Society of Automation (ISA). We also reached out to a select few individuals who are thought leaders or experienced in the ICS and critical infrastructure community, to further spread the survey. Even though convenience sampling implies an intrinsic risk of volunteer bias, we chose this method to identify individuals who were particularly interested or experienced in ICS vulnerabilities. This was mainly relevant to reach a representative sample despite the small size of the population with expertise on this topic.

There are currently no official estimates of the size of the ICS cybersecurity community, for multiple reasons. ICS cybersecurity is a young discipline, spread through diverse industries, that requires skills from multiple disciplines, and has only recently begun to be defined as a knowledge field. After exhaustive research, we decided to adopt as an estimate the number of members present in SANS ICS invitation-only forum, which is 6,300 [37]. However, we recognize that the forum does not only include members actively participating in ICS vulnerability management. Members range from security analysts and ICS owners, to sales representatives, managers, and anyone who has learned the basics about ICS security. As we were unable to capture data of how many people the survey did reach, we could not ascertain an accurate survey response rate.

The first section of the survey contained questions about the previous experience and demographics of the sample. The second section began by asking participants about their habits for accessing information pertaining to ICS vulnerabilities, then identified the main challenges they faced in using this information, and finally asked about their information needs. We added an additional field for comments and invited participants to provide their emails for follow-up interviews during the next phases of the research initiative. We employed descriptive statistics and exploratory data analysis to draw understanding from the participants' responses.

In 2015, Hollifield and Perez released a White Paper showing how designing usable human-machine interface (HMI) displays that fulfilled the needs of operators could improve their capacity to manage physical processes [38]. Our methodology seeks to adopt a similar approach for the design of ICS information-sharing platforms, recognizing that what currently exists follows patterns set by the IT community and does not meet the unique needs of ICS users. In the following section, we present an initial survey of users' needs and preferences to guide the creation of prototype tools for ICS vulnerability information-sharing.

5. FINDINGS AND DISCUSSION

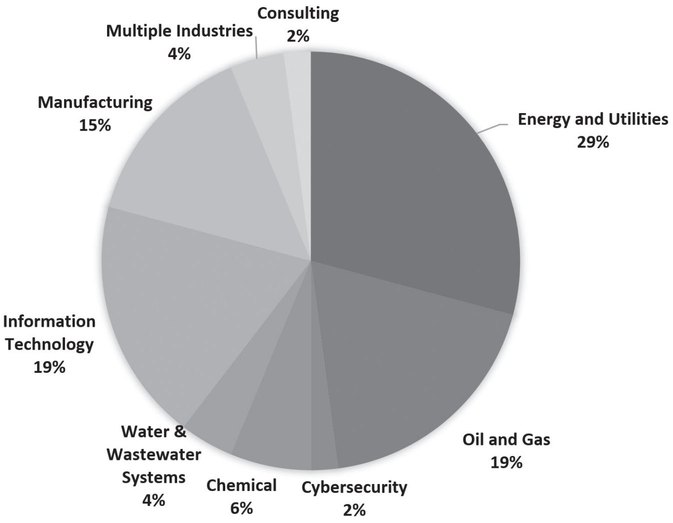
A. Sample Description

The survey captured 48 responses, of which four remained incomplete given that it was designed to exclude non-ICS stakeholders. While a bigger sample would provide higher statistical confidence, we consider that the present survey still provides highly valuable insights: as one of the first systematic efforts to identify the habits, challenges and needs of ICS stakeholders regarding ICS vulnerabilities present in critical infrastructure.

The survey was distributed in forums frequented by stakeholders from different backgrounds. Close to 98% of the individuals who elected to participate stated that they had technical backgrounds in areas such as engineering and computing science. More than 80% of the participants were employed in the private sector, but we also received responses from government, academia, and non-profit professionals. Close to 71% of the participants were currently occupied in the field of cybersecurity, followed by 15% from ICS engineering.

The main strengths of the sample were: a highly diverse group of participants from 15 different industries, with most participation from energy and utilities, oil and gas, information technology and manufacturing (as shown in Figure 1); and a reported medium to high confidence level in cybersecurity expertise from 94% of participants. The main limitation was the small size of the sample. This can be explained mainly by two factors: the previously discussed small size of the ICS cybersecurity community, and some individuals declining to participate due to concerns about sharing information. It is also possible that the lack of active discussion and interest impacted our response rate.

FIGURE 1. DISTRIBUTION OF THE SAMPLE BY INDUSTRY.



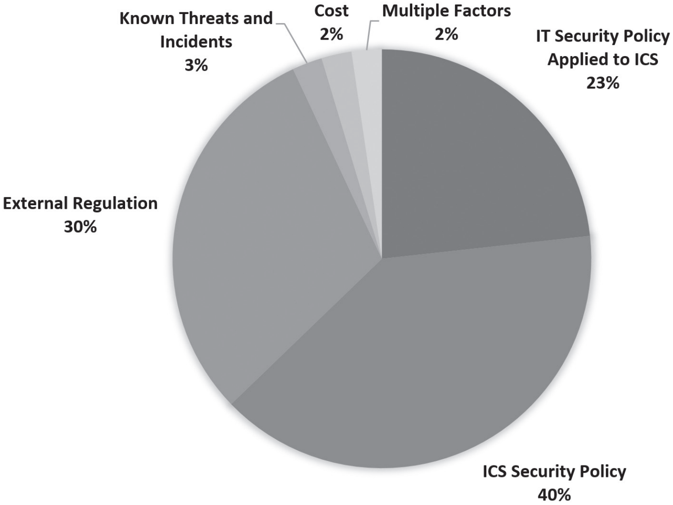
B. Habits, Challenges and Needs Pertaining to Information-Sharing for ICS Vulnerabilities

We categorized survey questions into three sections, to explore the current habits of the ICS community, the challenges they face, and their preferred mechanisms for fulfilling their ICS vulnerability information needs. An additional section is provided to share insights presented by survey responders beyond questionnaire requirements.

1) Habits

Most respondents were intensive consumers of information about ICS vulnerabilities. At least 61% accessed this information daily or weekly, and 20% monthly. The most common purposes for access were general awareness (learning about trends and new threats), research, vulnerability management, risk management and compliance with regulations. Figure 2 shows that despite the unique needs of ICS security, only 40% of the respondents considered ICS security policy to be the main factor driving ICS vulnerability management in their organization. In contrast, 30% considered it was mostly driven by government regulations, and 23% expressed it as IT security policy applied to ICS. This highlights the common adoption of IT resources to facilitate ICS cyber security, and the strong role of government regulations in vulnerability management.

FIGURE 2. MAIN FACTOR THAT DRIVES ICS VULNERABILITY MANAGEMENT IN RESPONDENTS' ORGANIZATIONS.



The primary avenues used by participants to access information about ICS vulnerabilities were ICS/US-CERT (77%), ICS vendor websites (57%), news and media (52%), and the NVD (39%). Participants demonstrated interest in multiple sources of information. Figure 3 illustrates the co-occurrence of source usage. The most common combinations included ICS/US-CERT, vendor websites, and the NVD. We highlight the prevalence of news and media as a source of information, given that a higher quality of information is regularly expected from validated sources such as CERTs and vendor websites. ICS/US-CERT and vendor websites both offer detailed vulnerability advisories, but lack support for checking multiple vulnerabilities at once. Finally, though the NVD contains the most information about vulnerabilities, identifying specific ICS vulnerabilities remains a challenge. Two survey participants noted limitations with this database, including improper association between vulnerabilities, product names as they are known by engineers in the field, and misrepresented risk ratings. These limitations result from the repository's original intention to share information about IT vulnerabilities. ICS products commonly have multiple components of firmware, hardware and software, which makes their naming more complex. In the case of risk ratings, most repositories utilize the CVSS, which does not account for damage caused by vulnerabilities to processes, people or equipment [39].

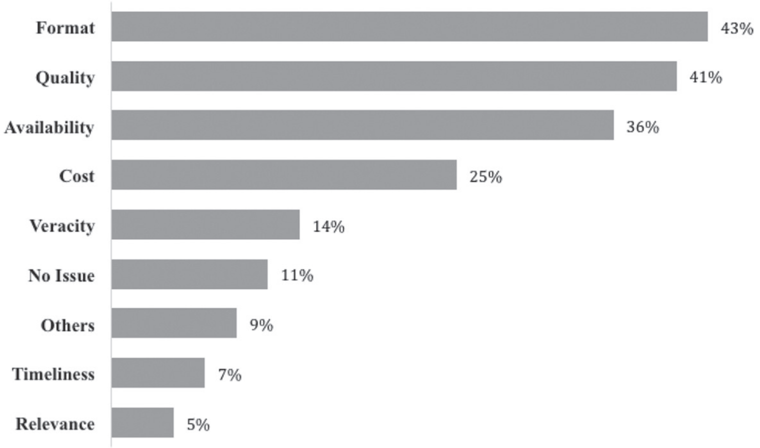
FIGURE 3. CO-OCCURRENCE OF PRIMARY AVENUES USED BY PARTICIPANTS TO ACCESS INFORMATION ABOUT ICS VULNERABILITIES.

	News & Media	ISACs	NVD	ICS/US-CERT	ICS Vendors	Private Industry	Others
News & Media	52%	14%	14%	39%	30%	11%	2%
ISACs	14%	30%	16%	27%	20%	16%	0%
NVD	14%	16%	39%	36%	27%	16%	2%
ICS/US-CERT	39%	27%	36%	77%	43%	23%	2%
ICS Vendors	30%	20%	27%	43%	57%	14%	0%
Private Industry	11%	16%	16%	23%	14%	25%	0%
Others	2%	0%	2%	2%	0%	0%	2%

2) Challenges

Close to 46% of the participants expressed dissatisfaction with the information they obtain through ICS vulnerability resources. At least half of those who expressed dissatisfaction also noted that their ICS security programs were mainly driven by risk management and compliance with regulations. This result can be driven by the high cost and complexity of regulatory requirements. When support from executives is limited, practitioners are challenged to find alternatives for compliance despite this. Figure 4 shows that the main barriers identified by participants in accessing the information they need about ICS vulnerabilities were the data format (43%), quality of information (41%), availability (36%), and cost of good information (25%).

FIGURE 4. MAIN BARRIERS TO FINDING INFORMATION ABOUT ICS VULNERABILITIES.

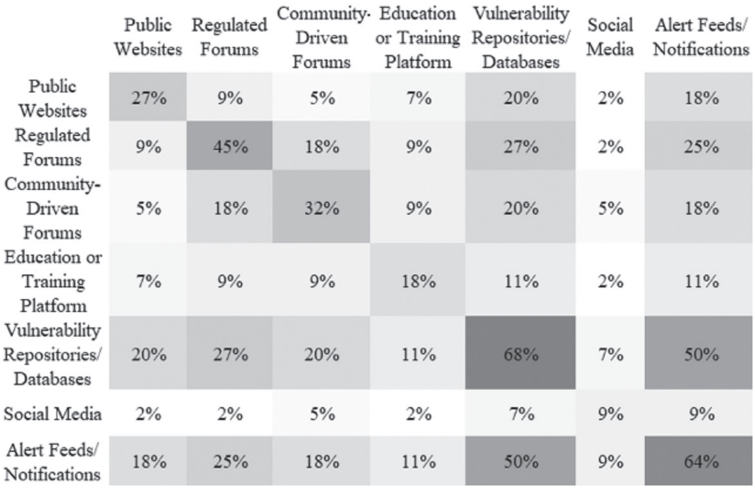


One of the participants who identified the format of information as one of the main challenges included a comment highlighting the inability of his organization to filter large amounts of data to identify risks pertaining to assets. In fact, the most commonly accessed resources (vendor advisories and ICS/US-CERT) are not accessible in single data repositories that enable analysis of multiple vulnerabilities at the same time. In the case of NVD, the large amount of information from IT vulnerabilities makes it difficult to address specific ICS needs. Interestingly, only 11% of the respondents indicated they found no barriers. This shows that even though 54% of the respondents considered they were satisfied with the information they had access to, 89% believed that information-sharing for ICS vulnerabilities had room for improvements.

3) Needs

The last section of the questionnaire was intended to learn about the needs and preferences of the ICS community to access and share information about known vulnerabilities. Figure 5 illustrates the most popular selections for ideal platform design and co-occurrence of multiple choices. These results highlight possible compatibilities between different platforms to inform the future design of solutions and address information-sharing needs. Participants expressed most interest in vulnerability repositories/databases (68%) and alert feeds/notifications (64%), with 50% expressing interest in both. The findings highlight the demand for an ICS vulnerability repository that provides a consumable format for analyzing multiple vulnerabilities at a time. Access to this repository would be preferred through newsfeeds and alerts (55%), an online dashboard (43.2%), application program interfaces (39%), XML or other markup languages (30%), or text reports (27%). Most participants (86%) prioritized quality of information when selecting information platforms over the design itself. Other factors that drove preferences were usability of the platform (55%), the veracity of sources (50%), and accessibility of the platform (41%). Other popular platforms included regulated forums (45%), and community-driven forums (32%).

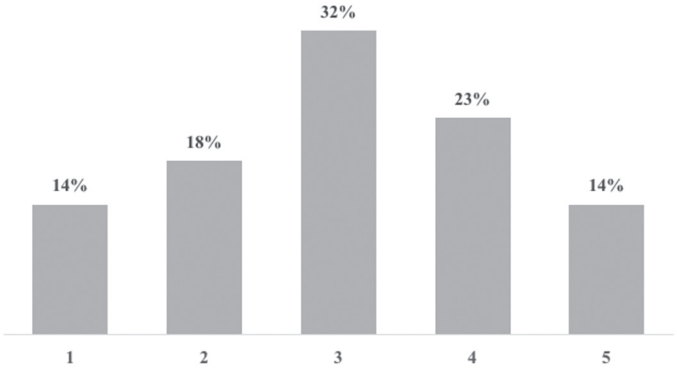
FIGURE 5. CO-OCCURRENCE OF PREFERRED TYPES OF PLATFORMS AS EXPRESSED BY PARTICIPANTS.



	Public Websites	Regulated Forums	Community-Driven Forums	Education or Training Platform	Vulnerability Repositories/ Databases	Social Media	Alert Feeds/ Notifications
Public Websites	27%	9%	5%	7%	20%	2%	18%
Regulated Forums	9%	45%	18%	9%	27%	2%	25%
Community-Driven Forums	5%	18%	32%	9%	20%	5%	18%
Education or Training Platform	7%	9%	9%	18%	11%	2%	11%
Vulnerability Repositories/ Databases	20%	27%	20%	11%	68%	7%	50%
Social Media	2%	2%	5%	2%	7%	9%	9%
Alert Feeds/ Notifications	18%	25%	18%	11%	50%	9%	64%

An additional finding (illustrated in Figure 6), indicated a normal distribution of participants expressing how comfortable they were sharing information about ICS vulnerabilities outside their organization on a scale from 1 to 5. The distribution corroborates that there is as yet no consensus on the topic among the community; though some members are open to sharing information, others are not. Willingness to share information about vulnerabilities may vary between stakeholders. For example, critical infrastructure organizations and ICS vendors commonly resist sharing information, while governments favor collaboration to improve the security of the community. Any solution that is implemented will require the consideration of both perspectives to become a widely used resource.

FIGURE 6. COMFORT WITH SHARING INFORMATION ABOUT ICS VULNERABILITIES OUTSIDE RESPONDENTS' ORGANIZATIONS RANKED FROM 1 (NOT COMFORTABLE) TO 5 (VERY COMFORTABLE).



4) Additional Highlights

From the 44 participants that completed the full survey, 52% provided their contact information to follow up through the research process. This shows a high level of engagement from participants in support of finding solutions to address the challenges discussed. One participant commented that some private sector products were beginning to offer more information about known vulnerabilities and potential mitigations. However, to the authors' knowledge, the listed solutions rely on comparing asset information with data from public repositories that use the Common Vulnerability Enumeration (CVE) format to identify matches. As a result, improvements in public repositories can result in a spillover to higher quality products for the private sector. Another relevant highlight was that vulnerability management requires a large amount of time and resources that is commonly understated by executives. Better quality information about ICS vulnerabilities may reduce the effort required for vulnerability management, increasing the level of preparedness of organizations against known threats.

6. MOVING FORWARD TO IMPROVE INFORMATION-SHARING FOR ICS VULNERABILITIES

Our survey provided a unique opportunity to gather insights from ICS stakeholders following principles from the user-centered design process to develop solutions that adapt to the needs of the industry. While IT software companies have long relied on user-centered methodologies to develop products and services, the ICS security community could still benefit from knowing what are the habits, challenges, and needs of this specific population dedicated to protecting critical infrastructure systems. By publicly releasing this information, we hope to promote and formalize conversations about ICS vulnerability platforms, and spark thoughts with regard to design alternatives. We highlight that addressing ICS vulnerabilities is not only relevant for the private industry, but holds value as a key component to safeguard national security by protecting critical infrastructure processes and assets.

In this first paper, part of a series to identify alternatives for ICS vulnerability information-sharing platforms, we performed exploratory user research on members from the ICS community. Our findings corroborated an interest from most participants in improving ICS vulnerability platforms. While the sample was divided into a normal distribution in terms of comfort with sharing information, there was a consensus on the importance of improving the format, quality, and availability of data. An interesting finding was that most participants prioritized quality over other attributes. Therefore, the first challenge is to identify what information is useful for practitioners, and how to obtain this data given limited resources.

The survey also reflected valuable findings to guide the development of such a solution. Results indicated that an ICS vulnerability repository/database would be highly accepted by the community, mainly in combination with alert feeds and notifications. To a certain extent, ICS/US CERT, ICS vendor resources, and some private organizations issue notifications about new vulnerabilities. Next steps should, however, improve the quality of shared information and offer access in multiple formats to fit the needs of different organizations. Another alternative spawning from this paper is the elaboration of hybrid information-sharing platforms combining features from different models. A particularly interesting experiment would be to combine a vulnerability repository with regulated or community-driven forums. Even though there are currently no forums specializing in sharing information about ICS vulnerabilities, these were a popular idea among respondents. This type of interaction could enable participants to discuss alternative mitigations and clarify misconceptions on known vulnerabilities.

This survey was the first step in recognizing and formally documenting the needs of ICS security practitioners with regard to vulnerability sharing. Conclusions may be known to some and novel to others. Regardless of this, it provides a first step in developing tools based on the needs of actual users. We hope this paper motivates the community to develop alternatives with which we can jointly improve our ability to address ICS vulnerabilities.

7. FURTHER RESEARCH

This research paper provides a precedent to invite the ICS community to develop further research on mechanisms and platforms for sharing information about ICS cyber security. We find the results particularly valuable in guiding the implementation of prototype tools and processes to better address the vulnerability management needs of the ICS community. Further research may also explore the challenges of inter-organizational information-sharing for ICS vulnerabilities and define high quality standards for this data. Finally, as expressed by one of the survey participants, we recognize that information-sharing about threats, incidents, and impacts should also be prioritized as a promising field of study.

Acknowledgments

Thanks to the members of the ICS community who responded or distributed the survey. We thank Sean McBride, Nathan Brubaker, Jeffrey Ashcraft, Alicia Bargar, Margaret Morganti, and Ryan Serabian for their support. Rhyner Washburn also thanks Dr. Steve Sin for his guidance.

REFERENCES

[1] NCCIC, “MAR-17-352-01 HatMan-Safety System Targeted Malware (Update A),” *ICS-CERT*, 2018.

[2] E. Byres and S. Howard, “Analysis of the Siemens WinCC / PCS7 ‘Stuxnet’ Malware for Industrial Control System Professionals,” *Tofino Security*, 2010.

[3] A. Cherepanov and R. Lipovsky, “Industroyer: Biggest threat to industrial control systems since Stuxnet,” 12 June 2017. [Online]. Available: https://www.welivesecurity.com/2017/06/12/industroyer-biggest-threat-industrial-control-systems-since-stuxnet/. [Accessed 05 December 2018].

[4] Federal Emergency Management Agency, “Critical Infrastructure and Key Resources,” [Online]. Available: https://emilms.fema.gov/IS520/PAN0101400text.htm. [Accessed 2 January 2019].

[5] J. Weiss, *Protecting Industrial Control Systems from Electronic Threats*, First Edition ed., New York: Momentum Press, 2010, pp. 63-86.

[6] Immunity, “CANVAS,” [Online]. Available: https://www.immunityinc.com/products/canvas/gleg-products.html. [Accessed 22 March 2019].

[7] dark-lbp, “Industrial Exploitation Framework,” [Online]. Available: https://github.com/dark-lbp/isf. [Accessed 22 March 2019].

[8] A. Ginter, *SCADA Security: What’s broken and how to fix it*, Calgary: Abterra Technologies Inc., 2016.

[9] M. Chapple and D. Seidl, *CompTIA CySA+*, Indianapolis: Sybex, 2017.

[10] E. D. Knapp and J. T. Langill, *Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems*, Massachusetts: Elsevier, 2015.

[11] L. Dandurand and O. Serrano Serrano, “Towards Improved Cyber Security Information Sharing,” in *5th International Conference on Cyber Conflict*, Tallin, 2013.

[12] International Association of Crime Analysts, “Information-Sharing Platforms,” International Association of Crime Analysts, Overland Park, 2014.

[13] N. Sardjoe, “The Interrelation of Information Sharing Levels: Intra-organizational and inter-personal influences on inter-organizational information sharing,” *TUDelft*, Delft, 2017.

[14] S. V. Sundar and D. E. Mann, “Effective Regional Cyber Threat Information Sharing,” *MITRE Corporation*, Bedford, 2016.

[15] R. Krishnan, R. Sandhu, J. Niu and W. H. Winsborough, “A Conceptual Framework for Group-Centric Secure Information Sharing,” in *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, 2009.

[16] N. Robinson and E. Disley, “Incentives and Challenges for Information Sharing in the Context of Network and Information Security,” *European Network and Information Security Agency* (ENISA), Heraklion, 2010.

[17] M. Ibrahim, “Interorganizational Trust and Interorganizational System’s Information Quality,” in *IQ*, 2005.

[18] E. Byres, “The Industrial Cybersecurity Problem,” *International Society of Automation (ISA) Research Triangle Park*, 2013.

[19] Industrial Control Systems Cyber Emergency Response Team (ICS-CERT), “Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies,” *Homeland Security*, 2016.

[20] D. Peterson, “Info Sharing Bubble Burst of Everything Is A Success,” 4 October 2012. [Online]. Available: http://www.digitalbond.com/blog/2012/10/04/info-sharing-bubble-burst-or-everything-is-a-success/. [Accessed 11 November 2018].

[21] FireEye iSIGHT Intelligence, “Overload: Critical Lessons from 15 Years of ICS Vulnerabilities,” *FireEye*, Milpitas, 2016.

[22] Dragos, “Year in Review 2018 - Industrial Controls System Vulnerabilities,” 2018. [Online]. Available: https://dragos.com/wp-content/uploads/yir-ics-vulnerabilities-2018.pdf. [Accessed 18 February 2019].

[23] European Parliament and the Council of the European Union, “Directive (EU) 2016/1148 of the European Parliament and of the Council,” *Official Journal of the European Union*, pp. 6-11, 2016.

[24] FIRST, “Vulnerability Database Catalog,” 17 03 2016. [Online]. Available: https://www.first.org/global/sigs/vrdx/vdb-catalog. [Accessed 20 November 2018].

[25] National Institute of Standards and Technology (NIST), “Statistics Results,” [Online]. Available: https://nvd.nist.gov/vuln/search/statistics?form_type=Basic&results_type=statistics&search_type=all. [Accessed 20 November 2018].

[26] ICS-CERT, “About Us,” [Online]. Available: https://ics-cert.us-cert.gov/about-us. [Accessed 20 November 2018].

[27] R. Wightman, “Industrial Control Vulnerabilities: 2017 in Review,” Dragos, Hanover, 2018.

[28] D. Kapellmann, “Vulnerability Assessments in ICS Environments: Lessons Learned and Wishlist of Improvements,” in *Industrial Control Systems Joint Working Group Spring Meeting*, Albuquerque, 2018.

[29] C. E. Bodungen, B. L. Singer, A. Shbeeb, S. Hilt and K. Wilhoit, *Hacking Industrial Control Systems Exposed*, New York: McGraw-Hill Education, 2017.

[30] SCADAhacker, “SCADAhacker,” [Online]. Available: https://www.scadahacker.com/. [Accessed 22 March 2019].

[31] The White House, “Protecting America’s Critical Infrastructures: PDD-63,” Washington DC, 1998.

[32] Cyber Threat Intelligence Network, “ISAOS,” [Online]. Available: https://ctin.us/site/isaos/. [Accessed 23 November 2018].

[33] Schneider Electric, “Cybersecurity Support Portal,” [Online]. Available: https://www.schneider-electric.com/en/work/support/cybersecurity/overview.jsp. [Accessed 22 March 2019].

[34] Siemens, “Siemens ProductCERT and Siemens CERT,” [Online]. Available: https://new.siemens.com/global/en/products/services/cert.html. [Accessed 22 March 2019].

[35] SecurityWeek, “SCADA/ICS,” [Online]. Available: https://www.securityweek.com/scada-ics. [Accessed 22 March 2019].

[36] ICS-CERT, “ICS-CERT,” Twitter, [Online]. Available: https://twitter.com/ICSCERT. [Accessed 22 March 2019].

[37] SANS, “SANS ICS Community,” [Online]. Available: https://ics-community.sans.org/. [Accessed 22 March 2019].

[38] B. Hollifield and H. Perez, “Maximize Operator Effectiveness: High Performance HMI Principles and Best Practices,” PAS, 2015.

[39] I. Barda, “Increase in CVE Reports vs Long Field-Development - How to Manage the Conflict,” in *ICS Cyber Security Conference*, Atlanta, 2018.

APPENDIX A: RESEARCH QUESTIONNAIRE

Information-Sharing for ICS Vulnerabilities.
Thanks for agreeing to take part in this important survey to better understand the needs and preferences of the ICS community related to the quality and availability of information-sharing platforms for ICS vulnerabilities. This survey consists of 21 questions and is designed to gather insights from different types of stakeholders.

1. What do you consider to be your primary background?
   a. Technical (e.g. engineering or computing sciences)
   b. Non-technical (e.g. policy or social sciences)

2. Which of the following options best describes your sector of work?
   a. Academia
   b. Private sector
   c. Government (including military)
   d. Non-profit

3. Which of the following options best fits your industry?
   a. Energy & utilities
   b. Oil & gas
   c. Manufacturing
   d. Chemical
   e. Water & wastewater systems
   f. Retail/commercial
   g. Legal/regulation
   h. Telecommunications
   i. Information technology
   j. Financial
   k. Healthcare

4. Which of the following options best describes your current occupation?
   a. ICS engineering
   b. Policy and regulation
   c. Cybersecurity
   d. Business/management
   e. ICS Equipment vendor

5. Do you consider yourself a stakeholder in the ICS community?
   a. Yes
   b. No

6. Rate your experience in cybersecurity:
   a. 1 – Not familiar
   b. 2
   c. 3
   d. 4 – Very knowledgeable

7. Do you access information about known ICS vulnerabilities?
   a. Yes (Continue to next section)
   b. No (Finish survey)

No Access to ICS Vulnerabilities

1) Why do you not have access to ICS vulnerability information?
   a. Not relevant to my current job
   b. I am unfamiliar with ICS vulnerability resources
   c. My organization has no vulnerability management program
   d. My organization prioritizes other security controls
   e. Lack of resources (time or funding)

2) Do you have any additional comments or recommendations?

Access to ICS Vulnerabilities

1) How often do you access information about known ICS vulnerabilities?
   a. Daily
   b. Weekly
   c. Monthly
   d. Quarterly
   e. Biannually
   f. Yearly
   g. Less than a year

2) For what purpose do you access this information? (Choose all that apply)
   a. General awareness: learning about trends and new threats
   b. Research: analysis, disclosure or assessment of ICS vulnerabilities
   c. Risk management & compliance: performing risk or vulnerability assessments
   d. Vulnerability management: mitigation of vulnerabilities in ICS

3) Based on your experience, what is the main factor that drives ICS vulnerability management in an organization?
   a. External regulation
   b. IT security policy applied to ICS
   c. ICS security policy

4) What are your primary avenues for accessing information about ICS vulnerabilities? (Choose all that apply)
   a. News and media
   b. Information-sharing and analysis centers (ISACs)
   c. National vulnerability database (NVD)
   d. ICS-CERT/US-CERT
   e. ICS vendor websites
   f. Private industry resource

5) Are you satisfied with the information you are getting through those services?
   a. Yes
   b. No

6) What are the main barriers you encounter to find the information you need? (Choose all that apply)
   a. Cost: good information is costly
   b. Availability: I can't find any information
   c. Format: information is not digestible
   d. Quality: information is subpar
   e. Veracity: sources are not trustworthy
   f. No issue: I do not find any barriers

7) What granularity of data would best satisfy your information needs related to ICS vulnerabilities
   a. 1 – Very broad (Only ID, name, description and resources)
   b. 2
   c. 3
   d. 4
   e. 5 – Very specific (In-depth description containing associated source code, scenarios, requirements for exploit, etc.)

8) What type of platforms do you think would best fit your organization to share or access information about known ICS vulnerabilities?
   a. Public websites
   b. Regulated forums
   c. Community-driven forums
   d. Education/training platform
   e. Vulnerability repositories/databases
   f. Social media
   g. Alert feeds/notifications

9) What factors mostly influenced your choice of best information-sharing platforms for ICS vulnerabilities?
   a. Accessibility of the platform
   b. Usability of the platform
   c. Privacy of the data exchange
   d. Quality of information
   e. Veracity of the sources

10) How comfortable are you sharing information about ICS vulnerabilities outside your organization?
   a. 1 – Not comfortable
   b. 2
   c. 3
   d. 4
   e. 5 – Very comfortable

11) What are the parameters you would want to have in an ideal ICS vulnerability repository/database? (Choose all that apply)
   a. Unique identifier (E.G. CVE)
   b. Vendor
   c. Affected products
   d. Affected versions
   e. Common vulnerability scoring system (CVSS)
   f. CVSS vector string
   g. Common weakness enumeration (CWE)
   h. Exploitability
   i. Risk score
   j. Researcher/author
   k. Critical infrastructure/Industry sectors affected
   l. Potential physical impact
   m. Countries/areas product is deployed

n. Vendor country of origin
o. Available patches/updates
p. Alternative mitigations
q. Tools for exploitation
r. References

12) How would you prefer to access information from this ICS vulnerability repository? (Choose up to two answers)
a. Text reports
b. Spreadsheets
c. Newsfeeds or alerts
d. Application program interface (API)
e. XL or other markup language
f. Online dashboard

13) Do you have any additional comments or recommendations?

14) May we contact you in the future to ask for additional insights and share the results from the survey? (If yes, please provide your email)

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Hidden Risks to Cyberspace Security from Obsolete COTS Software

Barış Egemen Özkan
Plans Branch Head
Cyberspace Operation Center
Mons/Belgium
BarisEgemen.Ozkan@shape.nato.int

Serol Bulkan
Professor
Marmara University
Istanbul/Turkey
sbulkan@marmara.edu.tr

Abstract: Obsolescence of Commercial Off The Shelf (COTS) hardware and software, with their shorter product life cycles, is one of the major concerns for cyberspace system/service providers. While hardware obsolescence has been widely studied, software obsolescence has received less attention. However, the increased number of cyber incidents globally calls for more attention to the use of COTS software in critical infrastructures and military systems: systems comprising 25+ product life cycles and dominated by sustainment concerns. The number of reported vulnerabilities of COTS software systems more than doubled in 2017 and continued to increase in 2018. It is already a challenge for system/service providers to keep up with the pace of vulnerabilities to sustain the resiliency of the systems. Increased use of COTS software in mission-critical systems exacerbates the situation because it forces system/service providers to manage the risk of not being able to receive security updates for obsolete software. In today's cyber conflict, where hybrid threats are enjoying the highly connected nature of cyberspace terrain enabled with globalization and newer technologies, if cyberspace security risks stemming from obsolete COTS software in critical systems are not addressed properly, they may easily become a national security problem. Such risks must be addressed comprehensively at both governance and management levels. This paper presents the sustainability, operational efficiency and cyberspace security risks of obsolete COTS software in critical infrastructures and military systems and proposes mitigations at both governance and management levels. At the management level, a Multi Criteria Decision Making methodology is proposed for system/service providers to balance the conflicting objective functions of reaching a cost-effective solution while maximizing the system's cyberspace security and efficiency.

Keywords: *software obsolescence risks, COTS, vulnerabilities, cyber security, sustainment-dominated systems, cyber conflict*

1. INTRODUCTION

Following the end of the Cold War, the small-scale consumer products market share drastically increased and came to dominate the market (Singh and Sandborn, 2006). While military market share of semiconductors in the 1970s was 35% of a $4.2 billion market, it dropped to 0.3% of a $316 billion market in the 2000s (Kelly, 2017). The high speed of technological advances and the ease of access to markets, as well as the dynamic nature of consumer needs, has made product life cycles even shorter, down to 2-5 years (Shen and Willems, 2014). These developments created a new phenomenon called Commercial Off The Shelf (COTS) products (Sandborn, 2008). COTS are hardware and software products or services available in the market for public use (Özkan & Bulkan, 2018) and they are cheaper than custom designed products, usable for multiple environments with well-defined interfaces, most likely available from multiple vendors and usually have faster product upgrade cycles. With these attractive attributes, COTS products became the indispensable choice of system designers to achieve cost-effective solutions.

As well as the many advantages of using COTS products in sustainment-dominated systems, there are also some disadvantages. Sustainment-dominated systems, including military, transportation, aviation and nuclear systems, have an average of 25 years of product life cycles (Özkan & Bulkan, 2016). Military assets such as the B-52 bombers of the US Air Force have been in use since the 1950s (US Air Force, 2015). Using COTS hardware and software with shorter life cycles on such systems generates a sustainability risk if and when the original vendor declares obsolescence or end-of-life/support for those COTS parts.

Although there have been numerous studies on hardware COTS obsolescence, software COTS obsolescence has not been equally studied (Sandborn, 2007), despite the ever-increasing security vulnerabilities of COTS software. There are studies seeking cost-efficient methods to sustain the systems (Wnuk, Gorschek, and Zahda, 2013; Rojo et al., 2010; Munoz et al., 2015; S. Rajagopal, J.A. Erkoyuncu, 2015). The number of reported vulnerabilities on COTS software is increasing (CVEDetails, 2019a). With the current pace of the accumulation of vulnerabilities, owners of non-obsolete systems are already having difficulty patching their systems to ensure cyber security. In addition to this issue, not being able to receive security updates from original vendors at all due to obsolescence leads to a serious silent risk if those vulnerabilities are exploited by cyber threat vectors.

In today's cyber conflict, traditional military threats of armed forces have been overshadowed by hybrid threats, in and through which cyberspace is highly utilized. The whole spectrum of cyber threat actors, including state-sponsored ones, are exploiting the intense digitization and globalization enabled by new technologies; and they are not shy about employing their means to create effects to deny, disrupt and even destruct. Such hybrid threats are challenging classical defense strategies, attribution and deterrence concepts. Increased use of COTS software in National Critical Infrastructures (NCI) and military systems is expanding our vulnerability surface for attackers to exploit, which may become a national security issue if not properly addressed.

In this paper the COTS software obsolescence section provides foundations with definitions of obsolescence in general and software obsolescence, and explains why we continue to use COTS software despite the disadvantages. The next section explains the cyber security risks of using obsolete COTS software, including those stemming from supply chain, and their impacts. This section also provides descriptive findings of several COTS software applications, still in use in numerous enterprises, for which the original vendor has already declared end-of-life and no longer provides support. The following section lays down the recommended mitigations for these risks. In the model section, a methodology is proposed to address the COTS software obsolescence with competing objectives including minimum cost, maximum operational availability and maximum cyber security. This section also suggests a practical approach for the proposed model. The last section sums up with recommendations and conclusions for the cyberspace security risks of COTS software obsolescence on NCI and military systems.

2. WHAT: COTS SOFTWARE OBSOLESCENCE

A. Obsolescence Defined

Obsolescence is the condition of no longer being used or useful, of being obsolete. The state of being obsolete may be voluntary or involuntary (Bartels et al., 2012). With voluntary obsolescence, the manufacturer plans the obsolescence and voluntarily stops support to shorten the repetitive purchase cycles. With involuntary obsolescence, however, neither producer nor consumer has intentions for such obsolescence (Sandborn and Myers, 2008).

Obsolescence may be due to logistical, functional or technological reasons. Logistical obsolescence is due to loss of ability to procure parts, material, manufacturing or software necessary to manufacture or support a product (Bartels et al., 2012), such as termination of access to software due to digital media obsolescence, formatting, or degradation (Sandborn, 2007). In functional obsolescence, the product still meets the functional requirements of the original design; however, the requirements or environmental factors have changed over time and the functionality that the product meets is no longer relevant. In technological obsolescence, a new product is delivered to the customer due to several possible reasons such as increase in capacity or processing power; it supersedes the older one. This type of obsolescence is the most common in information technology, such as CDs superseding floppydisks and DVDs superseding CDs (Özkan & Bulkan, 2016).

B. Software Obsolescence

While hardware obsolescence is better-known and more studied than software obsolescence, they must be considered together since they tightly depend on and affect one another. Software obsolescence occurs when the original vendor stops support, updates, upgrades and fixes for known bugs, which eventually makes the software unusable for consumers (S. Rajagopal, J.A. Erkoyuncu, 2015).

One of the fundamental drivers of COTS software obsolescence in information technology is the fear of losing market share, as was clearly stated by Bill Gates (Merola, 2006): “The only big companies that succeed will be those that obsolete their own products before someone else does”.

Another major reason for software obsolescence is the fast-degrading quality of software. The quality of software depends on its ability to meet consumer expectations. As consumers can rapidly change their requirements, partly due to the swiftly changing nature of the business environment and partly due to consumers’ lack of ability to specify their requirements clearly upfront, some requirements become obsolete even when the product is still in-house for design and development. For those reasons, vendors tend to deliver products within quicker schedules and issue more frequent updates to mitigate the quality defects of software. After a certain point, it becomes much more viable for vendors to cease support, declare a product obsolete and release a completely new product. This is a good business strategy for vendors, but definitely not for sustainment-dominated system owners who have to keep them up and running for many more years.

C. Why Continue to Use Obsolete COTS Software?

Despite these facts, military organizations continue to use COTS software. Many nations' military procurement strategies strongly support the use of COTS hardware and software over custom solutions. Following US Secretary of Defense William Perry's 1994 initiative, many nations started drifting away from the use of military specifications and began preferring COTS-based solutions (Gansler and Lucyshyn, 2008; Ministry of Defence, 2005; Turkey, 2010; US Navy, 2000).

Microsoft officially ended support for Windows XP in April 2014 after releasing its last major update in 2008. According to Netmarketshare.com (2019), 4.1% of all desktop users are still using Windows XP (around 80 million computers), which means that they are susceptible to security vulnerabilities that will never be fixed by the vendor.[1] Windows XP was 13 years old when it was declared obsolete and it still has more market share than Windows 8, Linux and many Mac OS versions.[2] According to CVEDetails, over 740 Windows XP vulnerabilities remain identified but unpatched (CVEDetails, 2019b). However, many systems, including Automated Teller Machines, schools, police stations, electronic voting machines, transportation systems, airport security systems and even casinos are still using Windows XP. In addition, there have been reports of military systems, including warships, still using this operating system (SpiceWorks, 2017). But the question is: why do individuals, companies and even nations continue to use obsolete software with known vulnerabilities that will never be fixed?

The first reason is that it still works. The complacency created by the software-in-use for many years is one of the major factors to continue with it rather than replacing it. The second reason is the need for efficiency. Hardware and software upgrades go hand-in-hand. However, such upgrades are not always feasible for sustainment-dominated systems and service providers due to the high costs of re-design, adaptation, implementation, re-certification and training. The time and cost implications for organizations to upgrade both software and hardware can become very complex to resolve. In addition, obsolete COTS software that has been in use for a long period may have led to numerous deep dependencies in complex systems and support arrangements. The need for compatibility among the systems forces their owners to continue with obsolete COTS software (Lapham and Woody, 2006).

[1] One and probably the last exception for this postulation was the WannaCry incident for which Microsoft issued a security update for Windows XP after the company ended its support.

[2] This premise is valid for the data retrieved in January 2019.

3. SO WHAT? THE RISKS OF OBSOLETE COTS SOFTWARE

A. Sustainability Risks

Due to their advantages, COTS software products are widely used in almost every part of our lives; including NCIs, government, military, personal systems. However, with their shorter life cycle, the obsolescence of COTS software at least creates a sustainability risk with significant impact on operations and maintenance costs.

Studies on the sustainability risks of COTS software obsolescence have focused on cost impacts and the techniques in-service to mitigate those impacts (Morris, 2000; Abts, Boehm, and Clark, 2000; Mckinney, 2001; Comella-Dorda et al., 2004; Sandborn, 2007; Rojo et al., 2010). These mitigation techniques include data preservation, managed integration, reengineering, reverse engineering, software license downgrade and redevelopment (Sandborn, 2007). Applying these techniques, however, incurs mitigation, redevelopment, requalifying, re-hosting and media management costs.

B. Operational Efficiency Risks

What is not as much studied as the sustainability risks of using COTS software is the impact on operational efficiency. The reasons for continuing to use obsolete COTS software create operational efficiency risks to the reliability, availability and maintainability (RAM) of the systems.[3] RAM is considered to be one of the major quality metrics of COTS products to measure operational efficiency.

The concept of RAM was developed predominantly for hardware systems, and parameters to measure RAM can easily be found in the datasheets of COTS hardware. Unfortunately, software systems do not enjoy the same level of predictable performance in their datasheets. While there are models to measure software quality with evaluation criteria and quality aspects which also include the RAM of the software (Miguel, Mauricio and Rodríguez, 2014), the features to measure the software RAM metrics are usually stated in the non-functional requirements section of the system specifications. However, they are habitually left blank or weakly specified due to the aggressive market conditions for the COTS software; hence they are rarely if ever tested thoroughly. Consequently, the measurement of RAM metrics for software COTS systems is generally left to the service-in-use phase of the software. Vendors look forward to verification and thorough certification tests by consumers during

[3] Reliability refers to the measure of the probability that failures will occur during operation of a system (PioneerEngineering, 2017). In other words, it is the probability of a system's ability to perform its intended function under defined conditions for a specified time interval without failure (ReliabilityWeb, 2018). Availability is the measure of a system's readiness for operation at a given time under given environmental conditions and is usually measured as point availability (Sebok, 2018). Operational availability is a slightly different term used in military literature to define the ratio of uptime to total time. Uptime is measured by adding standby, mission, relocation, pre-operation tests and operating times. Total time is the sum of uptime and maintenance time, which is composed of corrective and preventative maintenance activities (Pryor, 2008). Maintainability is the probability of being able to repair a system, in other words to perform corrective and preventative maintenance measures in a specified environment within a defined period of time (Sebok, 2018).

operational use and expect them to report the identified bugs and vulnerabilities for vendors to provide fixes via after-sale upgrades.

The downtime for obsolete COTS software is likely lengthy and, under severe conditions, mean time to repair (MTTR) becomes notionally infinite if the obsolete software is tightly dependent on another software component which has been upgraded without backward compatibility. The high figures of downtime and MTTR decrease the availability of the system.

When systems with obsolete COTS software are in use for operations, due to unfixed bugs which have been identified after obsolescence they will often stay longer in downstate or degraded and it will decrease the reliability of the system with decreased mean time to failure and increased MTTR figures. Those systems also suffer maintainability risks due to lack of vendor support to fix the problems that have been identified after obsolescence.

C. Cyber Security Risks

In addition to RAM, one other quality metric for software is cyber security (Altexsoft, 2017). Security refers to the protection of a system from inadvertent or malicious activity that could impair the confidentiality, integrity and accessibility of the data, service or function (Miguel, Mauricio and Rodríguez, 2014). As one of the major drawbacks of COTS software, not being able to fully specify cyberspace security requirements keeps increasing its vulnerabilities.

The number of vulnerabilities on information technologies utilizing COTS software continues to follow an ever-increasing trend. We have seen a conspicuous increase in 2017 with 14,714 identified vulnerabilities, and 2018 did not fall short either with 15,703 identified vulnerabilities[4] (See Figure 1). Both 2017 and 2018 vulnerability threat trends indicate that the scale of threat is increasing on internet-connected and mobile devices. Almost all internet and mobile devices software are COTS (Flexera, 2017; 2018) and pose a significant security risk.

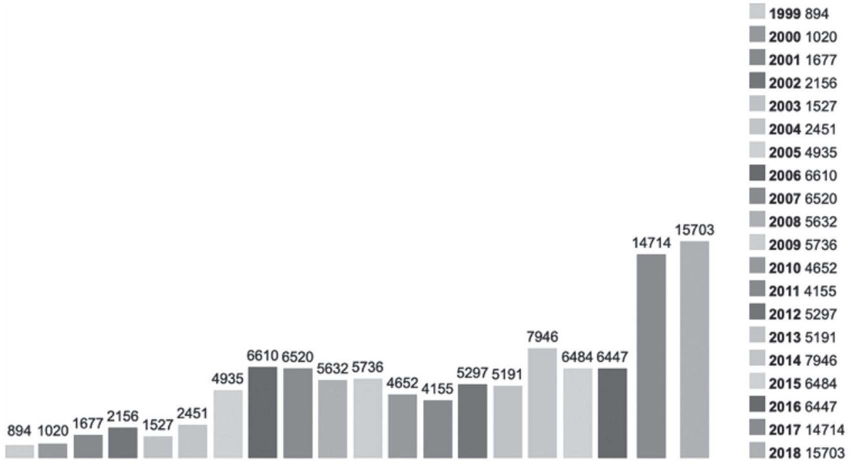
Using obsolete COTS software in systems, and particularly in NCIs, lowers the overall profile of cyberspace security. This is an especially significant concern for vulnerabilities which are discovered after the COTS software is declared obsolete. According to US-CERT, COTS software is risky to use because, compared to custom code, it is a very attractive point of attack: generic, well-known and widely available. Since COTS software comes as a black-box, it is no trivial exercise to mathematically model the security of COTS software and verify it. In addition, COTS software vendors are rarely held liable for direct and consequential damages (US-CERT, 2013).

[4] The significant increase is partly due to the actual rise in the vulnerabilities and partly due to enhanced cyber security awareness and maturity in consumers, original vendors and the third-party supply chain. The escalation of such awareness and interest has been an increased incentive to search for vulnerabilities (Flexera, 2018).

Software reuse is an effective strategy using existing working components rather than reinventing the wheel. This strategy brings higher yields, productivity, quality, lower costs and shorter time-to-market (Lee, 2003). It is not rare to see software components that were once used in an older version of software in the newer builds. The component reuse strategy in software design inevitably increases the number of unpatched vulnerabilities on the earlier version of the system at which a vulnerable component from obsolete software is reused in the newer versions. The obsolete software will cease to receive security updates from the vendor and this will increase the probability of exploitation of those later-found vulnerabilities. For example, Microsoft introduced New Technology File System (NTFS) withWindows NT 3.1 and still uses it as its primary file system in Windows 10. Another example is Microsoft's Internet Information Services (IIS) which was first introducedwith Windows NT 3.5.1. and is still used in Windows Server 2019 and Windows 10. Microsoft has declared end-of-life for some of those operating systems and no longer provides updates and patches (Microsoft, 2019).

A thorough survey conducted by these researchers on a particular service provider's approved product list revealed that a number of software applications providing support to critical operational functions are still using COTS software that is already announced as end-of-support by the original vendor. The list includes earlier versions of Adobe Flash Player (v29.0.0.113), Oracle database (11g), Microsoft Office (2007), and Microsoft SharePoint (2007), all of which are beyond their end-of-life and do not receive security patches from their vendors. Readers are advised to refer to many vulnerability exploits against obsolete software in Qualys lists (Qualys, 2019).

FIGURE 1. VULNERABILITIES PER YEAR (CVEDETAILS, 2019A)



With the current pace of the accumulation of vulnerabilities, system owners are already having difficulty patching up their systems. Slow configuration management processes to test patches for safety and interoperability are adding additional difficulty to the timely addressing of vulnerabilities. Finally, except for planned obsolescence, software obsolescence is not very predictable, hence, proactive management strategies developed for hardware obsolescence are not readily adaptable for software obsolescence.

Not being able to receive security updates from original vendors for those known vulnerabilities due to obsolescence leads to a serious silent risk. Especially in today's cyber conflict, cyber threat actors are very talented at disguising themselves in a hybrid threat environment inside highly complex cyberspace terrain. Considering that almost all the published breaches in recent years exploited known vulnerabilities (Gartner, 2017), increased vulnerabilities that will never be patched lead to a high probability of being exploited. The impact of exploitation depends on the system's characteristics. For example, consider a vulnerability stemming from an obsolete COTS software in a maritime harbor, railway or airport system which the military is planning to use during deployment for an operation. Imagine a power plant providing electricity to whole city, a SCADA used in a nuclear power plant or an electronic voting system still using Windows XP. If it is for NCIs, government or military systems, exploitation of those vulnerabilities will easily lead to a national security problem which must not be left alone but instead must be addressed and mitigated thoroughly.

Another risk vector for obsolete COTS software stems from the supply chain. COTS software creates inevitable dependency on certain vendors in the supply chain due to convenience and price advantage. If those vendors are from a foreign country that might be part of possible future conflicts, they can be tempted by foreign intelligence services to create backdoors through silent but deliberate planned obsolescence.

4. NOW WHAT? MITIGATING THE RISKS OF OBSOLETE COTS SOFTWARE

A. Governance-Level Mitigation

Due to the wide spectrum of impacts spanning from personal to national security, the cyber security risks of obsolete COTS software must be addressed with a comprehensive approach at all levels. At the strategic level, a whole-government approach is recommended to provide guidance for obsolescence risks in national cyber security policies. Additionally, while not very trivial, proactive risk mitigation strategies yield more efficiency than reactive models and require a systematic holistic approach as well.

Having reviewed all of the available national defense and cyber security strategies in open sources, it is common for all nations to draw attention to the protection of NCI and the risks of commercialization with intense digitization. However, with a few exceptions, none of those strategies is explicitly referring to the risks of obsolete COTS software on cyber security. Only a white paper on the defense of the Slovak Republic mentions obsolescence for its impacts on the military forces' readiness (Slovak Republic MoD, 2016). While the US Department of Defense's cyber strategy promotes the leverage of COTS capabilities, the US Department of Homeland Security's cyber security strategy explicitly points out the supply chain risks of COTS products.

As nations are becoming more global, connected and digitized, they become more fragile against cyber threats. Considering all the risks of obsolete COTS software mentioned above, it is a much bigger concern for those nations with a larger cyberspace footprint. For that reason, nations are advised to address the risks associated with obsolete COTS software in their cyber security policies, strategies and directives.

In addition, a central cross-domain consultant agency at governance level for public and private institutions may play a significant role to ensure a coherent mitigation approach among government, industry and military organizations. This central agency would also negotiate with original vendors to delay obsolescence, with certain incentives provided by government if needed.

One of the ways to achieve protection of cyberspace as described in national cyber security strategies is through effective relationships and close cooperation with science and technology organizations and academia. It is good practice to exploit the academic relationships and provide them with guidance to improve forecast models for the non-deterministic nature of software obsolescence. Such models will definitely improve proactive obsolescence management strategies.

In the intensely connected nature of cyberspace, where public and private organizations are mutually enabling and supporting each other, governance-level initiatives to enforce cross-domain mapping of all systems in a whole-of-government framework would ease the COTS obsolescence risk management activities at management level through informed assessment, prioritization and resource allocation.

The US Defense Standardization Program Office has been the custodian of a document: "SD-22 – Diminishing Manufacturing Sources and Material Shortages (DMSMS)" (Defense Standardization Program Office, 2016). It is a guidebook of "Best Practices for Implementing a Robust DMSMS Management Program" and primarily intended to provide program managers with increased awareness, robust obsolescence management processes, metrics for effective measurement and best practices. This document is mainly focused on cost-efficient solutions and slightly touching the cyberspace security risks of obsolete software. Increasing the accounts of the weight of cyber security risks stemming from use of obsolete COTS software and relevant mitigations in this prime reference document for COTS obsolescence will help to increase the awareness among program managers.

Additionally, keeping a definitive list at governance level of all foreign countries in the supply chain that may use software obsolescence as a way to create deliberate cyber security vulnerability in products can be a mitigation for the risks, especially on NCI and military systems.

B. Management Level Mitigation

In order to mitigate security impacts of obsolete software, the UK National Cyber Security Centre (NCSC) recommends to migrate away from them and apply short term mitigations (NCSC, 2017). NCSC recommends using only products still supported by the vendor. When the original vendor declares the obsolescence date, system owners are strongly advised to plan for prioritized migration to newer software. In order to mitigate the security impacts of obsolescence, systems owners are advised to not accept the new developments that will run on obsolete software and to reduce the dependencies on the obsolete software. System owners are also advised to decrease either the probability of the exploitation of the unpatched vulnerabilities or the impact of exploitation if the system is compromised. One method is to prevent malicious code or data access to the obsolete system from outside. This can be done by isolating the system from the enterprise and preventing or at least reducing the system access to untrusted services.

US-CERT recommends practicing a holistic approach to achieve a comprehensive mitigation of the risks stemming from COTS software (US-CERT, 2013). Since there is no such thing as 100% cyber security, it is fair to assume that there will always be more vulnerability than system owners can address. Therefore, the whole cyber security business is based on risk management and it aims to achieve and maintain cyber resilience for better defense, detection, response and recovery with a cognizant prioritization schema.

Mission mapping via thorough asset management is key to a resilience framework. It starts with identifying the COTS components and mapping them to information, functional services, processes and ultimately the critical outputs. The second step is to identify critical points in the mission map for an informed prioritization and cost-effectiveness. Some of those critical points are single points of failures, choke-points, entry-exit points to critical infrastructure, servers interfacing with the outside world, data centers and core systems for critical mission outputs. Identification of critical points is better achieved through continuing discussions among service providers who own the COTS software and IT infrastructure, the security community and the user community. Once the prioritization is sorted, the next step is to secure both COTS software and hardware to reduce the cyber security risks of using them. Increased redundancy for identified mission-critical points, preferably with different COTS products, is another risk mitigation method. Wrapping the COTS software that is mapped to mission-critical outputs to ensure that it will do only what it is supposed to do is an additional risk mitigation process.

Comparing these two big security organizations, NCSC's recommendations can be considered mostly procedural in order to maximize cyber security. US-CERT's recommendations are much more fit for a mission assurance framework to maximize operational efficiency. In contrast, almost all of the academic studies in the literature, as mentioned above, have focused on minimizing the sustainability cost. All of these approaches are right but not complete without each other. Migration from obsolete COTS software and the selection of mitigation techniques are complex problems for decision-makers. The better solution is a balanced approach to meet all of the objectives with risk informed trade-offs.

C. Balanced Model for Management Level Mitigation

There are at least three objectives for decision: minimize the cost, maximize the cyber security and maximize (or at least sustain) the operational efficiency. The contradictory nature of those objectives makes the decision making more complicated. First of all, high security comes with a significant bill. As we try to minimize the costs by abstaining from certain reactive mitigation measures, we may end up with decreased cyber security. On the other hand, as we try to maximize cyber security we may face with the discontinuity of services and hence decreased efficiency with a considerable impact on mission outputs. Therefore, applying mitigation for COTS software obsolescence is not a single objective decision making process but rather a Multi Criteria Decision Making (MCDM) problem with conflicting objectives.

In MCDM methodology, the decision maker has to use trade-offs and satisfy all objectives with the best effort based on his/her subjective criteria and preferences. The criteria reflect the desires of the decision maker, which points the direction to a better solution (Ehrgott & Xavier Gandibleux, 2002).

The aim of the MCDM is to define a set of candidate solutions in the problem space which will produce representative objective values in the solution space. The latter set in the solution space is called the Pareto Front (Özkan and Bulkan, 2018) and it holds the Pareto optimal solutions.

An improvement in one of the objective functions can only be achieved for Pareto optimal solutions if at least one other objective function's value is worsened. In that case, objective functions can be improved in value at the expense of degrading at least one other objective function's value.[5]

D. Practical Application of MCDM Approach

As listed above, the three objectives in mitigating the risks of obsolete COTS software are minimizing the costs, maximizing the operational efficiency and maximizing the security. Those three objectives are subject to a set of constraints. The first two constraints are resources for budget and time. The third constraint is the Key Performance Indicators (KPI) of the mission outputs derived from the mission mapping.

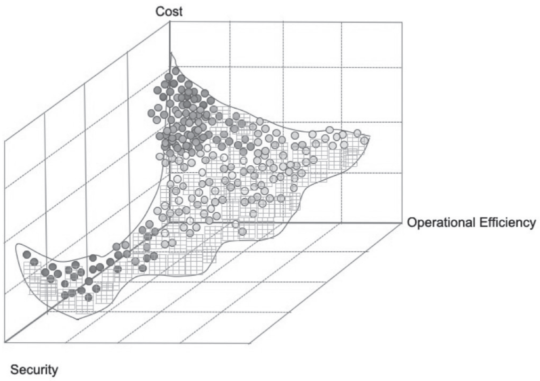
- $min\ \{cost(\bar{x})\}$
- $max\ \{operational\ efficiency\ (\bar{x})\ \}$
- $max\ \{security\ (\bar{x})\}$

subject to

- $budget(\bar{x}) \leq BUDGET$
- $time(\bar{x}) \leq TIME$
- $kpi(\bar{x}) \leq KPI$

BUDGET, TIME and KPI are model parameters and represent the constraints of resource implications and efficiency requirements. Decision variables are mitigation activity on obsolete COTS software and time of implementation. This model will produce a number of Pareto solutions in three-dimensional objective space, as shown in Figure 2.

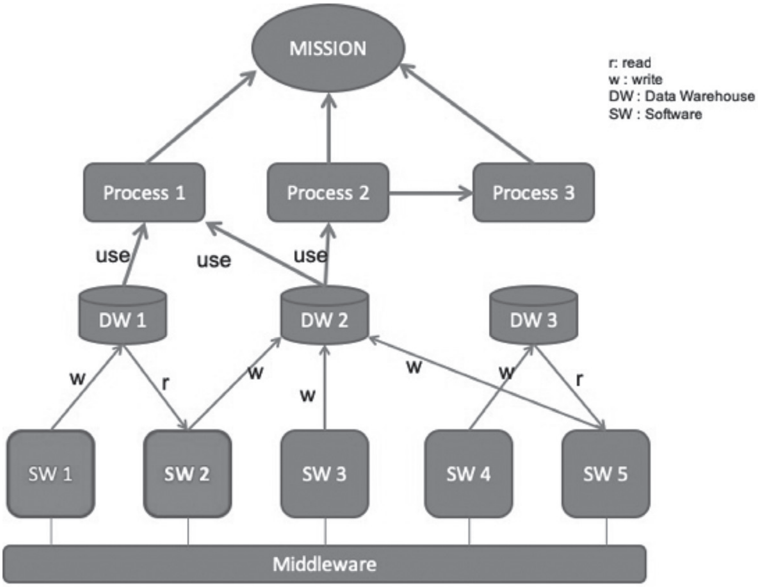
FIGURE 2. PARETO FRONT FOR THREE OBJECTIVES



[5] Implementation details and mathematical programming of MCDM problems is beyond the scope of this research. Interested readers are advised to follow Ehrgott and Xavier Gandibleux (2002) and Figueira, Greco, and Ehrgott (2005). A thorough study on the implementation of MCDM by evolutionary algorithms can be found in Özkan and Bulkan (2018).

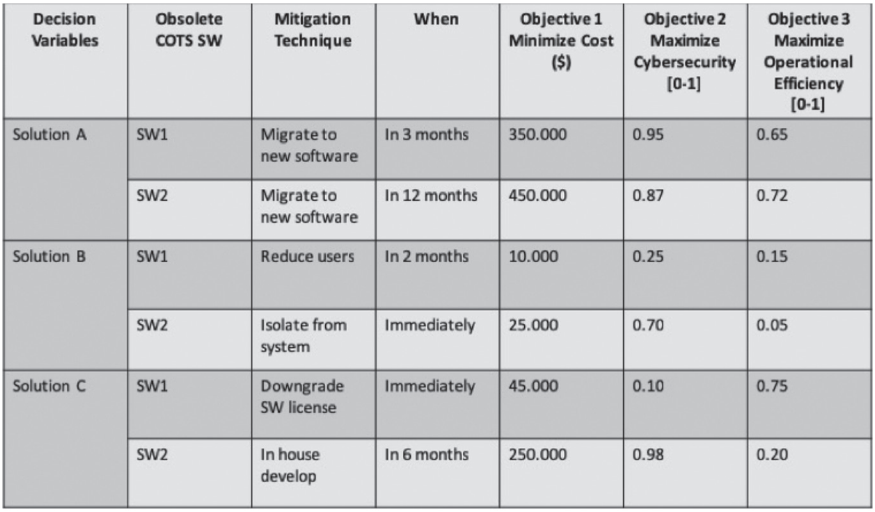
For the practical implementation of an MCDM problem, consider a system with mission mapping as given in Figure 3. In order to achieve mission objective, threeprocesses are used. Process 1 uses data from the first and second warehouses which all five software applications are either reading or writing. Process 2 uses data fromthe second data warehouse and only second, third and fourth software applications are reading and writing. Process 3 is accepting input directly from Process 2 with no connection to data warehouses. Consider that SW1 and SW2 are called obsolete. In this practical implementation, operational efficiency is measured by weighted product of reliability, availability and maintainability metrics of each SW and data warehouse on mission mapping. The cyber security value is measured by a function of unpatched vulnerabilities. Cost is the parametric value of each mitigation activity adjusted with inflation rate depending on the implementation time.

FIGURE 3. MISSION MAPPING OF A SYSTEM



Each of those solutions within the Pareto set represents an ordered triple of cost, security and operational efficiency objective values (See Figure 4). In this Pareto set, three optimal solutions with varying combinations of mitigation technique and time to implement mitigations are found. Each of those Pareto solutions yields to different objective values in the solution space. It is up to the decision maker to select a solution from the Pareto set based on his or her preferences.

FIGURE 4. OPTIMAL PARETO SET



Decision Variables	Obsolete COTS SW	Mitigation Technique	When	Objective 1 Minimize Cost ($)	Objective 2 Maximize Cybersecurity [0-1]	Objective 3 Maximize Operational Efficiency [0-1]
Solution A	SW1	Migrate to new software	In 3 months	350.000	0.95	0.65
	SW2	Migrate to new software	In 12 months	450.000	0.87	0.72
Solution B	SW1	Reduce users	In 2 months	10.000	0.25	0.15
	SW2	Isolate from system	Immediately	25.000	0.70	0.05
Solution C	SW1	Downgrade SW license	Immediately	45.000	0.10	0.75
	SW2	In house develop	In 6 months	250.000	0.98	0.20

The security community tries to maximize the security and minimize the risks while the user community tries to maximize the operational efficiency and mission outputs, and the service providers try to minimize the costs to increase their profit. This model enables the risk owner to select a Pareto solution for obsolete COTS software that balances the benefits of the three communities of interest.

5. CONCLUSION

COTS software is very appealing to use in complex systems, with numerous advantages such as cost and availability. However, due to market conditions, the product life cycle of such COTS software is very short compared to the longer product life cycles of sustainment-dominated systems, including military ones. Even though there are considerable disadvantages to using COTS software in military systems, the advantages outweigh them and national defense acquisition agencies continue to use COTS products. Increased use of COTS products makes them the capillaries of our NCIs and military systems and is expanding the vulnerability surface for attackers to exploit.

Increased use of COTS software in critical systems with longer product life cycle at minimum leads to sustainability costs, operational efficiency and cyberspace security risks. The sustainability costs are reactive or proactive mitigation costs due to the impacts of obsolescence. The operational efficiency risks are due to reliability, availability and maintainability risks stemming from obsolescence. The cyber security risks are the unaddressed vulnerabilities of obsolete software.

The security risks of using COTS software have significantly increased within the last two years, as we have seen a considerable increase in the number of reported vulnerabilities in COTS software. Considering the vast amount of existing vulnerabilities in COTS software, the risks associated with obsolete COTS software used in NCI and military systems are highly likely to have a considerable impact if not properly addressed.

In today's cyber conflict, cyber threat vectors have increased their competency and capacity to develop malicious activity for COTS software as well as their intention to use them. This makes the obsolete COTS software a significant element of cyber conflict. Such risks of cyberspace security may easily escalate to a national security issue if not properly addressed.

The cyber security challenges of the obsolete COTS software must be addressed holistically at both government and management levels. At the governance level a comprehensive whole-of-government approach must be pursued. National defense and cyber security strategies and directives are advised to explicitly include hidden risks of COTS obsolescence against NCI and the supply chain. Vendors from foreign countries for systems used in NCI and military systems must be especially closely monitored. A cross-domain central agency between public and private would serve to provide different clusters of organizations with best practices and common approaches. At the management level, each program manager or mission owner must balance the cost, security and operational efficiency objectives within a risk informed trade-off framework. Since those objectives conflict with each other, a MCDM methodology is proposed to find Pareto optimal solutions for all objectives. Decision-makers are compelled to manage the risks within a framework by balancing the needs of the security community, mission owners and the service providers in order to minimize the cost of services, maximize security and maximize operational efficiency.

REFERENCES

Abts, C., Boehm, B., & Clark, E. (2000). COCOTS: A COTS software integration lifecycle cost model-model overview and preliminary data collection findings. *ESCOM-SCOPE Conference*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.8295&rep=rep1&type=pdf

Altexsoft. (2017). What Software Quality (Really) Is and the Metrics You Can Use to Measure It. Retrieved January 5, 2019, from https://www.altexsoft.com/blog/engineering/what-software-quality-really-is-and-the-metrics-you-can-use-to-measure-it/

Bartels, B., Ermel, U., Pecht, M., & Sandborn, P. (2012). *Strategies to the Prediction, Mitigation and Management of Product Obsolescence. Strategies to the Prediction, Mitigation and Management of Product Obsolescence*. https://doi.org/10.1002/9781118275474

Comella-Dorda, S., Dean, J., Lewis, G., Morris, E., Oberndorf, P., & Harper, E. (2004). A process for COTS software product evaluation. *COTS-Based Software Systems*, (July), 86–96. https://doi.org/10.1007/3-540-45588-4_9

CVEDetails. (2019a). Vulnerabilities By Year. Retrieved January 5, 2019, from https://www.cvedetails.com/browse-by-date.php

CVEDetails. (2019b). Windows XP Vulnerability Statistics. Retrieved January 5, 2019, from https://www.cvedetails.com/product/739/Microsoft-Windows-Xp.html?vendor_id=26

Defense Standardization Program Office. (2016). SD-22 – *Diminishing Manufacturing Sources and Material Shortages (DMSMS) A Guidebook of Best Practices for Implementing a Robust DMSMS Management Program Defense Standardization Program Office*.

Ehrgott, M., & Xavier Gandibleux. (2002). *Multiple Criteria Optimization, State of the Art Annotated Bibliographic Surveys*. Kluwer Academic Publisher.

Figueira, J., Greco, S., & Matthias Ehrgott. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer.

Flexera. (2017). *Vulnerability review 2017*. Retrieved from https://resources.flexera.com/web/pdf/Research-SVM-Vulnerability-Review-2017.pdf

Flexera. (2018). *Vulnerability Review 2018*. Retrieved from https://resources.flexera.com/web/pdf/Research-SVM-Vulnerability-Review-2018.pdf

Gansler, J. S., & Lucyshyn, W. (2008). *Commercial off the Shelf (COTS)*. https://doi.org/10.1146/annurev.anthro.29.1.107

Gartner. (2017). Focus on the Biggest Security Threats, Not the Most Publicized. Retrieved from https://www.gartner.com/smarterwithgartner/focus-on-the-biggest-security-threats-not-the-most-publicized/

Kelly, S. (2017). Obsolescence Management in Long Term Projects. Retrieved January 5, 2019, from Obsolescence Management in Long Term Projects, 2017, World Codification Forum

Lapham, M. A., & Woody, C. (2006). Sustaining software-intensive systems, (May), 53.

Lee, E. (2003). Software reuse and its impact on Productivity , Quality and Time To Market, 1–6. Retrieved from https://pdfs.semanticscholar.org/08f2/af486985e34d4a46da42b8b4c322c7c22571.pdf

Mckinney, D. (2001). Impact of Commercial Off-The-Shelf (COTS) Software and Technology on Systems Engineering.

Merola, L. (2006). The COTS software obsolescence threat. *Proceedings - Fifth International Conference on Commercial-off-the-Shelf (COTS)-Based Software Systems, 2006*(Iccbss), 127–133. https://doi.org/10.1109/ICCBSS.2006.29

Microsoft. (2019). Microsoft Product Lifecycle. Retrieved January 5, 2019, from https://support.microsoft.com/en-ie/lifecycle/search?alpha=Microsoft Windows 2000 Server

Miguel, J. P., Mauricio, D., & Rodríguez, G. (2014). A Review of Software Quality Models for the Evaluation of Software Products. *International Journal of Software Engineering & Applications (IJSEA), 5*(6), 31–53. Retrieved from https://arxiv.org/pdf/1412.2977.pdf

Ministry Of Defence. (2005). Defence Industrial Strategy Defence Values for Acquisition. *Defence Studies, 8*(3), 286–310. https://doi.org/10.1080/14702430802252545

Morris, A. T. (2000). COTS score: An acceptance methodology for COTS software. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings, 1*, 4.B.2-1-4.B.2-8. https://doi.org/10.1109/DASC.2000.886948

Munoz, R. G., Shehab, E., Weinitzke, M., Bence, R., Fowler, C., Tothill, S., & Baguley, P. (2015). Key challenges in software application complexity and obsolescence management within aerospace industry. *Procedia CIRP, 37*, 24–29. https://doi.org/10.1016/j.procir.2015.08.013

NCSC. (2017). Obsolete platforms security guidance. Retrieved January 5, 2019, from https://www.ncsc.gov.uk/guidance/obsolete-platforms-security-guidance

Özkan, B., & Bulkan, S. (2016). COTS Parts Obsolescence Management of Sustainment Dominated Military Systems. In *10th NATO Operations Research and Analysis (OR&A)* Conference (pp. 1–18). https://doi.org/10.14339/STO-MP-SAS-OCS-ORA-2016

Özkan, B. E., & Bulkan, S. (2018). Obsolescence Management for Sustainment-Dominated Military Systems: Multiple Criteria Decision-Making Approach Using Evolutionary Algorithms. In *Operation Researches for Military Organizations* (p. 20). https://doi.org/10.4018/978-1-5225-5513-1

PioneerEngineering. (2017). Availability vs. Reliability Part 1. Retrieved January 5, 2019, from https://www.pioneer-engineering.com/resources/availability-vs-reliability-part-1

Pryor, G. A. (2008). Methodology for Estimation of Operational Availability as Applied to Military Systems. *ITEA, 29*(4), 420–428.

Qualys. (2019). Exploits Against Obsolete Software. Retrieved September 20, 2003, from https://www.qualys.com/research/exploits/

ReliabilityWeb. (2018). Understanding the Difference Between Reliability and Availability. Retrieved January 5, 2019, from https://reliabilityweb.com/tips/article/understanding_the_difference_between_reliability_and_availability/

Rojo, F., Roy, R., Shehab, E., & Cheruvu, K. (2010). Key Challenges in Managing Software Obsolescence for Industrial Product-Service Systems (IPS2). *CIRP IPS2 Conference*, 393–398. Retrieved from http://www.ep.liu.se/ecp/077/050/ecp10077050.pdf

S. Rajagopal, J.A. Erkoyuncu, R. R. (2015). Impact of Software Obsolescence in Defence Manufacturing Sectors. *Procedia CIRP, 28*, 197–201. https://doi.org/10.1016/j.procir.2015.04.034

Sandborn, P. (2007). Software obsolescence-Complicating the part and technology obsolescence management problem. *IEEE Trans on Components and Packaging Technologies, 30*(4), 886–888. https://doi.org/10.1109/TCAPT.2007.910918

Sandborn, P. (2008). Trapped on Technology ' s Trailing Edge. *IEEE Spectrum*, (May), 1–6.

Sandborn, P., & Myers, J. (2008). Designing engineering systems for sustainability. *Handbook of Performability Engineering*, 81–104. Retrieved from http://link.springer.com/chapter/10.1007/978-1-84800-131-2_7

Sebok. (2018). Reliability, Availability, and Maintainability. Retrieved January 5, 2019, from https://www.sebokwiki.org/wiki/Reliability,_Availability,_and_Maintainability

Shen, Y., & Willems, S. P. (2014). Modeling sourcing strategies to mitigate part obsolescence. *European Journal of Operational Research, 236*(2), 522–533. https://doi.org/10.1016/j.ejor.2014.01.025

Singh, P., & Sandborn, P. (2006). Obsolescence Driven-Design Refresh Planning For Sustainment-Dominated Systems. *The Engineering Economist, 51*(2), 115–139.

Slovak Republic MoD. (2016). *White Paper on Defence Of the Slovak Republic*.

SpiceWorks. (2017). 10 Computer Systems Still Using Windows XP in 2017. Retrieved January 5, 2019, from https://community.spiceworks.com/topic/2010831-10-computer-systems-still-using-windows-xp-in-2017-three-years-after-eos

Turkey. (2010). 2012-2016 Stratejik Plani, 53, 160. https://doi.org/10.1017/CBO9781107415324.004

U.S. Air Force. (2015). B-52 Stratofortress. Retrieved January 5, 2019, from https://www.af.mil/About-Us/Fact-Sheets/Display/Article/104465/b-52-stratofortress/

US-CERT. (2013). Security Considerations in Managing COTS Software. Retrieved January 5, 2019, from https://www.us-cert.gov/bsi/articles/best-practices/legacy-systems/security-considerations-in-managing-cots-software

U.S. Navy. (2000). Commercial-off-the-Shelf Policy.

Wnuk, K., Gorschek, T., & Zahda, S. (2013). Obsolete software requirements. *Information and Software Technology, 55*(6), 921–940. https://doi.org/10.1016/j.infsof.2012.12.001

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Applying Indications and Warning Frameworks to Cyber Incidents

Bilyana Lilly
RAND Corporation
Santa Monica, California, United States
blilly@rand.org

Lillian Ablon
RAND Corporation
Santa Monica, California, United States
lablon@rand.org

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

Adam S. Moore
RAND Corporation
Santa Monica, California, United States
amoore@rand.org

Abstract: Despite significant advancements in academia and public policy on identifying, deterring, and mitigating cyber incidents, there is a general discontent among NATO agencies, member states' governments, and intelligence agencies that their strategy against cyber incidents is primarily reactive and implemented *post factum*, rather than proactive and executed before such attacks occur. This issue could be addressed through the design and application of appropriate indications and warning (I&W) frameworks for the cyber domain. Currently, there is a lack of comprehensive understanding and generally accepted practice of how governments and international organizations can apply such I&W methodologies and integrate them with their existing capabilities and processes. A survey of the classic warning methodologies used by the U.S. intelligence community to address a range of non-cyber threats can inform the design of such robust frameworks. These mature intelligence methods can be adapted and perfected to adequately address threats in cyberspace. In this article, we examine some of these I&W frameworks and propose a high-level practical approach to cyber I&W that governments, NATO agencies and the private sector can use to design and structure their prevention, detection, and response mechanisms in order to effectively anticipate and defend against cyber threats. To demonstrate the utility of this approach, we apply it to an actual case: the November 14, 2018 spear-phishing campaign by Russia's APT29 against U.S. government agencies, think tanks, and businesses.

Keywords: *indications and warning, cyber warning, warning framework, threat intelligence, cyber, I&W, ATT&CK, APT29, cyber threat intelligence*

1. INTRODUCTION

In light of rapidly evolving technology and cyber threat landscapes, increased availability of commodity and modular polymorphic malware, as well as open-source hacking and post-exploitation tools, governments and international organizations face significant challenges in ensuring robust and effective defenses in the cyber domain. While traditional approaches of detecting and mitigating cyberattacks have been successfully applied to protect networks and maintain cyber resilience, these approaches are primarily reactive and retroactive, rather than proactive and implemented in advance of an impending cyber incident.[1] Cybersecurity representatives from governments, international organizations, and the private sector have expressed concern with this method and a desire to enrich it by designing a more forward-looking, practical approach to provide indications and warning (I&W) – or actionable intelligence and monitoring of potential threats – sufficiently in advance to enable the early detection and reaction to cyber incidents before they occur. The ability to design such an approach is hindered by the lack of a commonly accepted definition of cyber I&W, the highly classified nature of the field, and the layers of complexity introduced by constantly changing threats and networks.

In an attempt to address this problem, this research proposes a high-level yet practical strategic cyber I&W approach that governments, NATO agencies, and the private sector can apply to defend against cyber threats. The proposed approach is informed by mature I&W frameworks that the U.S. intelligence community (IC) has developed, refined, and consistently applied to monitor non-cyber threats throughout the Cold War and today. The practices of the U.S. intelligence community serve as an appropriate methodological foundation for a cyber I&W approach that can be introduced across NATO members and agencies, due to the availability of open-source literature and the broad influence of the U.S. IC in both NATO and among other Allied nations.

This article commences by first, outlining the evolution and history of I&W in the U.S.

[1] For the purposes of this article, cyber incident is defined as "actions taken through the use of computer networks that result in an actual or potentially adverse effect on an information system and/or the information residing therein." See U.S. Code of Federal Regulations, Title 48, Chapter 2, Subchapter H, Part 252, Subpart 252.2, Section 252.204-7012. Additionally, see CJCSM 6510.01b for a table of Incident Categories.

intelligence community. Second, it examines the existing definitions of cyber I&W and the divergent understanding among scholars and practitioners regarding how I&W can be applied to the cyber domain. As a third step, the research examines classic I&W frameworks for non-cyber threats and recent literature adapting I&W frameworks to cyberspace. Finally, on the basis of identified strengths in the existing approaches, the article offers a general practical approach to cyber I&W that governments, NATO agencies, and the private sector can consider adopting. To demonstrate the practical utility of our proposed approach, the research concludes by applying it to an actual cyberattack: the November 14, 2018 spear-phishing campaign by Russia's APT29 against U.S. government agencies, think tanks, and businesses.

The analysis is based on a mixed methods approach, including an examination of relevant publicly available literature such as articles, books, and reports. The literature consulted was compiled as a result of a systematic literature review of relevant databases including JSTOR, EBSCO, IEEE Xplore, and Web of Science. The research was also informed by a review of primary sources such as national cyber and military doctrines, and speeches by military and government representatives of NATO member states and NATO agencies. The arguments were further shaped and refined by a synthesis of insights gathered through correspondence and discussions with cybersecurity staff of international organizations, the U.S. government, and the private sector. This research is based on open-source literature and, due to the highly classified nature of the intelligence tradecraft, the scope, depth, and detail of the analysis and recommendations is limited. Therefore, this article should be considered as a starting point and general methodological framework of addressing the issue, accompanied by a set of recommendations, which should be adapted and refined further by agencies and decision-makers.

2. DEFINITIONS OF WARNING INTELLIGENCE

The conceptualization of indications and warning provides valuable insights into the evolution of threats and the utility of I&W approaches adopted to defend against them. The overview provided in this section describes the main elements of the I&W concept adopted and employed by the U.S. intelligence community since World War II, outlines variations in the definition of some of the key terms used in I&W frameworks in the cybersecurity community, and concludes by proposing a definition of cyber I&W.

Indications and warning is "an intelligence product upon which to base a notification of impending activities on the part of foreign powers, including hostilities, which may adversely affect military forces or security interests." (Watson, Watson and

Hopple 1990, 594; Grabo 1987, 5)[2] It includes "those intelligence activities intended to detect and report time-sensitive intelligence information on foreign developments that forewarn of hostile actions or intention against United States entities, partners, or interests." (Department of Defense 2013, p. GL-12) Warning intelligence is an analytical process that serves to assess continuously and report periodically on any developments which could indicate that a state or non-state actor is preparing an action which could threaten U.S. security interests and the interests of U.S. allies. It scrutinizes military, political or economic events, as well as other relevant and associated actions and developments or plans that could provide further insight into potential preparations for hostile acts. The analysis is an assessment of probabilities and provides a definitive (positive or negative) or a qualified (high, medium, low probability) judgement about the likelihood of the threat should it be brought to the attention of a policymaker. Warning intelligence is an art that requires understanding and continuous study of the capabilities, culture, history, and biases of potential adversaries. It applies to routine continuous monitoring and in crisis situations (Goldman 2002, iii-3).

In the context of warning intelligence, there is a fine distinction between the terms *indicator* and *indication*. An *indicator* is a theoretical or known development or an action which the adversary may undertake in preparation for a threatening act such as a deployment of forces, a military alert, a call-up of reservists, or the dispatch of a diplomatic communique. An intelligence organization anticipates an indicator's potential occurrence and adds it to a list of items to monitor, which is known as an "indicator list." Therefore, an indicator is a judgment based on collected evidence that an action of concern may happen. Information that an indicator is actually taking place constitutes an *indication*. The purpose of the *indication* is to provide insight into the adversary's potential course of action. Thus, the difference between an *indicator* and an *indication* is one between theory and practice; or expectation and an actual development (Goldman 2002, 3).

In contrast to the U.S. Department of Defense (DoD) and IC, the broader cybersecurity community has a different use of the term *indicator*. In this community, *indicators of compromise* (IOC) is used to refer to evidence indicating a breach in the security of a network (DeCianno 2014). This technical use of IOC is similar to the term *indication* described earlier. Throughout this article, we use the terms indicator and indication as they are defined in the U.S. DoD and IC. Another term, used later in this article, is *Priority Intelligence Requirements* (PIRs), which refers to an intelligence requirement to "focus information collection on the enemy or adversary and the [operational environment] to provide information required for decision making." (Joint Chiefs of Staff 2017)

[2] Indications and Warning has also been referred to as warning intelligence or indications intelligence.

Strategic warning does not have a universally accepted definition (Goldman 2002, 3). In its broad sense, *warning* is defined in the U.S. IC as “a notification of impending activities that may, or may be perceived to, adversely affect U.S. national security interests or military forces.” (JMIC 2001, 38) It is further defined as “[a] distinct communication to a decision maker about threats against U.S. and allied security, military, political, information, or economic interests. The message should be given in sufficient time to provide the decision maker opportunities to avoid or mitigate the impact of the threat.” (DIA Instruction 3000.001, 2014)

I&W has traditionally been focused on monitoring the behavior of potential adversaries on air, land, at sea, and in space. The distributed denial of service attacks on Estonia in 2007 placed cyber operations among the tools of statecraft and necessitated heightened focus on another class of monitoring targets from a relatively new environment: threats emanating in and through cyberspace. Today, warning intelligence incorporates a variety of threats and potential adversaries, both state and non-state actors that can initiate activities harmful to U.S. interests across multiple domains, including cyberspace. This wide spectrum of actors, methods and scenarios is reflected in a broader definition of threats, including any “discernible danger” that can inflict potential damage “to U.S. or allied persons, property or interests that occurs in a definable time in the future.” (DIA, Warning Fundamentals, 4)

Considering the gravity of threats to cyberspace, developing the capability to anticipate—not just react faster to—these threats would better position cyber defenders to accomplish their goals. Adapting I&W methodologies to the cyber domain would provide them with the means to do so; yet cyber I&W concepts and frameworks, as well as protocols on how to integrate these into the intelligence tradecraft, are still evolving (INSA 2018, 1; Correspondence with a cybersecurity expert, December 17, 2018). Neither NATO agencies nor the U.S. government provide publicly available comprehensive definitions of cyber I&W, perpetuating divergent understandings of cyber I&W frameworks.

Based on the literature and doctrine on I&W against non-cyber threats and interviews with cybersecurity experts, we propose the following general definition for cyber I&W frameworks and approaches:

> *An analytical process focused on collecting and analyzing information from a broad array of sources to develop indicators which can facilitate the prediction, early detection, and warning of cyber incidents relative to one’s information environment.*

When discussing the scope and purpose of I&W frameworks in the cyber domain in more detail, however, representatives from the private sector, NATO agencies, and the U.S. government define the concept differently. Some experts contend that I&W in cyberspace is primarily focused on gathering technical information on impending cyber threats, while others consider the concept to also include a survey of geopolitical developments that can influence a decision to initiate a cyber incident. Expert opinion also differs regarding the temporal parameters of the term. Some indicate cyber I&W frameworks should encompass monitoring the entire spectrum of cyberattack stages as outlined by Lockheed Martin's Kill Chain, to include detection of cyber incidents after the delivery stage.[3] Other cybersecurity experts see the utility of I&W frameworks as primarily focused on predicting incidents before they reach the delivery stage and while they are still in the reconnaissance stage, and even beforehand (INSA 2018, 3; Correspondence with cybersecurity experts and a NATO representative, December 4-21, 2018).

The U.S. Department of Defense's doctrine for cyberspace operations, DoD Joint Publication 3-12, provides useful clarity on the data-collection methods and techniques that warning intelligence applied to the cyber domain should include, and on the specific nature of cyber threats. The document stipulates that "cyberspace threat intelligence includes all-source analysis to factor in political, military, and technical warning intelligence. Adversary cyberspace actions may occur separate from, and well in advance of, related activities in the physical domains. Additionally, cyberspace threat sensors may recognize malicious activity with only a very short time available to respond. These factors make the inclusion of all-source intelligence analysis very important for effectively assessing adversaries' intentions in cyberspace." (Department of Defense 2018, IV-7) Yet, JP 3-12 and other U.S. doctrinal documents have not yet provided clear definitions and guidelines about how warning methodologies for cyber threats should be developed and how they should be incorporated in existing warning frameworks. Furthermore, existing U.S. documents fail to provide guidance for acceptable courses of action or responses given impending cyber threats.

3. CLASSIC I&W FRAMEWORKS

There are several well-known and widely-used I&W frameworks that the U.S. IC has been using to monitor and detect potentially threatening adversary behavior. Two such classic frameworks, summarized in this section, are the Lockwood Analytical Method for Prediction (LAMP) and the DoD's *Defense Warning Network Handbook* (Lockwood 2002, Joint Chiefs of Staff). These approaches can serve as the foundation in formulating a cyber I&W framework.

[3] The seven-step Lockheed Martin Kill Chain is a well-known framework for mapping the stages of cyber incidents in support of intelligence-driven defense. For more information, see Muckin and Fitch, 2019.

I&W entails a probabilistic analysis, in which an analyst attempts to provide an assessment which is as realistic and objective as possible, given data and time constraints. A knowledge of history, doctrine, and precedent is critical in this process (Goldman 2002, 13). Specifically, when compiling indicator lists, analysts draw primarily from three sources: logic or longtime historical precedent, lessons learned from the behavior of threat actors during a recent war or crisis, and specific knowledge of the military doctrine or practices of the threat actors (Goldman 2002, 26).[4] One of the seminal warning intelligence analysts, Cynthia Grabo, argued that a robust warning methodology should incorporate both military and political indicators, prioritize indicators, and examine a variety of data sources in context (Grabo 1987).

The LAMP is one such framework that applies structure to the warning intelligence problem (Lockwood 2002). It assumes that the future is a spectrum of changing relative probabilities and aims to determine the relative probability of alternative futures. It consists of the following 12 steps:

1. Define the intelligence question under consideration with sufficient specificity and narrowness of enquiry
2. Specify the actors involved in the problem
3. Study each actor's intentions and perceptions of the problem
4. Specify all possible courses of action for each actor
5. Determine the major scenarios
6. Calculate the total number of alternate futures
7. Perform a pairwise comparison of all alternate futures within each scenario to establish their relative probabilities
8. Rank the alternate futures for each scenario from highest relative probability to lowest relative probability
9. For each alternative future, analyze the scenario in terms of its consequences for the intelligence question
10. Determine focal events that must happen to realize each future
11. Develop indicators for each focal event
12. State the potential of a given alternate future to transpose into another alternate future (Lockwood 2002, 2010; Singh 2013).

LAMP provides significant leeway for defining the number of major scenarios and the breadth of problems with which one is concerned. Although it does not define the exact form of comparison (e.g., Delphi method, survey, Bayesian inference) to use when developing the relative rankings of alternative future scenarios, the framework clearly relies on the talents of the individuals engaged in the process and therefore could result in different outcomes. That said, it is amenable to evaluation and adjustment

[4] In this context, logic is tied to an actor's historical pattern of behavior - rather than based on an actor-agnostic theory, such as rational choice theory.

over time as events do (or do not) come to pass, providing a means to "grade" the probabilities.

The DoD's *Defense Warning Network Handbook* provides a similar set of steps as LAMP, but without the assignment of probabilities:

1) Identify anomalies/imagine alternatives
2) Produce scenarios
3) Identify conditions, drivers, and indicators
4) Determine warning threshold
5) Explore opportunities to influence or mitigate the threat
6) Communicate warning (Joint Chiefs of Staff).

As with LAMP, the DoD approach depends on the talents and experience of those engaged in the process. The U.S. IC has changed its intelligence approach over time, including having dedicated offices and analysts focused on warning, relative to other periods when warning was one of several duties assigned to analytic offices (Gentry and Gordon 2018). The two general warning frameworks provided here share a common approach that relies on speculating on potential futures which would be of concern, and crafting indicators which would provide early pointers towards that future coming to pass. These approaches rely on others within the military, intelligence, and defense communities to take action based on these warnings. Both frameworks offer a systematic way to monitor and detect threats and contain valuable components that can inform a cyber I&W framework; but are not sufficiently detailed to provide practical guidance for practitioners.

4. I&W FRAMEWORKS FOR CYBER THREATS

Experts have conducted promising initial research into adapting classic I&W frameworks or key components of the intelligence I&W cycle to the cyber domain. It is worth reviewing some of this research to demonstrate its applicability and build upon its strengths.

General I&W frameworks vary from cyber-specific frameworks in several areas, including in terms of the target of the analysis (i.e. physical/conventional/kinetic threats vs. cyber threats), but the classic frameworks can be adapted to address cyber threats. Another consideration is the partial divergence in analytical approaches. Specifically, classic intelligence analysis is primarily backward-looking and forensically focused, while cyber I&W framework may incorporate predictive analytical techniques that add a forward-focused analytical component. Nevertheless, the classic frameworks can inform the design of a robust I&W methodology for cyber incidents, while analytical processes, data-collection techniques, and methodologies can also be transferable across the two frameworks.

One such approach is a twelve-step adaptation of Lockwood's LAMP method by Robinson et al (2012):

1. Problem identification: determine the issue
2. Identify potential actors
3. Actor courses of action: viability and probability (include the Kill Chain here)
4. Determine scenario enablement
5. Manifested scenario focal events
6. Create focal event indicators: an adversary prepares for hostilities
7. Collect and monitor through indicators: assess emerging trends
8. Discern the probable scenario that is trending
9. Readjust for new manifestations of the scenario
10. Deception in indicators
11. Mental model avoidance: is it expectation or actuality, theory or current developments?
12. Strategic options analyzed against viable scenarios (Robinson, Astrich and Swanson, 2012).

More recently, the Intelligence and National Security Alliance (INSA) published a working group report that proposes a high-level conceptual framework against cyber threats, consisting of the following seven steps:[5]

1. Identify & prioritize assets – identify which data, devices, personnel, and facilities are most critical to the organization
2. Refine the threat – identify which top 10 or 15 cyber threats may inflict the most damage to the assets listed in step 1
3. Assess threat courses of action – design adversaries' Course of Action (COA) based on scenarios; can use the Lockheed Martin's Kill Chain or MITRE's ATT&CK methodology
4. Break down scenarios into IOCs
5. Plan and exercise countermeasures
6. Align to the intelligence cycle
7. Execute proactive countermeasures (INSA 2018, 12-7).

The valuable contribution from the INSA approach is to combine the outward focus of warning frameworks (i.e., what scenarios we are concerned about) with an inward

[5] INSA is a U.S.-based nonprofit organization founded in 1979 that provides a platform for the development and promotion of public-private solutions to national security challenges. For more information, see https://www.insaonline.org/about/.

focus on what those scenarios would impact. It begins by identifying and prioritizing the assets which an organization should seek to protect, and proceeds by understanding various threat actors' courses of action.[6]

5. COMPARISON OF EXISTING I&W FRAMEWORKS

Each of the four frameworks discussed provides insights into developing indications and warning for cyber threats. The two traditional intelligence processes, developed by the Defense Intelligence Agency and Lockwood, are general approaches which should be applied and made more specific to the cyber domain, but do provide a structured and logical approach. Robinson has attempted to do that with Lockwood's approach; while the INSA paper provides a different view on applying traditional intelligence community approaches. Below, we have mapped these four frameworks against general categories of analysis and action to highlight where they overlap and combine their elements into a synthesized approach.

Although not high-level cyber I&W frameworks, there are two other important approaches used to understand how malicious actors plan and conduct cyberattacks: Lockheed Martin's Cyber Kill Chain and MITRE's Adversary Tactics, Techniques, and Common Knowledge (ATT&CK). Both approaches start with the premise that understanding the steps a malicious cyber actor must accomplish to plan and execute an operation can help a cyber defender understand what activity to look for and the defensive measures to implement. The Kill Chain consists of seven steps: reconnaissance, weaponization, delivery, exploitation, installation, command and control, and action on objectives (Lockheed Martin 2015).

The ATT&CK framework was developed to provide a common taxonomy for mapping real-world observed behavior and techniques. It maps a technique to a stage of an operation and provides insight into what that technique is supposed to accomplish. Cyber Red Teams can use the framework to develop playbooks based on real-world experience, as well as show what techniques or exploits are most commonly used by Advanced Persistent Threats (APTs).[7] The framework, similar to the Cyber Kill Chain but with additional depth, maps techniques to the stages of an intrusion. In

[6] MITRE has developed a method for identifying critical cyber assets called Crown Jewel Analysis. Similar to mission assurance analysis, it starts with identification of critical missions and the assets those missions rely upon. See the MITRE Corporation. For more on this approach, see https://www.mitre.org/publications/systems-engineering-guide/enterprise-engineering/systems-engineering-for-mission-assurance/crown-jewels-analysis.

[7] For a general overview of the origins and use of the ATT&CK framework, see Strom, 2018. Playbook is a term used to describe a specific sequential collection of ATT&CK framework-mapped post-exploitation techniques employed by an adversary as they move through the Kill Chain phases of Installation, Command & Control and Actions on Objectives, under which MITRE's ATT&CK framework's 11 tactics logically fall. Each playbook is essentially a post-exploitation threat model, understanding that an adversary may use the same playbook for each operation or change technique combinations over time.

the case of ATT&CK, it has eleven stages tied to the desired objective for the stage: initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, exfiltration, and command and control.[8] More recently, MITRE has been developing a PRE-ATT&CK model to try to identify the stages of cyberattack planning prior to access to a network.

There are several insights to draw from this mapping. The frameworks vary in terms of the level of specificity they provide for a given step. The Lockwood approach, for example, provides several steps for developing scenarios, but Robinson's adaptation of Lockwood captures these in fewer steps. The DIA framework focuses ultimately on communicating warning (which we have placed in a general category of "acting" on indicators). While being less specific on recommending steps for generating scenarios, the DoD framework emphasizes the policy relevance of an I&W approach, while Lockwood does not address either tracking or acting. In comparison to the others, Robinson's framework is more focused on tracking, adjusting and acting on the indicators.

All frameworks contain valuable elements for a cyber I&W framework, but no one approach appears to incorporate the classic lessons of effective threat intelligence which Grabo, among others, advocated: such as conducting both technical and strategic assessment of threat actors and their environment, as well as clearly emphasizing the need to produce actionable information useful for policymakers. Therefore, the frameworks can be consolidated to inform the design of a cyber I&W approach that comprehensively addresses these issues and can be applied to the structure of an organization to inform decision-making.

6. RAND'S PRACTICAL APPROACH FOR CYBER I&W

RAND proposes the following approach for cyber I&W, which offers a practical, hands-on workflow for cyber defenders; synthesizes and adds onto many of the components of the other I&W frameworks; and would typically belong in an organization's Cyber Threat Intelligence (CTI) program. The steps of RAND's approach all take place in the first phase of cyber incident response: preparation (Kral 2011).[9] The approach explicitly accounts for both technical assessments (e.g., what are the most commonly used playbooks of APT actors that are likely to target a network?) and contextual, geopolitical assessments (e.g., what military, political, economic or social developments influence a decision to initiate an incident?) to understand the broader operating environment. Adding a focus on the strategic environment moves beyond the technical aspects of cybersecurity to attempt to understand the external factors that

[8] The full framework can be found at https://attack.mitre.org.

[9] The phases of incident response are as follows: preparation, identification, containment, eradication, recovery and lessons learned. Only the first phase, preparation, aligns with the predictive and anticipatory nature of I&W. See Kral 2011.

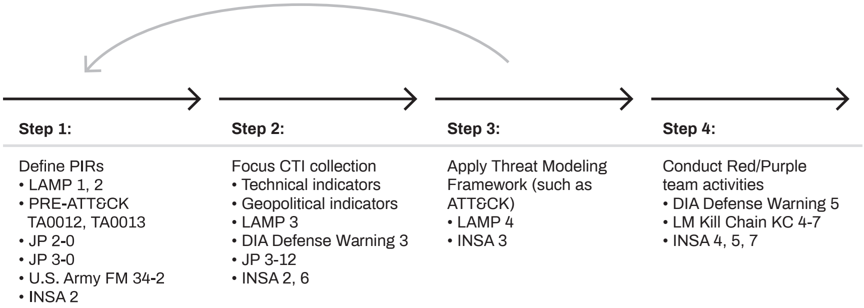
TABLE 1: A COMPARISON OF CLASSIC AND CYBER INDICATION AND WARNING FRAMEWORKS

General Actions	Classic I&W Frameworks		Cyber I&W Frameworks	
	Defense Warning Handbook	LAMP / Lockwood	Robinson (adopted from LAMP)	INSA
Framing Questions		1.Define the intelligence question under consideration with sufficient specificity and narrowness of enquiry	1. Problem identification: determine the issue	
Identify Threats		2. Specify the actors involved in the problem 3. Study each actor's intentions and perceptions of the problem 4. Specify all possible courses of action for each actor	2. Identify potential actors 3. Actor courses of action: viability and probability (include the Kill Chain here)	2. Develop a refined understanding of the most likely threats
Identify Assets to Defend				1. Identify and prioritize assets to be protected
Develop Scenarios	1) Identify anomalies/imagine alternatives 2) Produce scenarios	5. Determine the major scenarios 6. Calculate the total number of alternate futures 7. Perform a pairwise comparison of all alternate futures within each scenario to establish their relative probabilities 8. Rank the alternate futures for each scenario from highest relative probability to lowest relative probability 9. For each alternate future, analyze the scenario in terms of its consequences for the intelligence question	4. Determine scenario enablement 5. Manifested scenario focal events	3. Using structured analytic techniques, forecast likely attack scenarios
Develop Indicators	3) Identify conditions, drivers and indicators 4) Determine warning threshold	10. Determine focal events that must happen to realize each future 11. Develop indicators for each focal event 12. State the potential of a given alternate future to transpose into another alternate future	6. Create focal event indicators: an adversary prepares for hostilities	4. Decompose scenarios into indicators of likely adversary actions
Track Indicators	5) Explore opportunities to influence or mitigate the threat		7. Collect and monitor through indicators: assess emerging trends 8. Discern the probable scenario that is trending	6. Collect intelligence on indicators and adversary plans and intentions
Act on Indicators	6) Communicate warning		9. Readjust for new manifestations of the scenario 10. Deception in indicators 11. Mental model avoidance: is it expectation or actuality; theory or current development? 12. Strategic options analyzed against viable scenarios	5. Plan and exercise countermeasures to likely adversary actions 7. Execute proactive measures to counter anticipated attack vectors

indicate intent and timing behind adversary cyber activity. As such, corollary focused CTI collection combined with a strategic all-source approach may help answer "Why?" and "When?" questions, to give defenders further indications and warning as to the probability of a cyber incident.

RAND's Practical Approach for Cyber I&W begins with suggesting the use of Priority Intelligence Requirements (PIRs). This leads to an iterative loop with CTI collection, then to employment of systematically-constructed playbooks of adversarial techniques and behavior, by leveraging a threat modeling framework such as MITRE's ATT&CK. Finally, Red/Purple Team activities emulate relevant threats, check for visibility gaps, and allow mitigations to be designed.[10] This approach should be accessible and usable to cyber defense teams at all levels of capability maturity. Figure 1 shows our approach, followed by a high-level overview of each of the steps. Depending on an organization's resources and capabilities, much more depth can exist within each step as an organization's resources and capabilities allow.

FIGURE 1. RAND'S PRACTICAL APPROACH FOR CYBER I&W

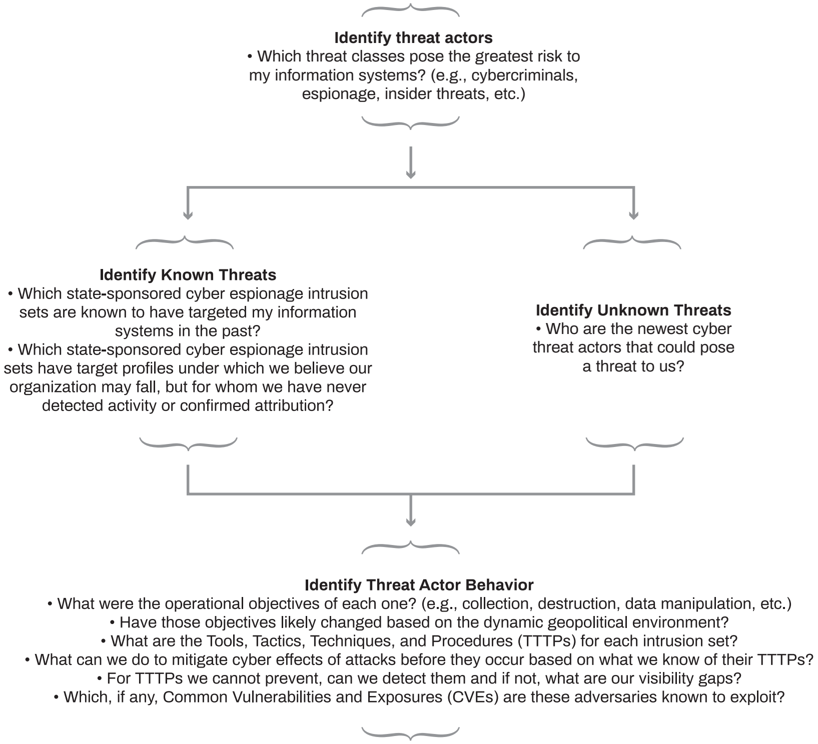


Step 1: Define PIRs

Anticipation of threats can be facilitated by a simplified approach of developing some basic PIRs from a cyber defense perspective. PIRs consist of a concise set of questions devised, prioritized, regularly updated, and continuously answered to better understand one's adversaries by allowing the defenders to focus their CTI collection. Examples of PIRs developed to facilitate discovery of I&W for cyber incidents within a cyber defense operation's CTI Program are shown in Figure 2.

[10] A Purple Team (red + blue) is a modification of a traditional Red Team, where the offensive cyber operations (Red Team operations) are conducted side by side with or by cyber defense analysts (Blue Team operations) against one's own network. This can have numerous benefits. Purple Teams work well in many organizations but not all; some still benefit from the hard separation, in which case an organization may choose to substitute our usage of Purple Teams with the traditional Red Team approach.

FIGURE 2: EXAMPLE PIRS TO FACILITATE DISCOVERY OF CYBER I&W



Identify threat actors

- Which threat classes pose the greatest risk to my information systems? (e.g., cybercriminals, espionage, insider threats, etc.)

Identify Known Threats

- Which state-sponsored cyber espionage intrusion sets are known to have targeted my information systems in the past?
- Which state-sponsored cyber espionage intrusion sets have target profiles under which we believe our organization may fall, but for whom we have never detected activity or confirmed attribution?

Identify Unknown Threats

- Who are the newest cyber threat actors that could pose a threat to us?

Identify Threat Actor Behavior

- What were the operational objectives of each one? (e.g., collection, destruction, data manipulation, etc.)
- Have those objectives likely changed based on the dynamic geopolitical environment?
- What are the Tools, Tactics, Techniques, and Procedures (TTTPs) for each intrusion set?
- What can we do to mitigate cyber effects of attacks before they occur based on what we know of their TTTPs?
- For TTTPs we cannot prevent, can we detect them and if not, what are our visibility gaps?
- Which, if any, Common Vulnerabilities and Exposures (CVEs) are these adversaries known to exploit?

Relative to other frameworks reviewed in this article, PIRs relate closely to LAMP steps 1 and 2, as well as the first two tactics described by MITRE's PRE-ATT&CK framework: priority definition planning (TA0012) and priority definition direction (TA0013). It also maps to what INSA's Framework for Cyber I&W lists as step 2.

Step 2: Use the Derived PIRs to Focus CTI Collection

CTI can often answer "Who?", "What?", "Where?" and "How?" questions, helping to understand adversaries' behaviors and tools, tactics, techniques, and procedures (TTTPs), thus strengthening I&W and cyber incident preparation or prevention. The findings from RAND's step 1 help drive CTI collection requirements and filter the mountain of tactical-level IOCs (e.g., malicious IP addresses, domains or hashes) that correspond to intrusion sets other than those targeting the organization, and strategic-level geopolitical developments (e.g., incoming national elections or recalling reservists) that could be indicative of probable adversary action to help scope and focus collection to what matters most to the organization, given the reality of limited resources.

Harnessing CTI in this way closely relates to LAMP step 3, DIA's Defense Warning Handbook step 3, and INSA's steps 2 and 6, and can also incorporate all-source intelligence collection for additional strategic context, to better help answer "When?" or "Why?" questions that may be defined in PIRs (see JP 3-12). Relevant CTI uncovered in response to PIRs such as, "Which, if any, CVEs are these adversaries known to exploit?" can also serve as vulnerability exploitation intelligence, informing enterprise patching prioritization efforts. Automated operationalization of IOCs from CTI is recommended - but describing this process is beyond the scope of this article.

Step 3: Apply Adversary Threat Modeling Framework: MITRE's ATT&CK

Step 3 of the approach is analogous to a narrowed LAMP step 4 and INSA's step 3. It takes the findings of which intrusion sets are targeting one's organization and enters them into an adversary threat modeling framework, such as MITRE's Enterprise ATT&CK Navigator (an interactive JavaScript-based version of the framework).[11] This helps prioritize an organization's focus on pre-mitigating probable attacks by being able to prevent or detect the specific techniques employed by one's adversaries. Once the most relevant TTTPs are identified, cyber defenders can use the information as inputs to a Red Team statement of work or Purple Team task list. MITRE's PRE-ATT&CK and ATT&CK framework has expanded upon Lockheed Martin's cyber intrusion Kill Chain, to originally include treating Kill Chain steps as akin to overarching tactics (represented as column heads) under which many techniques fall.

Step 4: Conduct Continuous Red / Purple Team Ops

The final step in RAND's Practical Approach for Cyber I&W is the culmination of all previous steps: it tests relevant adversary TTTPs and playbooks against the organization's environment. By this stage, the defenders know who their threats are, how they behave, the details of their tools (capability/how), when (opportunity), and why (intent) they might attack. In this step, if using the Purple Team concept, the defender emulates adversary behavior and current playbooks as closely as possible while tuning defenses to prepare for a potential similar incident. Performing these activities is akin to step 5 of the DIA Warning framework, and incorporates steps 4, 5, and 7 of the INSA framework. Another advantage this step has is that it allows cyber defenders to continuously discover, understand and test for detection visibility gaps, continuously improve their Security Information and Event Management (SIEM) and other detection content, and improve the security settings or architectural design details of an organization's network ahead of time. It also allows an organization to define and refine Courses of Action (COAs) to take during the containment phase of an attack, each of which can map to different phases of the Lockheed Martin cyber intrusion Kill Chain.

[11] https://mitre-attack.github.io/attack-navigator/enterprise/

7. CASE STUDY – NOVEMBER 14, 2018 APT29 SPEAR-PHISHING CAMPAIGN

Finally, we share a real-world example of an organization applying the RAND practical approach to the cyber I&W set forth in this article: integrated into normal cyber defense operations against the backdrop of strategic geopolitics, corresponding cyber espionage activity, and "friendly" government agencies conducting their own cyber I&W and counter-threat operations. The example involves the widespread November 14, 2018 post-midterm U.S. election phishing campaign, widely believed to have been perpetrated by the Russian-nexus intrusion set publicly known as APT29 (attributed to the Russian Foreign Intelligence Service (SVR), see Modderkolk 2018). We use "Organization Z" to denote one of the targets of the November attack, and describe examples of their cyber I&W actions to prepare for a probable attack.

Application of the cyber I&W process in this case resulted in Organization Z predicting and assessing with moderate confidence that APT29 would attempt a cyber intrusion against it, corresponding to the U.S. midterm elections, based on past adversary patterns. As some APT intrusion sets have shifted to or experimented with more generic or commodity malware or tools in an attempt to further obscure their origins for the purpose of making attribution more difficult, Organization Z had applied all steps of this approach not only to APT29's TTTPs, but also to tools more commonly used not just by legitimate Red Team operators, but some APT groups too. Organization Z's widening of scope for what tool to test for a Purple Team task is an example of efficiency when selecting a tool or technique from the ATT&CK framework in RAND's step 3 to test in RAND's step 4.

The tool selected in step 3 was based on answering step 1 PIRs: "Which state-sponsored cyber espionage intrusion sets are known to have targeted my information systems in the past?"; and "What are the TTTPs for each intrusion set?" The answers to these two PIRs resulted in the decision to focus Organization Z's specific CTI collection requirements in step 2. Multiple APT groups as well as Red Team operators use commodity tools. This is illustrative of an advantage that can be taken back by defenders in an analog of attacker/defender co-evolutionary adaptation, giving rise to increased cyber resiliency despite changing adversary tools and predictability.

This preparation resulted in Organization Z using threat emulation software, Cobalt Strike, on its network during internal Purple Team activities in preparation for a variety of threats.[12] This led to improved SIEM content, verification of detection and

[12] Cobalt Strike is a commercial, full-featured, penetration testing tool which bills itself as "adversary simulation software designed to execute targeted attacks and emulate the post-exploitation actions of advanced threat actors." Cobalt Strike's interactive post-exploit capabilities cover the full range of ATT&CK tactics, all executed within a single, integrated system. (https://cobaltstrike.com/downloads/csmanual38.pdf)

prevention capabilities, tool integration and automation.[13] During this test, detection of the type of beacon was confirmed, integration of security platforms demonstrated value, and SIEM content was created to notify Organization Z's cyber defense team via email if the selected events deemed critical occurred.

Just weeks after conclusion of the testing, and the day before election day and the expected intrusion attempt, on November 5, 2018, USCYBERCOM announced that "the Cyber National Mission Force, a unit subordinate to U.S. Cyber Command, posted its first malware sample to the website VirusTotal…"[14] The initial focus of uploads was unclassified malware samples attributed to Russia. The timing of this did not seem coincidental and appeared suggestive of a larger plan aimed to disrupt any potential Russian interference in the midterm elections, which was suspected based on Russia's interference in the 2016 U.S. presidential elections. November 6, 2018 (election day) passed without incident. Did USCYBERCOM, with all their resources, have their own cyber I&W that APT29 was going to perpetrate a large attack? Was the November 5, 2018 change in policy - uploading voluminous malware samples associated with Russian espionage - part of an attempt to disrupt the attack?

Eight days after the election, on November 14, the APT29's offensive cyber operation was finally conducted, but the initial tool used during the exploitation phase of the Kill Chain was a Cobalt Strike Beacon payload, with a modified Pandora malleable Command and Control (C2) (Dunwoody et al., 2018). It was previously unseen as a tool used by this intrusion set and is a widely available commodity tool, with the Pandora malleable C2 available as open-source code on GitHub. There are unanswered "Why?" questions in this case, but ultimately the intrusion attempt against Organization Z was unsuccessful and quickly contained.

USCYBERCOM was unable to stop this attack from happening entirely, but one of Organization Z's hypotheses as to why the attack was delayed by eight days was that USCYBERCOM disrupted the attack initially on or before November 6. It is possible that the uploaded malware samples or something else resulted in a change in tools by the adversary. The C2 domain was registered on October 15, 2018, yet it could have initially been intended for communication with another tool, beacon or malware specimen.[15]

[13] These details illustrate the tip of the iceberg on how an organization can go as deep as they have the resources for - chiefly based on their time and personnel availability - but additional expansion was beyond the scope of this article.

[14] "…Recognizing the value of collaboration with the public sector, the Cyber National Mission Force (CNMF) has initiated an effort to share unclassified malware samples it has discovered that it believes will have the greatest impact on improving global cybersecurity. For members of the security community, CNMF-discovered malware samples will be logged at this website: https://www.virustotal.com/en/user/CYBERCOM_Malware_Alert/".

[15] One can check domain registration dates and history by querying domain registration or passive DNS records.

As was revealed in late February 2019, there was a larger plan by USCYBERCOM, approved by the President and Congress and coordinated among numerous government agencies, to protect against election interference with an offensive cyber campaign. The authority was afforded by National Security Presidential Memorandum 13 (Nakashima 2018). The malware which USCYBERCOM uploaded to VirusTotal and the announcement about it seem to have formed only one small piece of a larger strategy that the public was able to glimpse at the time; and even though the specific malware uploaded was not likely to have involved new adversary tools, perhaps it had a psychological effect that affected adversary behavior and planning.

From an I&W perspective, this particular case study underscores the challenge of predictive analysis when many variables are at play, and also illustrates the interconnected and dynamic reality of the operating environment when other friendly agencies take calculated actions that possibly affect adversary behavior and disrupt some basic predictability which another organization may have established. It also highlights the potential increased resiliency that RAND's proposed practical cyber I&W approach can bring about. Perhaps resilience is more important than knowing precisely when and how an attack will occur, though a combination of the two would constitute the best case scenario from a defender's perspective.

8. CONCLUSION

Much can be learned from an examination of traditional strategic I&W intelligence frameworks, as well as the main methodological and analytical challenges that the I&W field has faced and already addressed, though significant differences in the cyber domain do exist when it comes to applying a practical workflow to operationalize collected intelligence. Despite these differences, however, both the cyber domain and traditional strategic I&W frameworks applied to the four other domains use overlapping methods and techniques for threat modeling and intelligence collection and exploitation, which can serve as a methodologically sound foundation for steps constituting a newly codified approach of addressing anticipated cyber threats.

RAND's proposed Practical Approach for Cyber I&W consists of four steps; each corresponds to and draws upon previously reviewed I&W frameworks. This overall approach accounts for collection, processing, and operationalization of filtered tactical, operational and strategic CTI, to determine and understand relevant adversaries within the context of the broader geopolitical environment as it relates to the network being defended. It also leverages MITRE's ATT&CK as an example of applying a threat modeling framework and to some extent, the PRE-ATT&CK extension.

An organization taking this approach to cyber I&W, integrating it into their cyber defense operations, and adding their own creativity and toolsets to expand, refine, and tailor the processes within each step can continuously improve readiness, prioritize limited resources, and enhance overall resilience to cyber incidents. This approach is also intended to be accessible by any cyber defense team at any capability maturity level; and as an organization's capabilities increase, they can iterate, automate, and expand processes in each step as they wish. For example, incorporating even more ideas from traditional I&W frameworks to develop new PIRs or improve COAs is easy to add to steps 1 or 4 respectively.

The November 14, 2018 spear-phishing campaign by Russia's APT29 against U.S. government agencies, think tanks, and businesses demonstrates how the proposed cyber I&W approach can be integrated into cyber defense operations and applied to achieve resiliency against cyber adversaries, despite inevitable unpredictability.

REFERENCES

Blackshaw, Amy. 2016. "Behavior Analytics: The Key to Rapid Detection and Response?" *RSA*. https://www.rsa.com/en-us/blog/2016-01/behavior-analytics-the-key-to-rapid-detection-response.

DeCianno, Jessica. 2014. "Indicators of Attack vs. Indicators of Compromise," *CrowdStrike*. December 9. https://www.crowdstrike.com/blog/indicators-attack-vs-indicators-compromise/.

Defense Intelligence Agency. 2014. "Instruction 3000.001, Enclosure 1, 27 May". In Defense Intelligence Agency, "Warning Fundamentals", Unclassified briefing, 3.

Defense Intelligence Agency. "Warning Fundamentals." Unclassified briefing.

Department of Defense. 2013. "Joint Intelligence, Joint Publication 2-0." October 22. http://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf.

Department of Defense. 2018. "Cyberspace Operations. Joint Publication 3-12." June 8. http://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_12.pdf?ver=2018-07-16-134954-150.

Dunwoody, Matthew, Andrew Thompson, Ben Withnell, Jonathan Leathery, Michael Matonis, and Nick Carr. "Not So Cozy: An Uncomfortable Examination of a Suspected APT29 Phishing Campaign." *FireEye*. November 19, 2018. https://www.fireeye.com/blog/threat-research/2018/11/not-so-cozy-an-uncomfortable-examination-of-a-suspected-apt29-phishing-campaign.html.

Gentry, John and Joseph Gordon. 2018. "US Strategic Warning Intelligence: Situation and prospects." *International Journal of Intelligence and Counterintelligence* 31(1): 19-53.

Goldman, Jan (ed.). 2002. *Anticipating Surprise: Analysis for Strategic Warning*. Center for Strategic Intelligence Research: Joint Military Intelligence College.

Grabo, Cynthia. 1987. *Warning Intelligence*. The Intelligence Profession Series: Association of Former Intelligence Officers.

Hernandez-Suarez, Aldo, et al. 2018. "Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using ℓ1 Regularization." *Sensors* 18(5): 1380.

Husák, M., et al. 2018. "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security." *IEEE Communications Surveys & Tutorials*: 1-21. https://ieeexplore.ieee.org/document/8470942/.

INSA. 2018. "A Framework for Cyber Indications and Warning". *Intelligence and National Security Alliance*. October. https://www.insaonline.org/wp-content/uploads/2018/10/INSA-Framework-For-Cyber-Indications-and-Warning.pdf.

JMIC. 2001. "Warning Glossary." This is not a publicly released document.

Joint Chiefs of Staff. 2017. "Joint Publication 3-0. Joint Operations. January 17, Incorporating Change 1, October 22, 2018." http://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_0ch1.pdf?ver=2018-11-27-160457-910.

Joint Chiefs of Staff. "Joint Staff J2. Defense Warning Staff. J2 Warning." *Defense Warning Network Handbook*. 4th ed. This is not a publicly released document.

Kral, Patrick. "The Incident Handler's Handbook." *SANS Institute InfoSec Reading Room*, February 21, 2012. https://www.sans.org/reading-room/whitepapers/incident/incident-handlers-handbook-33901.

Lockheed Martin. 2015. "Gaining the Advantage: Applying Cyber Kill Chain ® Methodology to Network Defense 2015". https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf.

Lockwood, Jonathan. 2002. "The Lockwood Analytic Method for Prediction (LAMP). An Innovative Methodological Approach to the Problem of Predictive Analysis." *LAMP-Method*. January. http://lamp-method.org/lampppt.ppt.

Lockwood, Jonathan. 2010. "The Application of LAMP." http://lamp-method.org/2.html.

Modderkolk, Huib. 2018."Dutch agencies provide crucial intel about Russia's interference in US-elections." *De Volkskrant*, January 25. https://www.volkskrant.nl/media/dutch-agencies-provide-crucial-intel-about-russia-s-interference-in-us-elections~a4561913/.

Muckin, Michael and Scott C. Fitch. 2019. "A Threat-Driven Approach to Cyber Security Methodologies, Practices and Tools to Enable a Functionally Integrated Cyber Security Organization," *Lockheed Martin Corporation*. https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Threat-Driven-Approach.pdf.

Nakashima, Ellen. "White House Authorizes 'Offensive Cyber Operations' to Deter Foreign Adversaries." *The Washington Post*. September 20, 2018. Accessed March 09, 2019. https://www.washingtonpost.com/world/national-security/trump-authorizes-offensive-cyber-operations-to-deter-foreign-adversaries-bolton-says/2018/09/20/b5880578-bd0b-11e8-b7d2-0773aa1e33da_story.html?utm_term=.510d18f07e45.

Robb, Drew. 2017. "Eight Top Threat Intelligence Platforms." *eSecurity Planet*. July 18. https://www.esecurityplanet.com/products/top-threat-intelligence-companies.html.

Robinson, Michael, Craig Astrich and Scott Swanson. 2012. "Cyber Threat Indications & Warning: Predict, identify and counter." *Small Wars Journal*. July 26.

Singh, Jai. 2013. "The Lockwood Analytical Method for Prediction within a Probabilistic Framework." *Journal of Strategic Security* 6(3): 83-99.

Strom, Blake. 2018. "ATT&CK 101." May 3. *The MITRE Corporation*. https://medium.com/mitre-attack/att-ck-101-17074d3bc62.

The MITRE Corporation, "Crown Jewels Analysis," *MITRE Systems Engineering Guide*. https://www.mitre.org/publications/systems-engineering-guide/enterprise-engineering/systems-engineering-for-mission-assurance/crown-jewels-analysis.

U.S. Code of Federal Regulations. "Title 48. Chapter 2. Subchapter H. Part 252. Subpart 252.2. Section 252.204-7012."

U.S. Cyber Command. 2018. "New CNMF initiative shares malware samples with cybersecurity industry." November 5. https://www.cybercom.mil/Media/News/News-Display/Article/1681533/new-cnmf-initiative-shares-malware-samples-with-cybersecurity-industry/

Watson, Bruce, Susan Watson and Gerald Hopple. 1990. "United States Intelligence: An Encyclopedia". New York: Garland Publishing, Inc., 594. In *JMIC Warning Glossary 2001*.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Recommendations for Enhancing the Results of Cyber Effects

Erwin Orye
NATO Cooperative Cyber Defence
Centre of Excellence
Tallinn, Estonia
erwin.orye@ccdcoe.org

Olaf M. Maennel
Department of Computer Science
Tallinn University of Technology
Tallinn, Estonia
olaf.maennel@ttu.ee

Abstract: Cyber effects[1] should be considered an important tool in the toolbox of warfare for the commander of a military operation. This paper discusses the key elements required to enhance decision-making in cyber operations. Many different parameters influence the outcome, but only some of them are internally controllable. This paper outlines how to predict the outcome of cyber effects and how to measure that outcome. It gives advice on developing cyber effect capabilities and reflects on how to integrate cyber effects in a mission as lateral support. The authors recommend a set of best practices for enhancing cyber effects in modern warfare.

Keywords: *cyber, effect, prediction, measuring, achieving*

1. INTRODUCTION

Although the defender and the attacker each control only a very small part of the cyberspace they use, whoever can influence the portion of cyberspace used by the adversary has the potential to control the adversary [1].

Better estimation of the effects achieved by cyber operations will allow for an enhanced decision-making process and ultimately, increased control over the adversary [2]. This higher-quality estimation will also improve the ability to predict side-effects,

[1] An "effect" is a direct or indirect objective (intended) outcome of an action. In warfare, the actions are intended to create effects, typically against the functional capabilities of a material target or to the behaviour of individuals.

both those that might be useful and those that are unwanted and could cause collateral damage. In this paper we discuss the strategic aspects to be taken into account in order to develop cyber effect capabilities and discuss the importance of predicting and measuring the outcomes of cyber effects.

How to integrate cyber effects in a mission is not yet well defined. In traditional warfare domains, such as land, sea and air, there are well-defined procedures and streamlined information-sharing mechanisms for lateral support from one nation to another.

First, we provide an overview of how cyber effects can be measured and predicted. Next, we discuss how cyber effects can be achieved and enhanced. Finally, the authors provide a series of recommendations stressing the important role of collaboration in enhancing cyber effects in modern warfare.

2. STEP 1: MEASURING AND PREDICTING CYBER EFFECTS

To enhance the effectiveness of cyber operations, a continuous evaluation of the impact is needed to recommend changes to tactics, strategies, objectives, and guidance. The end state of a campaign is an original estimate that will be constantly modified during an operation. Cyber effects should be estimated in order to identify and quantify the impact of cyber operations in warfare, which is essential to predict an end state. In order to make adjustments during the cyber operation, the outcomes, also referred to as battlefield damage estimation in kinetic[2] warfare, need to be measured.

A. Scope

There is no commonly accepted definition of cyber war and cyber warfare, which indicates the difficulty of reaching such a definition. The terms computer information warfare (IW), (offensive) information operation (IO), and network attack (CNA) are frequently used interchangeably [3].

Cyber war basically refers to a sustained computer-based cyberattack by a state, state-owned organisation (e.g. NSA[3] or national CERT[4]) or state-sponsored organisation against the IT infrastructure of a target state.

[2] "Kinetic warfare" is used in this paper as an umbrella term to cover warfare that uses weapons that have mechanic, kinetic, thermal, radiological, biological, or chemical effects.

[3] National Security Agency

[4] Computer Emergency Response Team

Cyber warfare could be defined in different ways:

- As defending and attacking information and computer networks, as well as denying an adversary's ability to do the same, or even dominating the information environment on the battlefield. It can include computer or network penetration, denial-of-service attacks on computers and networks, equipment sabotage through cyberspace, sensor jamming, and even manipulating trusted information sources to condition or control an adversary's thinking [4].
- As the use of computers or network-based capabilities by a state, or a group or person whose actions can be attributed to a state, in order to launch an attack on another state [3].
- By means of essential characteristics it has to fulfil: "a cyber attack reaching the level of an armed attack or cyber activity occurring in the context of armed conflict." The essential characteristics of an armed attack are: "the objective must be to undermine the function of a digital information system or network" and that it "must have a political or national security purpose" [5].

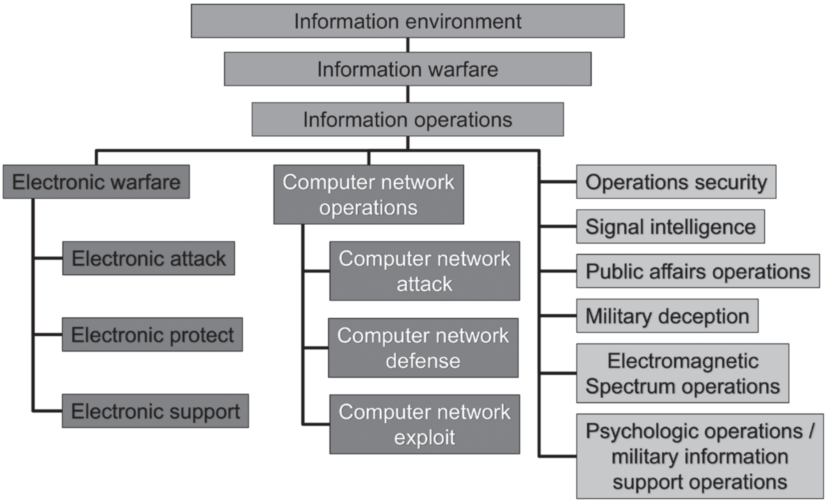
There is also no common agreed definition of information warfare. Figure 1 gives a non-exhaustive overview of information warfare and information operations. Definitions of information warfare could be:

- The use of information technology as an active weapon of war. This includes attempts to intercept, disrupt, and defend military-specific communications, information technology, and critical computer systems.
- The tactical and strategic use of information to gain an advantage.

There is still discussion over whether this is only conducting or defending against electronic attacks on computers and related information systems or whether it also includes the whole spectrum of possibilities for using information effectively in warfare and denying enemies the same capability [6].

Information operations are actions taken to affect adversary information and information systems while defending one's own. They are the integrated employment of the core capabilities of influence operations, electronic warfare operations, network warfare operations, together with specified integrated control enablers, to influence, disrupt, or corrupt adversarial human and automated decision-making while protecting our own.

FIGURE 1: INFORMATION OPERATIONS AND INFORMATION WARFARE [6]



Traditional warfare is characterised as a violent struggle for domination between nation-states or coalitions and alliances of nation-states, or, as Carl von Clausewitz put it in his book *On War* [7], war is "a mere continuation of policy by other means". This confrontation typically involves force-on-force military operations in which adversaries employ a variety of conventional military capabilities against each other in the air, land, maritime, and cyberspace domains. The objective may be to convince or coerce key military or political decision-makers, defeat an adversary's armed forces, destroy an adversary's war-making capacity, or seize or retain territory in order to force a change in an adversary's government or policies [8].

The term "hybrid operations" describes a specific subset of strategy that employs conventional military force supported by irregular and cyber warfare tactics. For hybrid threats, the methods and activities are multidimensional and the links between different actions are unclear. Sometimes they are even impossible to verify. Hybrid threats as such fall short of hybrid warfare, but if they are not detected or responded to, hybrid warfare can ensue [9],Hybrid warfare can be defined as blending conventional warfare and irregular warfare, potentially including cyber warfare and information warfare.

B. Cognitive and physical effects and orders of effects

Like traditional kinetic operations, a cyber operation is likely to cause cognitive effects.[5] Many of the cyber operations we see today have cognitive effects without important physical effects [10]. Cognitive effects of cyber operations include:

[5] Often one equates the terms kinetic and lethal and the terms non-kinetic and non-lethal. There might be a correlation between them, but the other combinations do also occur. This notion is important for effects-based operations in cyber warfare.

sowing confusion, changing behaviour, modifying trust, changing (public) opinion, manipulation, etc. One recent example of this is the Cambridge Analytica data analysis case [11], where social media was used to influence people's behaviour. However, not every cyber operation causes cognitive effects. For example, if a cyber attack, the aim of which is only to exfiltrate information, is not noticed by the target, it does not have cognitive effects until its discovery. Many cyber-targeted attacks are not discovered quickly by the target [12].

Effects have causes and can, in turn, cause further effects. A large number of cause-effect "chains" can be created, based on a single causal event. Cascading effects within the same IT systems are still considered to be first-order effects. Second-order effects are effects outside the IT environment, but within an independent mission (e.g. a factory or an organisation). Those effects represent the indirect effect caused by system failures triggered by the cyber operation: businesses or operations are interrupted, or at least degraded. These second-order effects are not desirable during covert or stealth cyber operations. Third-order effects are long-term. They represent the overall result of the first-order and second-order effects, which may be a change of behaviour in humans or institutions, an impact on international relations or a financial impact. These are the cyber effects that have a profound impact on ongoing operations, on the mission itself, and eventually even on strategic or political levels. In estimating the outcome of a cyber operation, one should not only consider first-order effects, but also examine the relationships between systems in order to estimate second- and third-order effects, which are potentially even more important for the mission and which could have an impact on different levels [13], as explained in subsection 3B.

An example of the different orders could be the NotPetya attack that took place in 2017, where the first-order effects were getting a malware through tax software that companies and individuals required for filing taxes in Ukraine. Among others, there were second-order effects on companies such as Maersk, which had interruptions in its operations that caused financial impact. The third-order effects were that different nations issued statements attributing NotPetya to Russian state-sponsored actors and the United States sanctioned Russian organisations that were believed to have assisted the Russian state-sponsored actors with the operation [14].

C. Measuring the effects of cyber operations

Cyber operations are able to create physical and cognitive effects, and can manifest in various ways: as first-order or higher-order effects; directly or indirectly; immediately or delayed. However, feedback (target damage assessment) as to the success or failure of a cyber effect reaching its destination, or whether the payload had been executed, is not always clear. Relevant questions for measuring the effects of cyber operations:

- Is it possible to detect disturbances in systems, even if the operation itself cannot be immediately detected and characterised?
- What are the effects – intended and actual – of the cyber operation on our own mission's effectiveness, as well as on our strategic interests?
- Is measuring first-order effects, for example by exfiltration of data, possible? Exfiltration of data can be by means of a command and control channel, a beacon installed on the target that provides information about the status of the cyber operation, an insider that leaks data or information, or other means.
- If there is no other way of directly measuring the outcome of a cyber operation, are there measurable second- or third-order effects?

Where possible, traditional kinetic battle damage assessment should also be used for cyber operations in order to integrate cyber effects as much as possible in the traditional warfare terminology [15].

D. Estimating the effects of cyber operations

Many different parameters influence the outcome of a cyber operation, but only some of them are controllable. Examples of such parameters are: the training of friendly forces and adversary personnel in cyber operations, the ability of the adversary to defend its IT infrastructure, the complexity of the systems involved, the accessibility of the targeted system, and so forth. The use of more parameters in identifying and quantifying the effects of cyber operations will result in better predictions.

No process currently exists that is capable of estimating the overall outcome of a cyber operation at mission level. Research has begun, but is still in the early stages of development. There is currently no way to describe dependencies between mission objectives, mission activities, and measurable outcomes. Integrating cyber operations into the overall mission is, for the moment, less effective than desired [2]. The existing, publicly-available modelling schemes deal with very specific scenarios based on attack graphs [16], game theory [17]–[19], extensions to traditional models of combat [20], modelling and prediction of several system properties using Monte Carlo models [21], or practical guides on how to better defend IT systems [22]. It is probably hard to create a model in the first place, but even if the model creation were doable, it would still be hard to measure and to validate it.

Analysis techniques are crucial in determining the decision metrics required to estimate the potential effects of cyber operations. In the development of decision metrics, the following is essential: physical or digital paths toward the final target, estimation of the probability of success, judgment about the amount of collateral damage that might be caused, and assessment of likely first-, second-, and third-order effects.

A framework for assessing cyber war that builds on the elements of risk assessment was proposed by Dorothy Denning [23]. However, for such a framework to be useful, as stated in the paper, there is a need for measurable metrics. In order to develop those decision metrics, one could use information security modelling and simulation tools to simulate a system's security baseline configuration and then test the outcome of the cyber effect in a simulated environment. There is a lack of publicly-available documentation about modelling methods and metrics for missions in cyberspace. Some likely reasons are:

- Test data is needed to validate a model or a technique, however data is not abundantly available for a mission and could be of a confidential nature.
- Impacts are often difficult to measure, even in laboratory conditions. Defining which parameters are relevant to measure is in itself a complicated matter, certainly for cognitive impacts, collateral damage and higher order effects.
- Cyberspace is highly dynamic, and often asymmetrical in nature [24].
- Rapid evolution of software and patching policies makes it harder to keep the cyber effects of technical solutions up to date and could reduce the time available for simulation.
- Most nations are developing internal capabilities due to the very sensitive nature of the topic. Therefore, there is no amplification factor of knowledge through information sharing.

E. Estimating the collateral damage of cyber operations

Like conventional kinetic weapons, cyber effects can cause collateral damage. In kinetic warfare, collateral damage is well understood, and policy, procedures, and national and international legislation are available. Collateral damage occurs when a military action causes unintended physical damage to civilian persons or objects [25]. Collateral damage in this context is not only used in relation to war and the laws of armed conflict, but also below this threshold with the prohibition of the use of force among states. When estimating the outcome of a cyber operation, collateral damage is a factor that must be taken into account. More and more, existing international law is accepted as applicable in cyberspace (at least by western societies) [26].

An example of regulation from the US Department of Defense clearly stipulates the procedure to avoid unnecessary collateral damage [27]:

- Identify the target.
- Determine whether protected persons or objects are within range of the target.
- Estimate the collateral damage that will occur.

- Determine whether there are other weapons that can accomplish the objective with less collateral damage.
- Evaluate whether the anticipated collateral damage exceeds the concrete and direct military advantage. Advantages that are hardly perceptible or would only appear in long-term view are to be disregarded.

This rule set is also applicable to cyber warfare or a cyberattack, but estimating the collateral damage that occurs might be very difficult to achieve in certain cases. Questions to take into consideration before planning cyber operations are:

- What to do when an operation unintentionally modifies data? What is the relevance and importance of the data concerned? Is the data related to lifesaving or loss of life? Can altering the data cause physical injury? Does the data contain private information? Has the data a military use?
- Is it possible to predict the outcome with confidence?
- What if the cyber effect, even after a long period of time, penetrates friendly forces in national industry or governmental institutions? In other words, is it possible that our own systems are vulnerable?
- Can the second- and third-order effects be predicted?

Although in theory IT systems should be deterministic as they are built on logic, in practice it is currently not possible to formally analyse a complete IT system, due to their complexity. They are often a system of systems and the components, hardware and software, are mostly built by different manufacturers. Even installation on site is often done by a wide range of employees from different companies. The effect of a cyber operation on such a complex system and the subsequent cascading effects are hard to predict.

Up to now, legal entities have not engaged much with this issue [28]. The *Tallinn Manual 2.0* states [29]in "Rule 113 – Proportionality" that "The issue is of particular relevance in the context of cyber attacks in that it is sometimes quite difficult to reliably determine likely collateral damage in advance". It has to be mentioned that stress, irritation, fear or inconvenience are not considered as collateral damage, but cyber operations can cause those effects. There are examples where measures were taken to limit the collateral damage: for example, by assuring that the cyber operation is specific enough to affect only the target and will become inactive after a specific date, collateral damage can be reduced [30]. The following questionnaire, based on one created by the Obama administration in 2014 [31], is useful for estimating the impact of cyber collateral damage:

- How much is the targeted vulnerability present in the core IT infrastructure, in critical IT infrastructure, in coalition members' IT infrastructure, and in national IT systems?
- Does the vulnerability, if left unpatched, pose significant risk for national or allied systems (military and civilian)?
- Does the cyber effect impact the complete system or only specific subsystems?
- How much harm could an adversary nation or criminal group do with knowledge of this vulnerability?
- Does patching this vulnerability provide information that can be used by adversaries?
- How likely is it that we would know if someone else was exploiting this vulnerability?
- How badly do we need the intelligence we think we can get from exploiting the vulnerability?
- Could we utilise the vulnerability for a short period of time before we disclose it?
- How likely is it that someone else will discover the vulnerability?
- Can the vulnerability be patched or otherwise mitigated?

3. STEP 2: ACHIEVING CYBER EFFECTS: ENDS, WAYS, AND MEANS

According to Major General Dennis Laich [32]:

> "Ends are defined as the strategic outcomes or end states desired. Ways are defined as the methods, tactics, and procedures, practices, and strategies to achieve the ends. Means are defined as the resources required to achieve the ends, such as troops, weapons systems, money, political will, and time. The model is really an equation that balances what you want with what you are wiling [sic] and able to pay for it or what you can get for what you are willing and able to pay."

This section will comment on how the traditional application of 'ends', 'ways' and 'means' deviates from traditional warfare in the context of cyber operations.

A. Preparation of cyber effects

Levels of war such as strategic, operational and tactical levels were introduced in order to enhance decision-making processes and to allow greater efficiency in the execution of tasks. The outcomes of kinetic warfare get more specific when moving from the strategic to the tactical level, i.e. the impact of the outcomes and the responsibilities are more limited. There is no known equivalent simplification in the planning of cyber warfare [13]. It is not even known to what extent cyber operations are achieving their effects [2], this, however, is the objective of this paper: to help nations to find it out.

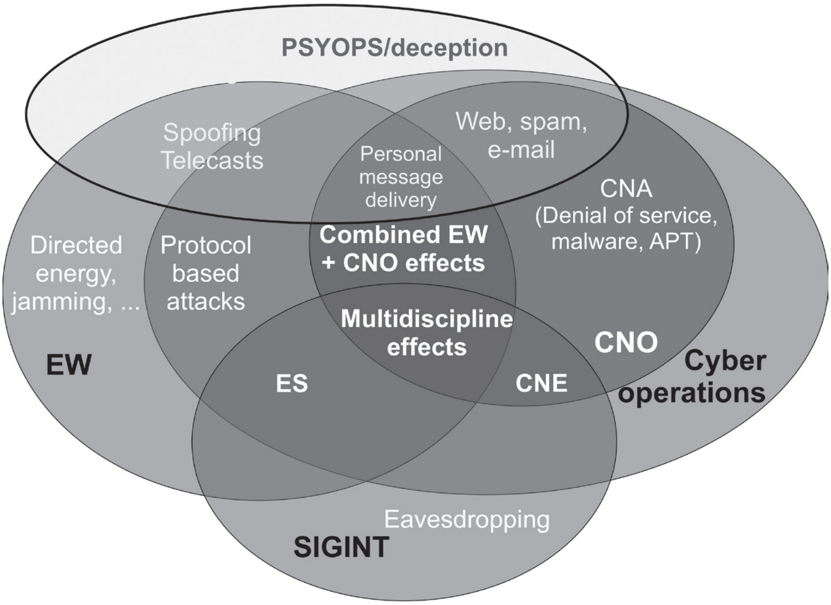
The preparation time of cyber effects can range from very long to almost immediate, depending on the effects to be achieved, the target system, access to information, cyber skills of personnel, etc. Physical distances are often irrelevant.

It is also important for nations to invest in training personnel, e.g. cyber operations training in military academies, exchange programmes of military cadets with technology universities for particular classes, etc. Certainly the areas of Cyber Network Operation, as explained in Figure 2, should have a focus.

It is important for nations to include cyber operations capabilities in their (grand) strategy, but there are not enough resources to prepare and provide cyber effects for all imaginable scenarios. Development of these capabilities may take a considerable amount of time; the focus on and prioritising of which cyber effects are key to an operation and should be performed well in advance of the operation.

Technical aspects, including technical skills, are a critical factor of a cyber operation. A cyber effect is linked to the technical characteristics of the chosen solution. It is often the case that a specific technological solution is most suitable to ensure a specific effect. Therefore, the achievable cyber effects are dependent on available technical know-how. Deciding on the areas in which technical knowledge should be developed has to be planned a long time in advance and incorporated in a strategy. Although not all IT-related capabilities, ranging from electronic warfare, signal intelligence to cyber operations, are subjects of this paper, a national strategy should not only include cyber effects, but all of them. Isaac Porche's [6] overview of the relevant functional areas is provided in Figure 2.

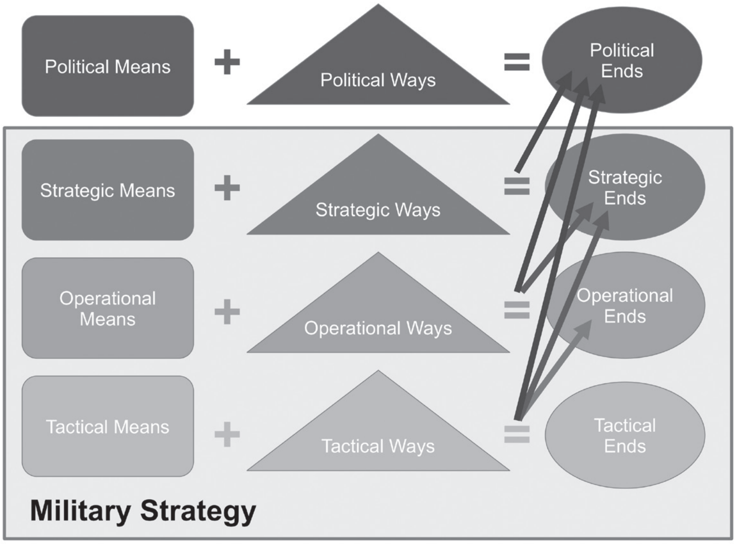
FIGURE 2: FUNCTIONAL VIEW OF CYBER EFFECTS. IT SHOWS THE OVERLAPS AMONG ELECTRONIC WARFARE (EW), SIGNAL INTELLIGENCE (SIGINT) AND CYBER OPERATIONS. COMPUTER NETWORK OPERATIONS (CNO) ENCOMPASSES COMPUTER NETWORK EXPLORATION (CNE) AND COMPUTER NETWORK ATTACK (CNA). ELECTRONIC SUPPORT (ES) ARE TECHNIQUES SUCH AS DIRECTION FINDING OF ELECTROMAGNETIC SOURCES [6].



B. 'Ends' in cyber operations

The use of cyber means and ways at one level of military strategy can potentially impact higher levels of strategy, even reaching the political, especially with regard to cognitive, economic or societal effects. This is not unique to cyber operations, but here this spill-over might be more difficult to predict. Therefore, cyber operations should be authorised by someone responsible for the highest potential level of impact. This is illustrated in Figure 3, based on a graphic by Murat Balci [13], which applies to kinetic warfare. The dotted lines present the spill-over that can happen in cyber operations.

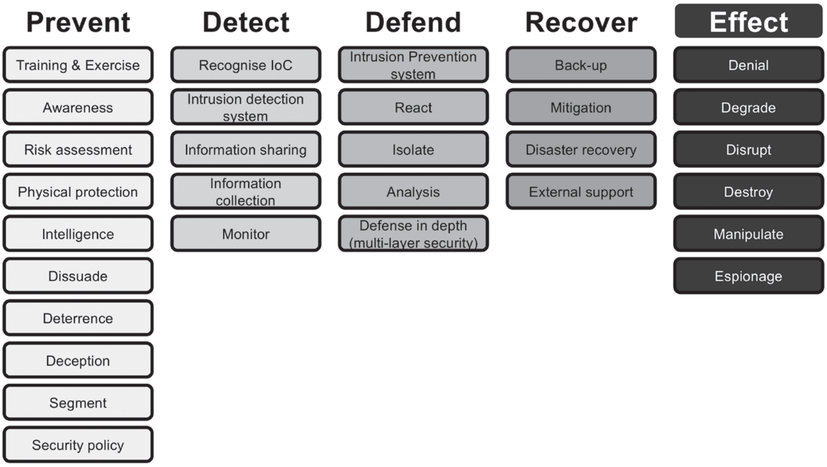
FIGURE 3: WAYS AND MEANS ON TACTICAL, OPERATIONAL, AND STRATEGIC LEVELS CAN RESULT IN ENDS ON A HIGHER LEVEL [13]



Cyber operations are able to destroy, degrade, deny, and disrupt information technology-dependent infrastructures and data. They can be used in espionage and manipulation. Often, they cause cognitive effects with few physical effects [33]. Cyber operations can be used against a specific target (e.g. Stuxnet [30], [34]) or indiscriminately (e.g. Wannacry [35]).

First, from a defensive point of view, there are tasks that should be undertaken or for which training should be provided, in order to deter and to detect adversary cyber operations. There are actions to be taken to defend against cyber operations and to recover from a successful adversary cyberattack, if defence has failed. It is a wise approach to start planning cyber operations only when all of those defensive tasks are taken care of. A non-exhaustive list of tasks in cyberspace in Figure 6 is proposed to the community for discussion. This list is developed from the concepts in NIST's "Framework for Improving Critical Infrastructure Cybersecurity": identify, protect, detect, respond and recover [36]. It is perfectly possible that some tasks are necessary for different purposes or that some tasks are not relevant in some situations.

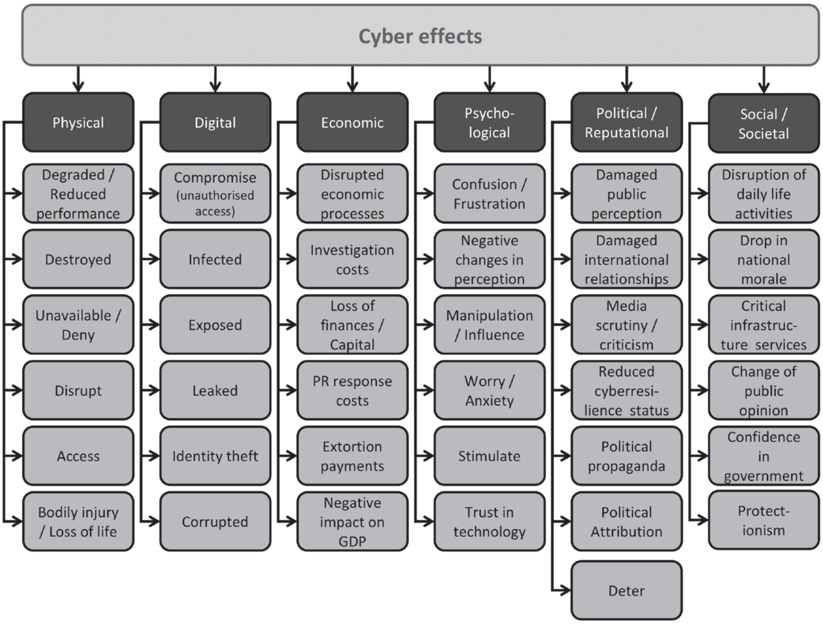
FIGURE 4: TASKS IN CYBERSPACE [36]



Prevent	Detect	Defend	Recover	Effect
Training & Exercise	Recognise IoC	Intrusion Prevention system	Back-up	Denial
Awareness	Intrusion detection system	React	Mitigation	Degrade
Risk assessment	Information sharing	Isolate	Disaster recovery	Disrupt
Physical protection	Information collection	Analysis	External support	Destroy
Intelligence	Monitor	Defense in depth (multi-layer security)		Manipulate
Dissuade				Espionage
Deterrence				
Deception				
Segment				
Security policy				

A taxonomy of cyber effects, based on previous work by Agrafiotis [37], is shown in Figure 5. This taxonomy takes into account further effects, such as economic and reputational, while the original taxonomy focused on cyberattacks on commercial enterprises. In this paper, the authors have designed a new taxonomy from the perspective of a nation-state.

FIGURE 5: ENCOMPASSING TAXONOMY OF CYBER EFFECTS [37]

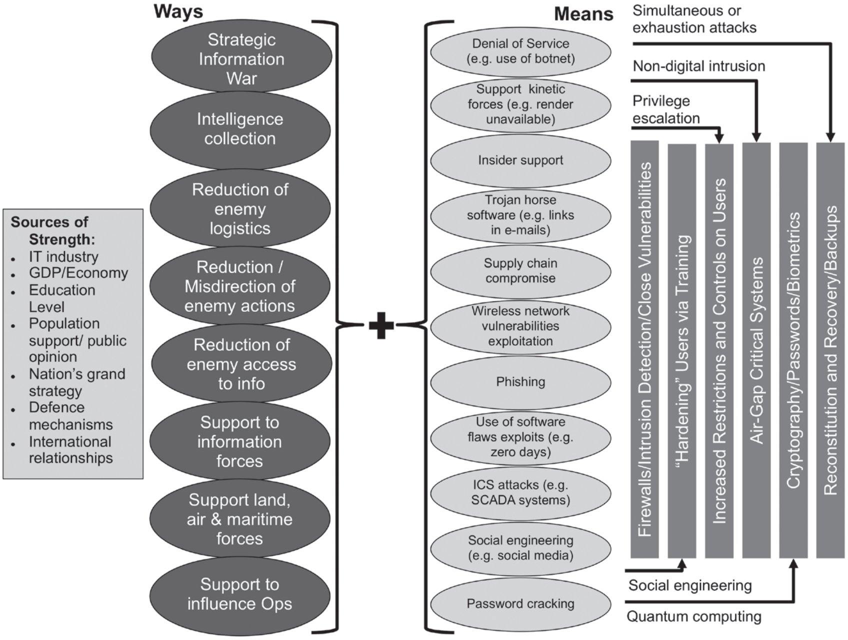


Physical	Digital	Economic	Psycho-logical	Political / Reputational	Social / Societal
Degraded / Reduced performance	Compromise (unauthorised access)	Disrupted economic processes	Confusion / Frustration	Damaged public perception	Disruption of daily life activities
Destroyed	Infected	Investigation costs	Negative changes in perception	Damaged international relationships	Drop in national morale
Unavailable / Deny	Exposed	Loss of finances / Capital	Manipulation / Influence	Media scrutiny / criticism	Critical infrastruc-ture services
Disrupt	Leaked	PR response costs	Worry / Anxiety	Reduced cyberresi-lience status	Change of public opinion
Access	Identity theft	Extortion payments	Stimulate	Political propaganda	Confidence in government
Bodily injury / Loss of life	Corrupted	Negative impact on GDP	Trust in technology	Political Attribution	Protect-ionism
				Deter	

C. ‘Means’ and ‘Ways’ in cyber operations

Figure 6 proposes a cyberspace superiority model based on the work of William Bryant [38]. The image shows that ways and means can be flexibly mixed in order to achieve the desired outcomes, and that some means are more useful to overcome some typical technical challenges. What adds to the complexity is that they do not get their strength only from military capabilities, but more broadly from the society.

FIGURE 6: CYBERSPACE SUPERIORITY MODEL, WHICH DESCRIBES MEANS, WAYS, AND DEFENSIVE BLOCKS [38]


Ways
Strategic Information War
Intelligence collection
Reduction of enemy logistics
Reduction / Misdirection of enemy actions
Reduction of enemy access to info
Support to information forces
Support land, air & maritime forces
Support to influence Ops
Sources of Strength:
IT industry
GDP/Economy
Education Level
Population support/ public opinion
Nation’s grand strategy
Defence mechanisms
International relationships
+
Means
Denial of Service (e.g. use of botnet)
Support kinetic forces (e.g. render unavailable)
Insider support
Trojan horse software (e.g. links in e-mails)
Supply chain compromise
Wireless network vulnerabilities exploitation
Phishing
Use of software flaws exploits (e.g. zero days)
ICS attacks (e.g. SCADA systems)
Social engineering (e.g. social media)
Password cracking
Simultaneous or exhaustion attacks
Non-digital intrusion
Privilege escalation
Firewalls/Intrusion Detection/Close Vulnerabilities
“Hardening” Users via Training
Increased Restrictions and Controls on Users
Air-Gap Critical Systems
Cryptography/Passwords/Biometrics
Reconstitution and Recovery/Backups
Social engineering
Quantum computing

D. Collaboration in cyber operations

There are a multitude of reasons why different actors would assist each other in achieving a particular end state on a specific target, and this has not changed with the emergence of cyberspace. There are huge differences in nations’ capabilities to develop and launch cyber operations, so it makes sense that some state actors are willing to offer cyber effects to other nations or organisations. The sharing of technical and operational details of cyber operations is very sensitive, which makes collaboration difficult. To some extent, the use of cyber operations could be compared with the use of Special Forces: there is an agreement on which effects one wants to achieve, but very little or no information will be shared about how this will be, or has been, executed.

Communities are built to share information about defending networks and computer systems. Most of them are public fora that share known vulnerabilities. They exist in the public domain, like MISP (NATO's malware information sharing platform), and in the private sector, where most software security companies post discovered vulnerabilities on their websites. In the open source community there are fora for sharing information, as well as initiatives like Metasploit [39], which is a framework that includes known vulnerabilities for the purposes of software penetration testing. The public sharing of this kind of information, even information gained from offensive penetration testing, is done from a defensive point of view.

In cyber warfare some cyber effects are usable only once. For example, if a cyber effect is based on the use of a zero-day vulnerability, sharing this information could render this exploit unusable if it were leaked or used elsewhere.

Coalition partners must be informed about capabilities. This is needed in order to understand expected cyber effects, to estimate both the probability of success, the expected collateral damage, and it creates trust. The aim of the exchange of additional information is not to replace any existing targeting procedures, but rather to enhance the 'capabilities analysis' aspect for cyber effects. A more legal approach to targeting in cyberspace can be found in 'Joint and combined targeting: system and process' [40].

Where possible one should express the desired effect by using existing terminology from kinetic warfare. Terminology describing effects, such as 'deny', 'degrade', 'disrupt', and 'destroy' can be used for cyber effects. Some additional effects, such as those mentioned in Figure 5, are more specific to cyber effects. This paper endorses the use, as much as possible, of existing terminology and procedures, because this integrates cyber effects more smoothly into the existing military decision-making process and facilitates the comparison of cyber effects with other means of achieving an objective.

4. RECOMMENDATIONS AND FUTURE WORK

A. Measuring and predicting cyber effects

1. Cyber effects can easily trigger outcomes, wanted or unwanted, on a strategic or a political level. Ensure that the use of cyber operations is authorised by the strategic level that aligns with the estimated ends of the cyber effect. This implies that the Rules of Engagement (RoE) delegation for cyber operations should be kept at a corresponding level. This diverges

from traditional kinetic warfare, where responsibility is delegated and use of force is limited in cascade, by using specific rules of engagement for each decision level.

2. A solid understanding of the expected outcome of the first-, second-, and third-order cyber effects will facilitate better decision-making as to whether a cyber effect is the best course of action to reach a specific goal, and will increase the effectiveness of its use.
3. Define the desired cyber effects with the fewest technical terms and use existing terminology whenever possible. A good understanding of the strengths and weaknesses of the technical possibilities of cyber effects is crucial for the cyber advisor, who should be capable of translating into non-technical language. Appropriate visualisation tools should be put in place in order to have more situational awareness; this will help with the battlefield damage assessment caused by a cyber effect.

B. Achieving cyber effects

1. The cyber operation will be one possible course of action for a commander. A cyber operation is not a 'silver bullet' that will provide a solution when traditional means are not able to achieve a desired end state. On the other hand, if a cyber or hybrid operation has been integrated into the planning process from the beginning, it could be the best option a commander has for executing a specific task or achieving a desired end state.
2. Before considering investment in cyber operations capabilities, the ideal situation is to verify that the mission's IT infrastructure is not vulnerable to cyber exploits, and that mission assurance is guaranteed from a cyber perspective. Therefore, as explained in Figure 6, if feasible, all cyber tasks in support of prevention, detection, defence, and recovery should be covered before enabling cyber operations, but everything depends on the risk assessment and the capabilities of the adversary.
3. Guaranteeing mission assurance from a cyber perspective implies that the mission's IT infrastructure is not vulnerable to the developed cyber exploits, but also to a possible counter cyber effect originating from the target as a response to the initial cyber operation, also called a 'hack-back'.
4. On a national political level, having a long-term political vision of what type of cyber capability should be developed is key. There are not enough human resources nor financial means to build every possible cyber capability in advance. On a political level, the following questions should be addressed: "What is the main focus of the cyber capabilities?" and "How much effort will be invested in research, and for which types of cyber capability?"

5. When cyber effects are needed, there could be too little time for development if they have not been prepared and trained for in advance. Creating high-impact, targeted, cyber capability with limited collateral damage requires substantial preparation time.
6. Ensure that a cyber advisor is present at every level of decision-making. Cyber operations are technical in nature and it is a challenge to translate those technical aspects into operational language in terms of 'means', 'ways' and 'ends'.
7. Make sure that the cyber domain is involved in the planning process as early as possible, from the beginning of the planning of every campaign. This will optimise the outcome of cyber effects.

C. Future work

1. Until now there has been little input from the legal community with regard to collateral damage in cyberspace. Legal specialists should develop this topic in more detail.
2. If a cyber effect will be delivered by a supporting nation, there is a need to streamline the coordination and exchange of information. Nations delivering voluntary sovereign contributions need to receive information from the mission commander, and the mission commander needs feedback from the supporting nation in order to integrate this in the planning process. Information sharing about this among allies is not yet well developed. A framework should be developed that defines the essential elements of information to be exchanged, and describes how to do this. Due to the sensitive nature of the technical details of cyber capabilities, more focus should be put on ends and effects rather than on technical aspects, without neglecting the specificities of a cyber effect. The level of detail of the information coming from the national sovereign contribution to the mission must allow the mission commander to be confident that the desired end goals will be achieved, that mission assurance is not compromised, and that collateral damage is under control.

5. CONCLUSIONS

Cyber operations are advisable when the effects are planned well in advance, and when one's own systems are well protected. Achieving cyber effects should take into account that it is difficult to estimate the spill-over to other levels of warfare. Support from one nation to another or to a coalition in order to achieve cyber effects sounds promising, but publicly known procedures how to achieve this do not exist yet.

Whoever can influence the portion of cyberspace used by the adversary has the potential to control the adversary. Cyber effects can, in specific circumstances, be the most effective tool to disrupt, degrade, corrupt, influence, etc. an adversary's ability to conduct military operations. The ability to accurately estimate the impact of cyber effects is currently limited. Just as in kinetic warfare, estimations and measurements of outcomes of cyber effects are essential in planning operations because they allow decision-makers to optimise the outcome and to limit the unwanted effects or collateral damage.

REFERENCES

[1] R. C. Parks and D. P. Duggan, "Principles of Cyberwarfare," in *Workshop on Information Assurance and Security*, 2001, pp. 122–125 on page 122.

[2] S. Musman, A. Temin, M. Tanner, D. Fox, and B. Pridemore, "Evaluating the Impact of Cyber Attacks on Missions", MITRE Corp., 2010.

[3] J. Döge, "Cyber Warfare Challenges for the Applicability of the Traditional Laws of War Regime" *Arch. des Völkerrechts*, vol. 48, no. 4, pp. 486–501, 2011on page 488.

[4] S. A. Hildreth, "CRS Report to Congress: Cyberwarfare" 2001 in footnote 3. [Online]. Available: https://fas.org/sgp/crs/intel/RL30735.pdf. [Accessed: 19-Feb-2019].

[5] O. A. Hathaway et al., "The law of cyber-attack" *Yale Law Sch. Fac. Scholarsh. Ser.*, pp. 817–886, Jan. 2012 on page 883.

[6] I. R. Porche et al., "Redefining Information Warfare Boundaries for an Army in a Wireless World" RAND Corporation, 2013 on page 15.

[7] K.V. Clausewitz, *On war*. Translated to English in 1943 by Jolles, on page 16.

[8] US Air Force, US Air Force doctrine, 2015 in Vol 1 Basic Doctrine. [Online]. Available: https://www.doctrine.af.mil/Core-Doctrine/Vol-1-Basic-Doctrine/. [Accessed: 19-Feb-2019].

[9] The European Centre of Excellence for Countering Hybrid Threats, "Blog: Hybrid threats – what are we talking about?" www.hybridcoe.fi, 2017. [Online]. Available: https://www.hybridcoe.fi/hybrid-threats-what-are-we-talking-about/. [Accessed: 19-Feb-2019].

[10] M. C. Libicki, "The Convergence of Information Warfare" *Strateg. Stud. Q.*, vol. 11, no. 1, pp. 49–65, 2017 on page 54.

[11] C. Cadwalladr and E. Graham-Harrison, "How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool" *The Guardian*, International edition, 17-Mar-2018.

[12] Mandiant, "M-TRENDS2018" 2018 in "2017 by the numbers", "Dwell time". [Online]. Available: https://www.fireeye.com/content/dam/collateral/en/mtrends-2018.pdf. [Accessed: 19-Feb-2019].

[13] M. Balci, M. Canan, and G. Kucukkaya, "Defining military levels for cyber warfare by using components of strategy: ends, ways, and means" in 21st ICCRTS "C2 in a Complex Connected Battlespace"'," 2016, pp. 1–13 on page 5.

[14] Council on Foreign Relations, "NotPetya" 2017. [Online]. Available: https://www.cfr.org/interactive/cyber-operations/search?keys=not+petya. [Accessed: 19-Feb-2019].

[15] R. A. Martino, "Leveraging traditional battle damage assessment procedures to measure effects from a computer network attack" Air Force Institute of Technology, 2011 on page iv.

[16] I. Kotenko and A. Chechulin, "A cyber attack modeling and impact assessment framework" in *5th International Conference on Cyber Conflict (CyCon)*, 2013.

[17] B. K. Mishra and H. Saini, "Cyber Attack Classification using Game Theoretic Weighted Metrics Approach" World Appl. Sci. J., vol. 7, no. *Special Issue of Computer & IT*, pp. 206–215, 2009.

[18] B. Edwards, A. Furnas, S. Forrest, and R. Axelrod, "Strategic aspects of cyberattack, attribution, and blame" *Proc. Natl. Acad. Sci.*, vol. 114, no. 11, pp. 2825–2830, 2017.

[19] M. Jones, G. Kotsalis, and J. S. Shamma, "Cyber-attack forecast modeling and complexity reduction using a game-theoretic framework" in *Control of Cyber-Physical Systems*, Springer, 2013, pp. 65–84.

[20] F. Yıldız, "Modeling the effects of cyber operations on kinetic battles" *Engineering*, pp. 32–37, 2014.

[21] P. Johnson, J. Ullberg, M. Buschle, U. Franke, and K. Shahzad, "An architecture modeling framework for probabilistic prediction" *Inf. Syst. E-bus. Manag.*, vol. 12, no. 4, pp. 595–622, Nov. 2014.

[22] W. S. Powell, “Methodology for Cyber Effects Prediction” in Black Hat USA, 2010.

[23] D. E. Denning, “Assessing Cyber War” in *Assessing War*, L. J. Blanken, H. Rothstein, and J. J. Lepore, Eds. Georgetown University Press, 2015, pp. 266–284.

[24] A. Phillips, “The asymmetric nature of cyber warfare” *USNI News*, Feb-2013. [Online]. Available: https://news.usni.org/2012/10/14/asymmetric-nature-cyber-warfare#more-785. [Accessed: 23-Feb-2019].

[25] S. Mele, “Cyber-Weapons: Legal and Strategic Aspects (Version 2.0)” *Instituto Italiano Di Studi Strategici,* 2013 on page 13. [Online]. Available: https://papers.ssrn.com/sol3 /papers.cfm?abstract_id=2518212. [Accessed: 19-Feb-2019].

[26] B. J. Egan, “International Law and Stability in Cyberspace” *Berkeley J. Int. Law*, vol. 35, no. 1, pp. 169–181, 2017.

[27] DoD Joint Chiefs of Staff, No-strike and the collateral damage estimation methodology. U.S.A., 2012.

[28] S. Romanosky and Z. Goldman, “Cyber Collateral Damage,” *Procedia Comput. Sci*., vol. 95, pp. 10–17, 2016.

[29] M. N. Schmitt, *Tallinn Manual 2.0* on the international law applicable to cyber operations. 2017 on page 475.

[30] N. Falliere, L. O. Murchu, and E. Chien, “W32.Stuxnet Dossier” 2011.

[31] M. Daniel, “Heartbleed: Understanding when we disclose cyber vulnerabilities” White House blog, April, 2014. [Online]. Available: https://obamawhitehouse.archives.gov/blog/2014/04/28/heartbleed-understanding-when-we-disclose-cyber-vulnerabilities. [Accessed: 23-Feb-2019].

[32] D. Laich, “Ends = Ways + Means” mglaich, 2010. [Online]. Available: http://mglaich.blogspot.com/2010/07/ends-ways-means.html. [Accessed: 23-Feb-2019].

[33] S. Goel, “Cyberwarfare: Connecting the Dots in Cyber Intelligence” *Commun. ACM*, vol. 54, no. 8, pp. 132–140, Aug. 2011.

[34] J. R. Lindsay, “Stuxnet and the Limits of Cyber Warfare” *Secur. Stud*., vol. 22, no. 3, pp. 365–404, 2013.

[35] S. Mohurle and M. Patil, “A brief study of Wannacry Threat: Ransomware Attack 2017” *Int. J. Adv. Res. Comput. Sci*., vol. 8, no. 5, pp. 1938–1940, 2017.

[36] NIST (National Institute of Standards and Technology), “Framework for improving critical infrastructure cybersecurity” Feb. 2014 on page 19.

[37] I. Agrafiotis, J. R. C. Nurse, M. Goldsmith, S. Creese, and D. Upton, “A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate” *J. Cybersecurity*, pp. 1–15, 2018 on page 8.

[38] W. D. Bryant, “Cyberspace Superiority: A Conceptual Model,” *Air Sp. Power J*., pp. 25–44, 2013 page 37.

[39] F. Holik, J. Horalek, O. Marik, S. Neradova, and S. Zitta, “Effective penetration testing with Metasploit framework and methodologies” *CINTI 2014 - 15th IEEE Int. Symp. Comput. Intell. Informatics, Proc*., pp. 237–242, 2014.

[40] M. N. Schmitt, J. Biller, S. C. Fahey, D. Goddard, and C. Highfill, “Joint and combined targeting: system and process” 2016 pages 13 to 19. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2830229. [Accessed: 19-Feb-2019].

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Rough-and-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing

Jason Healey
Senior Research Scholar
Columbia University, SIPA
New York, NY USA
jh3639@columbia.edu

Neil Jenkins
Chief Analytic Officer
Cyber Threat Alliance
Arlington, VA USA
neiljenkins@cyberthreatalliance.org

Abstract: This paper addresses the recent shift in the United States' policy that emphasizes forward defense and deterrence and to "intercept and halt" adversary cyber operations. Supporters believe these actions should significantly reduce attacks against the United States, while critics worry that they may incite more adversary activity. As there is no standard methodology to measure which is the case, this paper introduces a transparent framework to better assess whether the new U.S. policy and actions are suppressing or encouraging attacks.[1]

Determining correlation and causation will be difficult due to the hidden nature of cyber attacks, the veiled motivations of differing actors, and other factors. However even if causation may never be clear, changes in the direction and magnitude of cyber attacks can be suggestive of the success or failure of these new policies, especially as their proponents suggest they should be especially effective. Rough-and-ready metrics can be helpful to assess the impacts of policymaking, can lay the groundwork for more comprehensive measurements, and may also provide insight into academic theories of persistent engagement and deterrence.

Keywords: *cyber deterrence, metrics, cyber conflict, cyber operations, threat intelligence, cyber policy, persistent engagement*

[1] This work was funded in part by the Office of Naval Research under the OSD Minerva program: Grant number N00014-17-1-2423.

1. INTRODUCTION

The United States has significantly shifted its policy regarding the Department of Defense (DoD)'s role in cyberspace to emphasize "persistent presence," to remain in "in foreign cyberspace to counter threats as they emerge" and to "intercept and halt cyber threats."[2] The belief is that, over time, these actions will cause nation state adversaries (particularly Russia, China, Iran, and North Korea) to become less effective; they will be forced to expend more resources on defense and will choose not to engage the United States.

Beyond such active engagement with adversaries, the new policy also seeks to impose costs through deterrence, both outside and inside cyberspace. The measures outside cyberspace include actions like sanctions and indictments, while those inside include gaining access to systems that adversaries value, to hold them at risk with offensive cyber operations. In the words of John Bolton, the National Security Advisor, the White House has "authorized offensive cyber operations […] not because we want more offensive operations in cyberspace, but precisely to create the structures of deterrence that will demonstrate to adversaries that the cost of their engaging in operations against us is higher than they want to bear."[3]

Supporters believe these are long-awaited steps which will significantly reduce transgressions against the United States. Critics believe such counteroffensive activities may only inflame nation state adversaries, who could see them not as a mild corrective but as a fresh insult which demands a response. There is currently no standard methodology to measure whether the new U.S. policy and actions are suppressing or encouraging attacks. While it would be routine for a military command like U.S. Cyber Command to have measures of effectiveness for specific military operations, this is not necessarily true for assessments of the policy outcomes. To this end, Representative James Langevin (D, RI-2) is pushing for such measures: "Much like the traditional battlefield, we must measure the impact of our operations to assess our warfighting effectiveness towards the larger objectives and ensure our strategic vision reflects the realities of engagement in cyberspace."

[2] Nakasone, Paul M. 2019. "An Interview with Paul M. Nakasone," Joint Forces Quarterly. https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-92/jfq-92.pdf.

[3] Bolton, John. 2018. "Transcript: White House Press Briefing on National Cyber Strategy - Sept. 20, 2018." Washington DC (September 8). Available at https://news.grabien.com/making-transcript-white-house-press-briefing-national-cyber-strateg.

[4] Langevin, James R. 2019. "Opening Statement: FY 2020 Budget Request for Military Operations in Cyberspace." March 13. https://armedservices.house.gov/_cache/files/d/5/d5f94725-3373-40ef-803c-1f0ff8f106a8/577D710BF48F37825B2656EE1AF6891A.opening-statement---ietc-chairman-langevin-3-13-2019.pdf.

This paper is intended to help bridge this gap and has four related goals:

1. *Stimulate the conversation*. Despite significant commentary and research on the new policy, there has been little discussion on how to assess if it is working as expected or not.
2. *Propose basic metrics* which might suggest if the new policy is working as expected to dissuade attacks or is actually encouraging them. Even simple metrics might make some causal explanations more or less likely, even though determining strong correlation (much less causation) may be distant goals.
3. *Introduce a basic framework* of terms and concepts. Security threat analysts, policymakers, and researchers need an analytical structure to make it easier to weigh evidence and make conclusions.
4. *Encourage more complex, data-driven approaches* from those who may be able to determine causation or correlation, such as the U.S. Intelligence Community and the commercial cybersecurity threat intelligence community.

It is obviously hard to prove whether any kind of policy to influence adversaries is working or not. It is still not definitively settled if the lack of Cold War nuclear attacks between the United States and the Soviet Union was the result of deterrence or a lucky coincidence. It has been three years since the U.S. and Chinese presidents agreed to limit cyber espionage for commercial benefit, and the cyber-threat and policy communities continue to debate if the Chinese did in fact reduce such espionage and whether any such changes were meaningful or due to the agreement (or other U.S. actions such as indictments).[5] For both nuclear and cyber attacks, it is fortunately easier to measure failure than success. Successful policies may succeed quietly but fail explosively. A skyrocketing increase in Chinese espionage operations after the Obama-Xi agreement would have been a far clearer signal than the apparent decrease.

The metrics in this paper have obvious shortcomings: they cannot prove causality and cannot usually be based on specific U.S. actions, which will likely be classified – such as threats expressed privately to adversaries or counter-offensive disruptions by U.S. Cyber Command. Usually, only the overall policy and pace of adversary attacks will be known, at least from public sources. Attribution and adversaries' decision calculus in many cases cannot be understood quickly. These shortcomings can all be minimized with the right framework and by comparison with additional data sources to address key questions.

Rough-and-ready metrics, including determining the direction and magnitude of any changes over time, are needed to assess the impacts of cyber policymaking. The U.S. military is already conducting these operations, so policymakers need good enough

[5] Segal, Adam. 2016. "The U.S.-China Cyber Espionage Deal One Year Later." Council on Foreign Relations, September 28. https://www.cfr.org/blog/us-china-cyber-espionage-deal-one-year-later.

metrics now to assess the overall effort.[6] With only marginal changes in terminology, this framework can also be useful for efforts to advance "digital peace," such as those by Microsoft and France.[7]

This paper proceeds by first introducing the new U.S. government policies and examining issues of measurement. Then, it discusses several frameworks, starting from simple, illustrative examples, to more fuller descriptions of categories of transgressions. It addresses shortcomings of the framework before a short conclusion and recommendation for future work.

2. THE DOD POLICIES

The new DoD strategy is based on "persistent presence", in part to "intercept and halt" adversary operations – imposing costs on their *current* operations – as well as outright deterrence so that they will choose not to undertake *future* operations, as they fear the costs imposed by the United States will be "higher than they want to bear."[8]

As an example of what this might mean in practical terms, if Iranian cyber operators were gathering resources to conduct further disruptive campaigns against the United States financial sector (as they did in 2011-2012), U.S. Cyber Command could seek to disrupt their efforts in foreign cyberspace, up to and including counter-offensive cyber operations.[9] In the short term, this would impose "tactical friction", dissuading the Iranians as they have to defend themselves and expend resources to rebuild the disrupted capabilities and infrastructure. These operations for persistent presence and forward defense would be heightened with actions specifically aimed at cyber deterrence, such as U.S. Cyber Command holding Iranian critical infrastructure at risk of a counter-attack with offensive cyber operations.

While such actions for persistent presences are an innovation in cyber conflict, applying concepts of deterrence is not. A recent Defense Science Board task force characterized cyber deterrence as actions "affecting the calculations of an adversary … to convince adversaries not to conduct cyber attacks or costly cyber intrusions."[10] Deterrence in cyberspace is a complex web of deterrence by denial (actions that reduce

[6] Barnes, Julian E. 2018. "U.S. Begins First Cyberoperation Against Russia Aimed at Protecting Elections." The New York Times, October 23. https://www.nytimes.com/2018/10/23/us/politics/russian-hacking-usa-cyber-command.html.

[7] See Microsoft, Digital Peace, https://digitalpeace.microsoft.com/; and Paris Peace Forum, held November 2018, https://parispeaceforum.org/.

[8] Nakasone. 2019. Bolton. 2018.

[9] Perlroth, Nicole, and Quentin Hardy. 2013. "Bank Hacking Was the Work of Iranians, Officials Say." The New York Times, January 8. https://www.nytimes.com/2013/01/09/technology/online-banking-attacks-were-work-of-iran-us-officials-say.html.

[10] Defense Science Board, Department of Defense. 2017. "Task Force on Cyber Deterrence." Defense Science Board, 3, 4. https://www.acq.osd.mil/dsb/reports/2010s/DSB-cyberDeterrenceReport_02-28-17_Final.pdf.

the effectiveness of attacks, most notably by improving cyber defenses, network protection and security, and resilience) and deterrence by cost imposition (actions that increase the costs of the adversary when attacking, such as public attribution and shaming, diplomatic actions, economic sanctions, and the risk of a cyber or physical counterattack).[11]

The U.S. Government has used these components of deterrence over the last several years, with various levels of success. One publicly available report notes that network defense alone: "will not be sufficient to deter determined and sophisticated state-sponsored adversaries" and "the United States will also undertake a new effort to increase deterrence of state actors through cost imposition and other measures".[12] The new policy accordingly joins typical defense actions (like information sharing), with specific deterrent actions and operations for persistent presence, not meant to deter future attacks but to disrupt those underway or expected.

The assessment of this newly forceful DoD policy has broadly split into two camps, which we dub "hawks" and "owls." The hawks accept that a more forceful U.S. response with offensive cyber operations will work the way Bolton predicts, imposing "negative feedback" leading to a reduction in transgressions by adversaries. The owls are more cautious, worried that offensive cyberattacks – even if justified – may instead create "positive feedback," inciting more attacks in return.

There is sparse evidence supporting either position and the debate on whether the new policy will garner negative or positive feedback will not be settled through discussion and opinion, no matter how many op-eds are written. Rather, analyzing the success or failure of the new policy requires an evidence-based approach with a repeatable and transparent framework.

3. MEASURING CYBER DETERRENCE AND PERSISTENT ENGAGEMENT … INDIRECTLY

There have been few, if any, significant efforts to comprehensively measure these effects of persistent engagement or cyber deterrence on adversary behavior.

Scholars and practitioners have perhaps been dissuaded because the discussion for the past decade has been focused on cyber deterrence and it is "virtually impossible

[11] Note that cyber deterrence in this context only applies to a limited subset of adversaries: those tied to nation states, especially Russia, China, Iran, and North Korea (and any proxy or irregular groups supported by states). It does not apply to non-state groups like criminals or hacktivists.

[12] Department of State. 2018. "Recommendations to the President on Deterring Adversaries and Better Protecting the American People from Cyber Threats." May 31. https://www.state.gov/documents/organization/282253.pdf.

to know if deterrence is working" in cyberspace.[13] When the discussion is framed as deterrence, Henry Kissinger stated what is still a common position: "Since deterrence can only be tested negatively, by events that do not take place, and since it is never possible to demonstrate why something has not occurred, it became especially difficult to assess whether the existing policy was the best possible policy or just a barely effective one."[14] Yet, this is mostly true only in measuring the success of deterrence. Its failure would have been obvious soon after detonation of the first atomic warheads in an enemy's heartland; but because nuclear war never happened, this was not a practicable distinction.

The upside for cyber conflict is that – unlike with nuclear weapons – engagements and campaigns are constantly happening in cyberspace: what DoD is now calling "persistent engagement". These operations are tracked over time – though not to determine the success of different policies, such as persistent engagement and deterrence. The downside is that much of the activity of engagement and campaigns is hidden, cause and effect blend and overlap, and the identity of the adversary is obscured.[15]

Past cyber incidents show a range of "knowability" of the impact of adversary actions. On the more knowable end of that range, in responding to election interference in 2016, the administration of President Barack Obama took response actions off the table out of concern that Russia would escalate "against America's critical infrastructure—and possibly shut down the electrical grid" or engage in "hacking into Election Day vote tabulations."[16] This is knowable because principals involved in the Situation Room themselves confirmed the impact of Russian capabilities.

The second case, the 2015 agreement by President Barack Obama and President Xi Jinping of China, is not as clear; but, as will be discussed below, the debate can be addressed with data and an analytical framework. It is accordingly far more tractable than determining any effects of President Obama's warning to President Vladimir Putin over election interference. After the warning, in Hangzhou, China in September 2016, the U.S. government detected "no further evidence of Russia cyber-intrusions

[13] Sulmeyer, Michael. 2018. "How the U.S. can Play Cyber-Offense." Foreign Affairs, March 22. https://www.foreignaffairs.com/articles/world/2018-03-22/how-us-can-play-cyber-offense.

[14] Kissinger, Henry. 1994. Diplomacy. Simon and Schuster. p608.

[15] In the Cold War, academics and policymakers generally knew far more about U.S. operations and capabilities than those of the Soviets. In cyber conflict, the reverse is true: the DoD and Intelligence Communities publicly discuss adversary operations against the United States, while highly classifying their own operations against the same adversaries, masking critical issues such as determining of cause and effect.

[16] Healey, Jason. 2018. "Not the Cyber Deterrence the United States Wants." Council on Foreign Relations, June 11. https://www.cfr.org/blog/not-cyber-deterrence-united-states-wants. Subsequently confirmed in conversation with Gen. Clapper, former Director of National Intelligence; and Peter Clement, Columbia University, 21 February 2019.

into state election systems."[17] Such systems are typically not as tightly monitored as corporate networks, so there may be little data from which to draw conclusions. Barring exquisite intelligence or confirmation by the Kremlin, determining whether the warning caused any decrease is highly problematic.

Despite these drawbacks, this newly muscular U.S. strategy, like all policies, needs to be measured as best as possible to determine its effectiveness. Although they may not be definitive, rough-and-ready measurements of the scope and number of cyber incidents can suggest the impact of persistent engagement and deterrence. There has been some work in this space, especially by Brandon Valeriano and Ryan Maness, but it has been focused on academic questions of deterrence and not yet on policy effectiveness.[18]

The simplest metric framework is to describe different levels and types of cyber transgressions and simply tot them up. The next paragraphs describe such examples; these can be used as mere illustrations of the concept, especially to highlight that big data sets are not required (and may just overcomplicate the analysis, hiding more obvious signals), but can also be reasonable frameworks in their own right. Each metric should be tied directly to the goals of the policymakers.

A. Three Basic Frameworks

The Federal Government uses a standard five-tier severity score, the Cyber Incident Severity Schema, to assess the gravity of incidents.[19] The Schema rates cyber incidents according to observed effect, impact, affected sectors, and attribution (if known). Users can make qualitative judgments on these categories or attempt to score and weight them. If the general policy goal is to impose costs on adversaries and reduce the number and scope of significant incidents, this could be operationalized by tracking the number of attacks rated level 3 and above (those that are "significant cyber incidents" and likely to result in impacts to "public health or safety, national security, economic security, foreign relations, civil liberties, or public confidence"). The metric is then a simple algorithm along the lines of, "*if* cyber_level = {3, 4, 5} then count = count +1," tracked over time.[20] This metric might also be tied to the higher threshold of "use of force" to fit more neatly with the stated goal of the DoD

[17] Isikoff, Michael, and David Corn. 2018. "'Stand Down': How the Obama Team Blew the Response to Russian Meddling." Huffington Post, March 9. https://www.huffingtonpost.com/entry/stand-down-how-the-obama-team-blew-the-response-to-russian-meddling_us_5aa29a97e4b086698a9d1112.

[18] Valeriano, Brandon, and Ryan C Maness. 2015. Cyber War versus Cyber Realities. Cyber Conflict in the International System. Oxford University Press.

[19] U.S.-CERT, Department of Homeland Security. n.d. "NCCIC Cyber Incident Scoring System." https://www.us-cert.gov/NCCIC-Cyber-Incident-Scoring-System. Brandon Valeriano has pointed out that in his experience, a 10-point scale would allow finer grained analyses; but for now, DHS uses only five points.

[20] The authors intend to use the scheme to assess a number of past examples, possibly including Target, Iranian DDoS of the finance sector, Shamoon, Sony, OPM, Ukrainian power outage/Black Energy, Russian election interference, and Cloud Hopper.

strategy: to prioritize "deterring malicious cyber activities that constitute a use of force against the United States, our allies, or our partners."[21]

As attribution becomes clearer for each incident, this metric becomes more useful, as there can be a separate count for each of the United States' major adversaries in cyberspace, especially China, Russia, Iran, and North Korea. As the U.S. Government already assesses this score for all incidents to which they respond, there is no additional cost to tracking the trends over time to determine if the new policy has measurable impact. It is most useful for tracking discrete events, such as denial of service attacks or malware outbreaks (like NotPetya) and individual espionage incidents (OPM) than for less easily counted espionage campaigns (Cloud Hopper) or implanting malicious software for future use (Havex/Black Energy).[22]

This may be sound overly simple, but based on our knowledge and interviews, such tracking and measurement is less routine than may be imagined. Even marginal gains can be immediately useful to policymakers.

Bolton explained that the 2015 intrusion by China into the Office of Personnel Management was just "the kind of threat to privacy from hostile foreign actors that we're determined to deter".[23] This policy goal can be operationalized to a rough-and-ready metric framework by developing three elements:

1. A general description of an "OPM-type" incident, such as by scope, duration, intensity or against international laws or norms, U.S. red lines, or explicit agreement between states.
2. A measured baseline of such incidents. These can largely be drawn from headlines as they are both relatively few in number and quickly become public knowledge.
3. Tracking new developments to see whether the number of "OPM-type" incidents increases or decreases after the new policy comes into force.

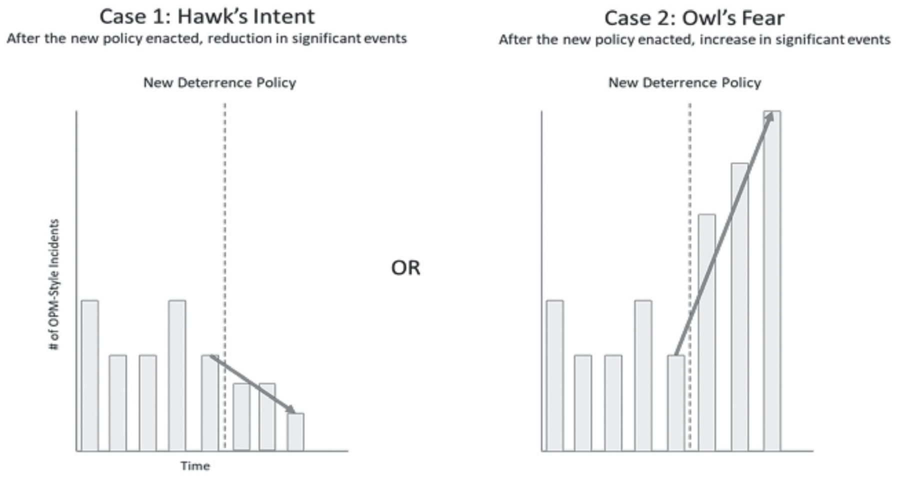
Although this simple metric could not meaningfully "prove" whether or not Bolton's threatened deterrence worked, if there was a decrease in such incidents (see case 1), then the evidence might indeed support Bolton's policy. But if there is a sharp increase in OPM-style incidents (see case 2), this suggests that the policy might be counter-productive. Further analysis is needed to check the competing hypotheses. The increase in case 2 is particularly significant, as the hawks suggest that the new U.S. policies

[21] Defense Science Board. 2017.

[22] NotPetya was the devastating 2017 ransomware outbreak caused by Russia, while OPM refers to the espionage incidents involving the U.S. Office of Personnel Management in 2015. Cloud Hopper was a large-scale Chinese espionage operation against managed service providers. Havex/Black Energy were malware implanted widely in energy grids but not, except for the notable exception of Ukraine, actually used to cause disruption.

[23] Bolton. 2018.

and actions should have a substantial impact on adversary operations. Therefore, any movement of the trend in the opposite direction has far more significance. Failure is louder than success.



Case 1: Hawk's Intent

After the new policy enacted, reduction in significant events

Case 2: Owl's Fear

After the new policy enacted, increase in significant events

The above two frameworks rely on data more than context. A more applied framework centers on measuring the effects of U.S. policies and actions meant to deter or dissuade a specific adversary from a specific kind of transgression. This more detailed metric is particularly useful for campaigns and implants, as it tracks, rather than individual incidents, volumes of activity over time from a specific adversary. The best example is the FireEye assessment that there had been a "notable decline in China-based groups' overall intrusion activity against entities in the U.S. and 25 other countries", with an especially sharp decline after the Obama-Xi agreement in 2015.[24] That company's cyber threat intelligence analysts measured the number of "active networks compromised" by 72 suspected China-based groups (see Chart x below).

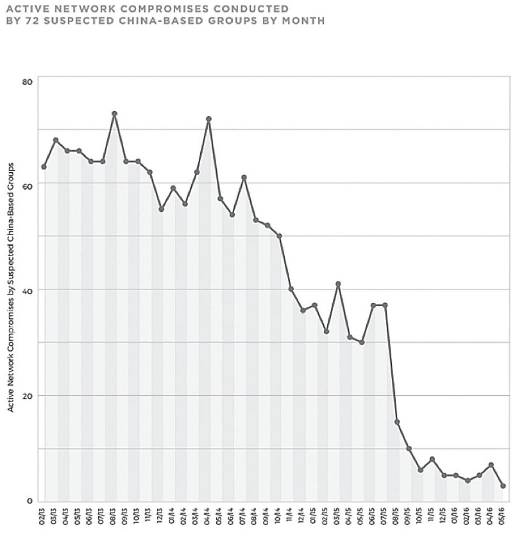
The 90+ percent decrease, according to FireEye, was due to "ongoing political and military reforms in China, widespread exposure of Chinese cyber activity, and unprecedented action by the U.S. government." It is noteworthy that two of these three reasons are related to U.S. counters: the public naming-and-shaming of "widespread exposure"; and "unprecedented" indictments and the threat of sanctions.

[24] FireEye iSight Intelligence. 2016. "Redline Drawn: China Recalculated Its Use of Cyber Espionage." FireEye. https://www.fireeye.com/content/dam/fireeye-www/current-threats/pdfs/rpt-china-espionage.pdf.

There is still debate about three key issues:

1. Was there was an actual decrease in Chinese espionage operations for commercial purposes? Perhaps the number of incidents held steady but the bulk were not detected, due to improved Chinese stealth. This is generally a question for cyber threat analysts.
2. How much of any Chinese response was the result of the U.S. policy? Perhaps the Chinese primarily acted for their own reasons, in response to domestic Chinese pressures, and U.S. policies had little additional impact. This is a question best answered by China experts.
3. Did the decrease matter? Perhaps the few networks still being compromised were those most critical to national security, so the overall impact was not meaningfully diminished. This is a question best answered by the policymakers themselves.

It has been over three years since the Obama-Xi agreement, yet there has been little if any structured work that has pulled out and analyzed these separate strands, rather than addressing whichever one supports the authors' preconceived ideas about China or the efficacy of agreements.



The evidence to answer these three questions, if not definitive, is certainly suggestive. We know many of our colleagues will disagree, some vehemently, with these assessments. This only reinforces the key point: that estimative conclusions must be systematically addressed in a transparent framework and tied to policymakers' goals. Isolating each element allows more transparency and repeatability, so that different analysts with different sources of information can develop individual and collective assessments of whether U.S. policies and actions are working or not.

Assistant Attorney General John Carlin confirmed FireEye's assessment of the direction and magnitude of Chinese activity: "Consistent with their agreement, they largely ceased state-sponsored hacking that targeted a private US company for the direct economic benefit of a Chinese competitor."[25] Even as late as November 2018, Rob Joyce, the former White House cybersecurity coordinator and NSA executive, with access to unique sources of intelligence, felt that although Chinese activity had returned, it had still dropped "dramatically" since the agreement of three years before.[26] We assess with medium confidence that it is very likely that there was a significant drop in Chinese activity.

On whether U.S. pressure worked, Xi has indeed been centralizing power in the Communist Party and his own person while cracking down on corruption. Either or both drives may have led Xi to clamp down on barely authorized cyber operations for commercial purposes. Even so, the Chinese, Carlin felt, "saw they had a big potential embarrassment brewing", while another Justice official noted "they were highly motivated to do the right thing".[27] According to Michael Daniel, then the White House cyber coordinator, the agreement was due to "steady, sustained pressure through a number of channels, including direct diplomacy, indirect diplomatic activity, public statements, and law enforcement actions" and "the Chinese were also concerned about potential additional actions that the U.S. could have taken, such as economic sanctions".[28] In addition, "President Xi had an upcoming visit to the United States and the Chinese wanted to make cybersecurity a positive topic, rather than a source of tension during the visit". Yet, with high confidence, we analyze it as very likely that U.S. pressure helped push Xi's decision – and subsequent Chinese action – in the preferred direction.

Regarding whether the decrease mattered, much of the original U.S. complaint was not only that the Chinese were stealing secrets for commercial purposes, but that the sheer quantity of such transgressions themselves was destabilizing. This is not an intelligence assessment but a policy judgment; but we believe this was a win for the United States. A reduction in the number of incidents directly relates to a decrease in perceived aggression: fewer incidents affected fewer companies. This is a strong benefit, in line with the U.S. policy goals, even if there was an impact from the remaining incidents. For other kinds of transgressions (see below), the policy goal must be not just fewer incidents, but zero, with any adversary action unacceptable. Espionage for commercial purposes is not usually such a zero tolerance issue.

[25] Graff, Garrett M. 2018. "How the US Forced China to Quit Stealing—Using a Chinese Spy." Wired, October 11. https://www.wired.com/story/us-china-cybertheft-su-bin/.

[26] Reuters. 2018. "US Accuses China of Violating Bilateral Anti-Hacking Agreement." CNBC, November 8. https://www.cnbc.com/2018/11/09/us-accuses-china-of-violating-bilateral-anti-hacking-agreement.html.

[27] Graff. 2018.

[28] Daniel, Michael. 2019. Interview with Michael Daniel (January 8).

B. The Model Applied to Other Transgressions

As illustrated in the China example above, measurement frameworks work best when the policy goals are clearly stated. As cyber incidents can take so many forms, this next section will articulate different kinds of transgressions, both to simplify the lexicon for policymakers and to define different categories for measurement. Analysts can, as above, rate the severity of transgressions and assign these to one or more categories. These are samples: there may be a larger set, especially as technology and adversary attacks develop over the decades.

Reckless Incidents: Some cyber transgressions have shown a "lack of regard for the danger or consequences," falling well outside the norms and having global effects.[29] These include attacks that have cascading or systemic effects, which cause significant cyber effects well beyond the intended target or original goal; or attacks that largely only affect their intended target, but that target itself is particularly critical or with a high potential for mistake or miscalculation and possibility of massive damage. As the true intent of the attacker may not be known, this will often be an analytical judgement based on effects and impact.

Coding, in the social sciences, means to apply categories to facilitate analysis. To determine if an incident should be coded as "reckless," how widespread the disruptive effects were (such as local to intended target, regional outside of intended target, or global), or the sensitivity or criticality of the intended target, might be assessed. For example, the NotPetya and WannaCry attacks, from Russia and North Korea respectively, both had globally disruptive impact. The Chinese "Great Cannon" denial of service against Github affected not only that software repository site, but developers globally who depended upon it.[30] All might be coded as "reckless."

Brazen Incidents: As seen in Bolton's response to the OPM espionage incident, some cyber incidents have a scope, duration and intensity that necessitates a significant national security response by the attacked nation: "This must not stand".[31] Some of these brazen attacks may cross a specific threshold, such as causing death or physical destruction or defying international law and norms. But "brazen" is not a legal threshold but a political judgment, as even espionage could be brazen if of an appropriate scope, duration, or intensity. As with "reckless," the intent of adversaries cannot be known and what might seem "brazen" to the defender might seem reasonable (or even just deserts) to others.

Possible coding for brazen transgressions includes the number of deaths; a measure of disruption or destruction (such as economic cost or number of systems "bricked");

[29] Oxford. 2009. "Recklessness." In Oxford English Dictionary. Oxford University Press.

[30] Marczak, Bill, Nicholas Weaver, Jakub Dalek, Roya Ensafi, David Fifield, Sarah McKune, Arn Rey, John Scott-Railton, Ron Deibert, and Vern Paxson. 2015. "China's Great Cannon." The Citizen Lab. April 10. https://citizenlab.ca/2015/04/chinas-great-cannon/.

[31] Thanks to Christopher Painter for the recommendation to tie "brazen" to incidents that necessitate a response.

and whether the incident violated a norm previously agreed to by the attacker, a global norm, a national “red line,” or none of these. The Chinese intrusion into OPM has been mentioned above as a possible brazen attack, while Russian interference in the 2016 U.S. elections is an even more obvious candidate.[32] The U.S. has conducted its own brazen attacks, most notably the Stuxnet malware attack (conducted with Israel) against Iranian uranium enrichment.[33]

Destabilizing Presence: Some systems are so critical and hazardous that *any* foreign cyber presence is extraordinarily high-risk and potentially destabilizing. For example, gaining access to the command and control systems of a nation’s nuclear weapons could precipitate a nuclear war. Access to the control systems of a nuclear power plant or a massive dam could be similarly high-risk, as even a simple key stroke error could cause a disaster. To a lesser degree, gaining access and pre-positioning malware in another nation’s electrical grid could be destabilizing because it could lead to a sudden, strategic strike.[34]

Possible coding for this category is far simpler, as it depends on the degree of adversary presence: zero, limited, or widespread. Perhaps the most worrying example is the Black Energy and Havex malware implanted by Russia in U.S. and European electrical systems, including nuclear power plants.[35] A variant was subsequently used to disrupt the Ukrainian power grid, in what was certainly also a brazen incident, highlighting that these categories are not mutually exclusive.[36]

Disproportionate Response: Nations are frequently subjected to intrusions and low-level disruption from other states. Another kind of transgression, related to those above, is when a nation’s response is far out of scale to the harm done to it. This is likely a small category, but is included here as the policy response to a disproportionate response should be different from the response to a pure brazen attack.

Possible coding for disproportionate response could include comparing the level of the initial incident (such as number of systems disrupted) with the response. For example, it is possible that the Iranians conducted the large-scale Shamoon attack

[32] Office of the Director of National Intelligence. 2017. “Background to ‘Assessing Russian Activities and Intentions in Recent US Elections’: The Analytic Process and Cyber Incident Attribution. https://www.dni.gov/files/documents/ICA_2017_01.pdf.; Sanger, David E. 2018. The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age. Crown. pxviii.

[33] Zetter, Kim. 2014. “An Unprecedented Look at Stuxnet, the World’s First Digital Weapon.” Wired, November 3. https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/.

[34] Clarke, Richard A, and Robert K Knake. 2011. Cyber War: The Next Threat to National Security and What to Do About It. Ecco. p244.

[35] F-Secure Labs. 2014. “BlackEnergy & Quedagh: The Convergence of Crimeware and APT Attacks.” https://www.f-secure.com/documents/996508/1030745/blackenergy_whitepaper.pdf. Constantine, Lucian. 2014. “New Havex Malware Variants Target Industrial Control System and SCADA Users.” PC World, June 24. www.pcworld.com/article/2367240/new-havex-malware-variants-target-industrial-control-system-and-scada-users.html.

[36] Jackson Higgins, Kelly. 2016. “Lessons from the Ukraine Electric Grid Hack.” Dark Reading, March 18. https://www.darkreading.com/vulnerabilities---threats/lessons-from-the-ukraine-electric-grid-hack/d/d-id/1324743.

against Saudi Aramco and Qatari Rasgas because their own energy infrastructure had been hit by a similar Wiper worm only weeks beforehand.[37] It seems likely to have been a disproportionate response rather than a fresh transgression. Distinguishing tit from tat is an important analytical distinction.

Attacker Infrastructure: This category stands apart, as it does not capture the output metrics of actual transgressions, but the impact of U.S. operations on adversary attack infrastructure – hop points, command and control servers, development and test environments, capabilities and the like. One hope for the new U.S. cyber doctrine is for U.S. operations to have a "strategic effect as the 'tactical friction' the adversary experiences through continuous engagement by the United States compels them to shift their resources (and thinking) toward their own vulnerabilities and defense."[38] This implies, in part, operations against adversary attack infrastructure to impose that friction, which can be directly measured. It would not be surprising if U.S. Cyber Command were using such measures to assess the effectiveness of their actions, but this can also be tracked by commercial cybersecurity companies.

Mandiant, in its groundbreaking report on the APT1 group, noted that the Chinese espionage team had "937 Command and Control (C2) servers hosted on 849 distinct IP addresses in 13 countries," and were "logging into their attack infrastructure from 832 different" Internet addresses.[39] Tracking these same metrics over time allows a rough measure of U.S. operational effectiveness. This could include total infrastructure disrupted, by category and by ratio of the total known infrastructure, and the mean time to rebuild.

C. Addressing the Shortcomings of This Approach

It can be relatively straightforward to use this framework as a rough-and-ready measure of the effects of U.S. policies and actions. This requires the three steps mentioned in the OPM example above: a description of the transgression (such as brazen, reckless), followed by a coding of past incidents fitting the description to create a baseline, followed by the addition of new incidents. Deep dives to analyze data for specific transgressions by specific adversaries helps to provide critical context and to differentiate between competing hypotheses. The overall results can be compared to deterrent policies and actions (as well as other, non-cyber developments between the nations) to see any suggestions of correlation.

Still, the interaction of international affairs and cyberspace is hidden, complex and

[37] Perlroth, Nicole. 2010. "In Cyberattack on Saudi Firm, U.S. Sees Iran Firing Back." The New York Times, October 24. https://www.nytimes.com/2012/10/24/business/global/cyberattack-on-saudi-oil-firm-disquiets-us.html.

[38] Harknett, Richard J. 2018. "United States Cyber Command's New Vision: What It Entails and Why It Matters." Lawfare, March 23. https://www.lawfareblog.com/united-states-cyber-commands-new-vision-what-it-entails-and-why-it-matters.

[39] Mandiant. 2013. "APT1: Exposing One of China's Cyber Espionage Units." p4. https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf.

ever-changing, challenging the disentanglement of multiple causes and effects. Any methodology to measure policy impacts – not just the one presented here – will share the following shortcomings, each of which can be effectively minimized.

The most obvious shortcoming is that the effect of U.S. actions may be swamped by technical developments. An increase in the number of reported incidents could be due to new classes of vulnerabilities, a flood of new and insecure Internet-of-things devices, or improvements in detection and defense. The deployment of more secure infrastructure would lead to fewer attacks, as would an increase in adversary use of "living off the land" and obfuscation techniques. This class of shortcoming can be controlled for by assessments and metrics from cyber threat analysts directly tracking adversary operations (such as those following Chinese espionage before and after the Obama-Xi agreement). Any variances can be investigated by comparing against the trend lines of different adversaries (if, say, the Iran trend declines but the trend for China increases). Competing hypotheses ("this increase means little because of more deployment of insecure IoT devices") can be compared against actual observed adversary behavior.

A second set of shortcomings include that many attacks (and adversary motivations) are hidden and data can be hard to come by and analyze. Geopolitical events could cause adversaries to decrease or increase their use of cyber capabilities for strategic ends, regardless of U.S. counter-offensive operations. Fortunately, having an exact enumeration of the events in each category matters less than the direction and magnitude of the trends.

The advocates of persistent engagement and deterrence suggest it should have a substantial, perhaps unprecedented impact on adversary behavior. Anything other than a correspondingly strong reduction, such as that seen after the Obama-Xi agreement, suggests that the policy may not be working as intended. If the trend significantly worsens, it may be that a hypothesis that the new policy is inciting adversaries is a better fit to the curve. But it could also mean that any deterrent effect is being swamped by other signals, perhaps an overall rise in global incidents or a significant worsening of tensions with the adversary nation. Either of these can be checked against a control, such as the overall trend of global incidents and bilateral relations (such as US-China). Other controls can include target states (if the United States sees a decrease of brazen attacks from China while the United Kingdom and France see increases).

These shortcomings can also be addressed by more sharing of intelligence, assessments, and data sets. Different communities have different strengths. Academic researchers generally can only rely on open-source material, especially media reporting, but bring rigor and strict methodologies; while commercial cyber threat analysts have long continuity following targets and have deep access to proprietary data (as the FireEye team did for the report on Chinese commercial espionage). U.S. government analysts, especially those in the Intelligence Community, can rely on classified sources but will miss much of the data held by commercial threat analysts (or by states, cities, and counties) and can overlook information not coming from classified sources.

A third set of shortcomings deal with methodological factors. The timescale to discover cyber incidents hampers assessment, as incidents are often not publicized until well after they are conducted, complicating efforts to ascribe cause and effect.[40] There may be so few truly dangerous attacks on a regular basis that an increase or decrease of a small number of incidents leads to an enormous percentage increase or decrease. These can be dealt with through appropriate structuring of the framework and coding of the data.

4. CONCLUSION AND FUTURE WORK

According to Michael Daniel, former White House cyber coordinator, the Trump administration "is willing to take more risks than previous administrations, but the proof will be in the results".[41] We can't assess what we don't try to measure. Together, the frameworks in this paper can act as a check on whether these new, riskier U.S. cyber policies and operations are succeeding in suppressing incoming attacks, or inciting them.

The shortcomings in the previous section are generally not specific to this paper and would pertain to *any* attempt at measuring the new U.S. policies. Some of the people reviewing this paper suggested that the U.S. Government – especially the intelligence community – would be uniquely placed to conduct these assessments. But the Federal Government cannot easily measure attacks from adversaries, as it lacks access to most victim data, which can be held by cybersecurity companies and organizations like the Cyber Threat Alliance. Moreover, the U.S. Government cannot easily even know all its own operations against adversaries: some will be covert actions, others espionage, while others are "traditional military operations." Each is held in a separate compartment and few individuals have the full picture.

Of course, we still encourage all parties to attempt to measure. The U.S. Government should conduct its assessment, with different agencies using their own processes,

[40] This delay can be seen in large data breaches, such as the Marriott incident that occurred around the same time as the OPM incident in 2014 but was only publicized in 2018; as well as offensive cyber effects operations, such as U.S. Cyber Command reportedly blocking internet access to the Internet Research Agency on the day of the 2018 elections, which went unreported until February 2019.

[41] Nakashima, Ellen. 2018. "White House authorizes 'offensive cyber operations' to deter foreign adversaries." The Washington Post, September 20. https://www.washingtonpost.com/world/national-security/trump-authorizes-offensive-cyber-operations-to-deter-foreign-adversaries-bolton-says/2018/09/20/b5880578-bd0b-11e8-b7d2-0773aa1e33da_story.html?utm_term=.f0d9d4720f36.

sources, and methods. The National Intelligence Council or Cyber Threat Intelligence Integration Center may be natural homes for much of this activity. As no one should be allowed to grade their own homework, this process should not be owned by either the National Security Agency or U.S. Cyber Command. The National Security Council must be the ultimate arbiter, deciding if the new operations are meeting the goals set by policymakers.

Rough-and-ready metrics such as those presented here can at least begin to indicate the direction and magnitude of changes over time, allowing indirect measurement to determine whether the policies are suppressing adversary attacks or inciting them. As this project moves forward, we will seek to improve and further refine the framework presented here for a usable pilot project for the commercial cyber threat intelligence community. These companies regularly assess the impact and quantity of foreign cyber operations; with a more standard and transparent methodology, they can help create a public understanding of the impact of U.S. actions on cyberspace, which has taken a central position in supporting our economy and society.

Additional research should also be done on historical antecedents of persistent engagement. Though the comparisons are inexact, persistent engagement has similarities to other examples where the military and intelligence forces of the two blocs during the Cold War were in routine belligerent contact: anti-submarine warfare; espionage-counterespionage; freedom of navigation operations; and intelligence, surveillance, and "exciter" flights against each other's homelands.

Acknowledgments

The authors wish to acknowledge Jennifer Gennaro of Columbia University's School of International and Public Affairs for her assistance. We would also like to thank the staff and attendees of the inaugural CyberWarCon conference (held on 28 November 2016 in Arlington, Virginia), where we presented an early version of this work; and the participants of a workshop at SIPA on 25 February 2019. Michael Daniel, Gregory Rattray, Chris Painter, and Alexandra Friedman also provided valuable support and input. This work was funded in part by the Office of Naval Research under the OSD Minerva program: Grant number N00014-17-1-2423

REFERENCES

Barnes, Julian E. 2018. "U.S. Begins First Cyberoperation Against Russia Aimed at Protecting Elections." *The New York Times*, October 23. https://www.nytimes.com/2018/10/23/us/politics/russian-hacking-usa-cyber-command.html.

Bolton, John. 2018. “Transcript: White House Press Briefing on National Cyber Strategy - Sept. 20, 2018”. Washington DC (September 8). Available at https://news.grabien.com/making-transcript-white-house-press-briefing-national-cyber-strateg.

Clarke, Richard A, and Robert K Knake. 2011. *Cyber War: The Next Threat to National Security and What to Do About It*. Ecco.

Constantine, Lucian. 2014. “New Havex Malware Variants Target Industrial Control System and SCADA Users.” *PC World*, June 24. www.pcworld.com/article/2367240/new-havex-malware-variants-target-industrial-control-system-and-scada-users.html.

Daniel, Michael. 2019. Interview with Michael Daniel (January 8).

Defense Science Board, Department of Defense. 2017. “Task Force on Cyber Deterrence.” Defense Science Board, 3, 4. https://www.acq.osd.mil/dsb/reports/2010s/DSB-cyberDeterrenceReport_02-28-17_Final.pdf.

Department of Defense. 2018. “Cyber Strategy 2018.” https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF.

Department of State. 2018. “Recommendations to the President on Deterring Adversaries and Better Protecting the American People from Cyber Threats.” May 31. https://www.state.gov/documents/organization/282253.pdf.

FireEye iSight Intelligence. 2016. “Redline Drawn: China Recalculated Its Use of Cyber Espionage.” FireEye. https://www.fireeye.com/content/dam/fireeye-www/current-threats/pdfs/rpt-china-espionage.pdf.

F-Secure Labs. 2014. “BlackEnergy & Quedagh: The Convergence of Crimeware and APT Attacks.” https://www.f-secure.com/documents/996508/1030745/blackenergy_whitepaper.pdf.

Geller, Eric. 2018. “Trump Scraps Obama Rules on Cyber Attacks, Giving Military Freer Hand.” *Politico*, August 18. https://www.politico.com/story/2018/08/16/trump-cybersecurity-cyberattack-hacking-military-742095.

Graff, Garrett M. 2018. “How the US Forced China to Quit Stealing—Using a Chinese Spy.” *Wired*, October 11. https://www.wired.com/story/us-china-cybertheft-su-bin/.

Harknett, Richard J. 2018. “United States Cyber Command’s New Vision: What It Entails and Why It Matters.” *Lawfare*, March 23. https://www.lawfareblog.com/united-states-cyber-commands-new-vision-what-it-entails-and-why-it-matters.

Healey, Jason. 2018. “Not The Cyber Deterrence the United States Wants.” *Council on Foreign Relations*, June 11. https://www.cfr.org/blog/not-cyber-deterrence-united-states-wants.

Isikoff, Michael, and David Corn. 2018. “‘Stand Down’: How The Obama Team Blew The Response To Russian Meddling.” *Huffington Post*, March 9. https://www.huffingtonpost.com/entry/stand-down-how-the-obama-team-blew-the-response-to-russian-meddling_us_5aa29a97e4b086698a9d1112.

Jackson Higgins, Kelly. 2016. “Lessons from the Ukraine Electric Grid Hack.” *Dark Reading*, March 18. https://www.darkreading.com/vulnerabilities---threats/lessons-from-the-ukraine-electric-grid-hack/d/d-id/1324743.

Kissinger, Henry. 1994. *Diplomacy*. Simon and Schuster.

Langevin, James R. 2019. “Opening Statement: FY 2020 Budget Request for Military Operations in Cyberspace.” March 13. https://armedservices.house.gov/_cache/files/d/5/d5f94725-3373-40ef-803c-1f0ff8f106a8/577D710BF48F37825B2656EE1AF6891A.opening-statement---ietc-chairman-langevin-3-13-2019.pdf.

Mandiant. 2014. “APT1: Exposing One of China’s Cyber Espionage Units.”

Marczak, Bill, Nicholas Weaver, Jakub Dalek, Roya Ensafi, David Fifield, Sarah McKune, Arn Rey, John Scott-Railton, Ron Deibert, and Vern Paxson. 2015. “China’s Great Cannon.” *The Citizen Lab*. April 10. https://citizenlab.ca/2015/04/chinas-great-cannon/.

Microsoft. 2018. Digital Peace. https://digitalpeace.microsoft.com/.

Nakashima, Ellen. 2014. “U.S. Rallied Multinational Response to 2012 Cyberattack on American Banks.” *The Washington Post*, April 11. https://www.washingtonpost.com/world/national-security/us-rallied-multi-nation-response-to-2012-cyberattack-on-american-banks/2014/04/11/7c1fbb12-b45c-11e3-8cb6-284052554d74_story.html?utm_term=.be386c400c97.

—. 2018. “White House authorizes ‘offensive cyber operations’ to deter foreign adversaries.” *The Washington Post*, September 20. https://www.washingtonpost.com/world/national-security/trump-authorizes-offensive-cyber-operations-to-deter-foreign-adversaries-bolton-says/2018/09/20/b5880578-bd0b-11e8-b7d2-0773aa1e33da_story.html?utm_term=.f0d9d4720f36.

Nakasone, Paul M. 2019. “An Interview with Paul M. Nakasone.” *Joint Forces Quarterly*. https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-92/jfq-92.pdf.

Office of the Director of National Intelligence. 2017. “Background to ‘Assessing Russian Activities and Intentions in Recent US Elections’: The Analytic Process and Cyber Incident Attribution.” https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Oxford. 2009. “Recklessness.” In *Oxford English Dictionary*. Oxford University Press.

Paris Peace Forum. n.d. Paris Peace Forum. https://parispeaceforum.org/.

Perlroth, Nicole. 2010. “In Cyberattack on Saudi Firm, U.S. Sees Iran Firing Back.” *The New York Times*, October 24. https://www.nytimes.com/2012/10/24/business/global/cyberattack-on-saudi-oil-firm-disquiets-us.html.

Perlroth, Nicole, and Quentin Hardy. 2013. “Bank Hacking Was the Work of Iranians, Officials Say.” *The New York Times*, January 8. https://www.nytimes.com/2013/01/09/technology/online-banking-attacks-were-work-of-iran-us-officials-say.html.

Reuters. 2018. “US Accuses China of Violating Bilateral Anti-Hacking Agreement.” CNBC, November 8. https://www.cnbc.com/2018/11/09/us-accuses-china-of-violating-bilateral-anti-hacking-agreement.html.

Sanger, David E. 2018. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. Crown.

Segal, Adam. 2016. “The U.S.-China Cyber Espionage Deal One Year Later.” *Council on Foreign Relations*, September 28. https://www.cfr.org/blog/us-china-cyber-espionage-deal-one-year-later.

Sulmeyer, Michael. 2018. “How the U.S. can Play Cyber-Offense.” *Foreign Affairs*, March 22. https://www.foreignaffairs.com/articles/world/2018-03-22/how-us-can-play-cyber-offense.

The White House. 2018. “National Cyber Strategy.” https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf.

U.S. Cyber Command. 2018. “Achieve and Maintain Cyberspace Superiority: Command VIsion for U.S. Cyber Command.” April. https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf?ver=2018-06-14-152556-010.

U.S.-CERT, Department of Homeland Security. n.d. “NCCIC Cyber Incident Scoring System.” https://www.us-cert.gov/NCCIC-Cyber-Incident-Scoring-System.

—. n.d. “US-CERT Federal Incident Notification Guidelines.” https://www.us-cert.gov/incident-notification-guidelines.

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

Zetter, Kim. 2014. “An Unprecedented Look at Stuxnet, the World’s First Digital Weapon.” *Wired*, November 3. https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

The All-Purpose Sword: North Korea's Cyber Operations and Strategies

Ji Young, Kong
ROK Air Force First Lieutenant
Department of Information Security
Korea University, Agency for Defense Development
Seoul, Republic of Korea
jykong27@gmail.com

Jong In, Lim
Professor
Department of Information Security
Korea University
Seoul, Republic of Korea
jilim76@gmail.com

Kyoung Gon, Kim
Industry-University Cooperation Professor
Department of Information Security
Korea University
Seoul, Republic of Korea
anesra@gmail.com

Abstract: According to a 2013 briefing from the South Korean National Assembly by the South Korean National Intelligence Service, North Korean leader Kim Jong-un stated, "Cyberwarfare is an all-purpose sword that guarantees the North Korean People's Armed Forces ruthless striking capability, along with nuclear weapons and missiles." Kim has secretly executed all-purpose cyberattacks to achieve his agenda, regardless of North Korea's diplomatic and economic situation. The "all-purpose sword" has been adapted to the different purposes it has pursued against North Korea's adversaries, such as creating ransomware for financial gain, a cyberweapon to destroy computer systems, and an invisible espionage tool to accumulate sensitive information. This paper is divided into three parts. The first section discusses the will of North Korea to use cyber warfare for different purposes by explaining how its administrative agencies take charge of different fields but carry out cyber operations to achieve their goals. The second section describes and analyzes the interconnectivity in North Korea's suspected cyber operations: specifically, Campaign Kimsuky, Operation KHNP, Operation DarkSeoul, Operation Blockbuster, the Bangladesh Central Bank Heist, and Wannacry. The operations will be categorized by operational goals, showing North Korea's success at achieving its various purposes by these means. In the last section, we suggest a future cyber strategy direction for North Korea based on our analysis of its tactics, techniques and procedures; and how North Korea cooperates with other countries, including countermeasures for countries around the world.

Keywords: *North Korea, North Korean cyber forces, state-sponsored cyber operations, mixing tactics, techniques and procedures, cyber strategies*

1. INTRODUCTION

Kim Jong-un's interest in cyber warfare predated the start of his regime. Kim Jong-il, the former North Korean leader, perceived the advantage of having a networked military after monitoring the 1991 Gulf War, the 1999 Kosovo War, and the 2003 Iraq War (Jun, LaFoy, and Sohn 2015). Subsequently, Kim Jong-il had stressed the importance of building cyber capabilities. In the *Electronic Warfare Reference Guide* published by the Korean People's Army's Military Publishing House in 2005, he stated, "If the Internet is like a gun, cyber-attacks are like atomic bombs"; and "modern war is decided by one's conduct of electronic warfare," thus "cyber units are my detached force and backup power." (Ahn 2011) Moreover, after the Iraq War, he convened a high-level meeting and asserted:

> If warfare was about bullets and oil until now, warfare in the 21st century is about information. War is won and lost by who has greater access to the adversary's military technical information in peacetime, how effectively one can disrupt the adversary's military command and control information, and how effectively one can utilize one's own information. (Kim 2010)

Based on his father's work, Kim Jong-un, the current leader of North Korea, established and extended specific mission-oriented cyber units. Having taken a degree majoring in computer science and the military, he emphasized the importance of cyber warfare. In February 2013, he visited the cyber units of the Reconnaissance General Bureau (RGB), and proclaimed that "With intensive information and communication technology, and the brave RGB with its [cyber] warriors, we can penetrate any sanctions for the construction of a strong and prosperous nation." (Lee 2013)

North Korea aims to develop its asymmetric military power by enlisting elite soldiers for their cyber capabilities while minimizing Internet dependency in the country. A North Korean defector who had been a high-ranking official testified that the country annually sends 50 to 60 elite soldiers abroad to study computer science, who later work as cyber attackers in the RGB and other cyber units (Han 2016). As a result, there are an estimated 6,800 trained cyberwarfare specialists in the North's cyber units (ROK Ministry of National Defense 2018). Meanwhile, the North chooses to protectively control the Internet rather than provide open information, since Kim believes, along with China, that open information would harm his regime. Accordingly, only a few people can access and use the Internet; additionally, the North developed its own intranet in 1996, which is separate from the Internet and only accessible within its territory (Lim, Kwon, Jang, and Baek 2013).

Therefore, cyber warfare is an optimal choice for North Korea considering its costs and effects. It ensures continuous effects during peacetime and wartime, with covert and low-cost cyber operations that achieve various missions from the upper leadership of North Korea, without leaving irrefutable physical traces as conventional military forces would.

Although public interest has increased, few studies have been conducted on North Korea's cyber capabilities and strategies due to information limitations. Lim, Kwon, Jang, and Baek (2013) analyzed the North's cyber capabilities and proposed 10 cyber strategies, based on technical, political, and international aspects, for South Korea to counteract North Korea. Jun, LaFoy, and Shon (2015) made policy recommendations to the U.S. and the U.S.–ROK alliance after analyzing the approach of North Korea's cyber operations, based on conventional military strategies, specific institutions within the government, and the technology and industrial base. However, these related works had limitations in considering North Korea's cyberwarfare as extended capabilities, i.e. taking North Korea's cooperation with third-party countries to conduct operations into account.

Ha and Maxwell (2018) suggested using case studies to explain North Korea's capabilities and its avoidance of sustained Cyber-Enabled Economic Warfare Operations because of its primary strategic objective of prolonging the Kim regime's survival and its desire to remain within the gray zone.[1] However, this suggestion has limited ability in emphasizing the importance of viewing cyber units under North Korea's military command structure. Accordingly, the authors emphasize the

[1] U.S. Special Operations Command defines the "gray zone" as a realm of competitive interactions among and within state and non-state actors that fall between the traditional war and peace duality. See U.S. Special Operations Command (2015).

significance of the Kim regime's use of cyber warfare, by evaluating North Korea's military command structure, cyber operations, and relations with other countries.

2. ORGANIZATIONS OF CYBER OPERATIONS IN NORTH KOREA

FIGURE 1. NORTH KOREA'S MILITARY COMMAND STRUCTURE (ROK MINISTRY OF NATIONAL DEFENSE 2018; JUN, LAFOY, AND SHON 2015; PARK 2018; MOK 2017B)
Remark: The dark-shaded units are directly relevant to cyber operations; the lighter-shaded units may have the potential to conduct cyber capabilities.

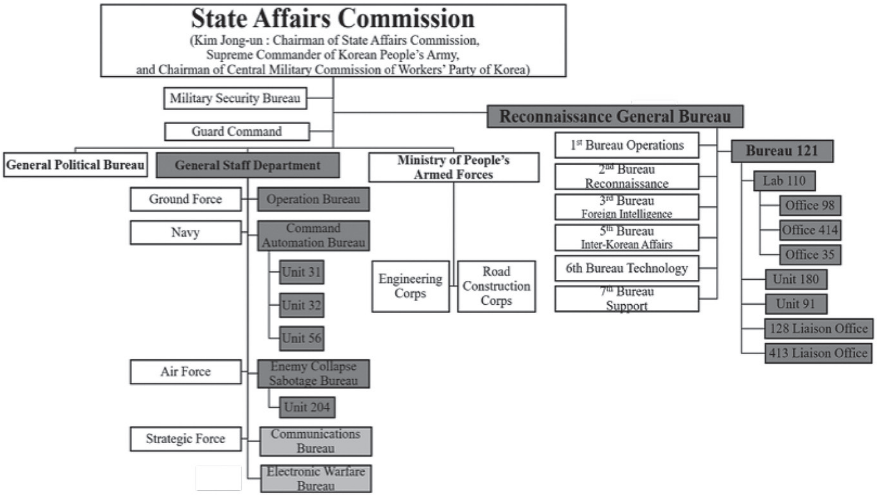


Figure 1 shows North Korea's military command structure based on the Defense White Paper of the ROK, published in December 2018. In addition, it illustrates other recent, open-source media regarding newly formed units that are relevant to cyber operations. Kim Jong-un, who serves as Chairman of the State Affairs Commission (SAC) (국무위원회 위원장), Supreme Commander of the Korean People's Army (KPA) (인민군 최고사령관), and Chairman of the Central Military Commission of the Workers' Party of Korea (WPK) (당 중앙군사위원회 위원장), maintains practical command and control over the North Korean military. As Chairman of the SAC, Kim oversees the affairs of the North Korean state and decides which policies are important to the country. As the Supreme Commander of the KPA, Kim commands the General Political Bureau (총정치국, GPB), General Staff Department (총참모부, GSD) and the Ministry of the People's Armed Forces (인민무력성, MPAF). Specifically, the

GPB oversees party organs within the military and is responsible for issues related to political ideology; while the GSD is responsible for conducting military operations; and the MPAF for administering military diplomacy, military logistics, procurement, and finance. Furthermore, as the Chairman of the Central Military Commission of the WPK, Kim deliberates and decides what measures are necessary for implementing military policy and provides guidance for overall defense affairs at a party level (ROK Ministry of National Defense 2018).

As shown in Figure 1, military units that carry out cyber operations in North Korea's military command structure are largely divided into two groups: the GSD of KPA; and the Reconnaissance General Bureau (정찰총국, RGB). This division illustrates the RGB's high degree of strategic importance. Since it is independent of the GSD and MPAF, this indicates that it acquires tasks and reports directly to the upper leadership of the SAC, Kim Jong-un, in both peacetime and wartime (Bechtol Jr. 2018).

Reconnaissance General Bureau: The RGB was formed in 2009; it is equivalent to the U.S. Directorate of National Intelligence (Madde 2018). The RGB reports directly to the SAC: it used to report directly to the senior leadership of the National Defense Commission, which was replaced by the SAC in the 2016 constitutional revision (Bechtol Jr. 2018). Since the SAC tasks the RGB with North Korea's terrorist, clandestine and illicit activities, and the RGB conducts these tasks independent of North Korea's conventional military (the KPA), previous studies have suggested that North Korea sees cyber capabilities as extending beyond military assets (Ha and Maxwell 2018, 11).

Under the RGB, Bureau 121 is the primary office tasked with disruptive cyber operations, such as infiltrating computer networks, hacking to extract foreign intelligence, and deploying viruses on adversary computer networks (Chung and Lee 2017, 21). According to the Radio Free Asia interview with Kim Heungkyang, the leader of North Korean Intellectuals Solidarity, his report. "The actual state of North Korea's Cyberwarfare Reinforcement and Counterstrategies for South Korea", submitted to the National Assembly Defense Committee of South Korea on September 30, 2016, introduced the newly formed organizations after reorganizing Bureau 121: specifically, Lab 110, Unit 180, Unit 91, 128 Liaison Office, and 413 Liaison Office (Mok 2017b).

Lab 110 is the key cyber unit under the RGB; it applies cyberattack techniques to conduct intelligence operations. The South Korean military discovered that Lab 110 was an expansion and reorganized adaptation of Unit 121 under the RGB, credited with researching computer command systems and electronic jamming in 1998 (Kim 2014). According to Park's presentation at DragonCon 2018, Lab 110 is divided into three offices according to their function. Office 98, located in Pyongyang, primarily collects information on North Korean defectors, organizations that support them, overseas research institutes related to North Korea, and university professors in South Korea. Office 414, located in Pyongyang and Shenyang, China, gathers information on overseas government agencies, public agencies, and private companies. Office 35 is in Pyongyang and concentrates on developing malware, researching and analyzing vulnerabilities, exploits, and hacking tools (Park 2018).

Unit 180 specializes in conducting cyber operations to steal foreign money from outside North Korea. Hackers in Unit 180 generally operate overseas to obscure the link between their operations and North Korea (Ha and Maxwell 2018); the state offers them every support in coming and going abroad to conduct their operations. Unit 91 focuses on cyberattack missions targeting isolated networks, particularly on South Korea's critical national infrastructure such as KHNP and the ROK Ministry of National Defense. Moreover, Unit 91 targets stealing confidential information and technology to develop weapons of mass destruction with a "super striking power," as ordered by Kim Jong-un (Mok 2017b).

The term "Liaison Office" usually denotes an office responsible for "escorting and communicating with any commando or special operations forces sent to infiltrate South Korea" in North Korea. 128 Liaison Office and 414 Liaison Office are likely responsible for maintaining communications with espionage networks in South Korea, including relaying missions and receiving reports rather than directly impacting targets using cyber capabilities. Specifically, 128 Liaison Office works on hacking foreign information intelligence websites and studies cyber strategies, while 414 Liaison Office cultivates cyber experts to conduct cyberwarfare (Jun, Lafoy, and Sohn 2015).

General Staff Department: The GSD is responsible for the operational command and operational planning of the Korean People's Army (Jun, LaFoy, and Sohn 2015). Its primary goal with cyber capabilities is to integrate emerging tools and weapons of cyber capabilities into North Korea's warfighting strategy (Park 2018). The Operations Bureau does not directly perform cyber operations but may serve an important role in making key decisions related to cyber force planning, defining and disseminating cyber strategy, and mission. The Command Automation Bureau is responsible for conducting cyberwarfare operations. Units 31, 32, and 56 are responsible for malware development, military software development, and command and control software development, respectively. The Enemy Collapse Sabotage Bureau is tasked with information and psychological warfare (Jun, LaFoy, and Sohn 2015).

3. CYBER OPERATIONS ATTRIBUTED TO NORTH KOREA

TABLE 1. CATEGORIES OF OPERATIONS BASED ON OBJECTIVES

Objectives	Cyber Operations	Period	Remarks
Information Espionage	Campaign Kimsuky	2009–2018	Variants and affiliations have been found up until 2018
	Operation KHNP	2014.12.15	Causing social chaos in South Korea
Cyber Terrorism[2]	Operation DarkSeoul	2013.3.20	The attackers shared TTPs[3] with malicious activities from 2007
	Operation BlockBuster	2014.11.24	*The Interview* movie that plotted the assassination of Kim Jong-un kindled the attack The FBI attributed this attack to North Korea
Financial Warfare	Bangladesh Central Bank Heist	2016.02.04–05	Stealing bank credentials and sending fraudulent transactions to SWIFT
	WannaCry	2017.05	Demanding a ransom for files taken hostage The FBI attributed this to North Korea

Table 1 shows the notable cyber operations attributed to North Korea, categorized into three objectives: Information Espionage, Cyber Terrorism and Financial Warfare. Although some analysts include psychological warfare to describe the objectives of cyberwarfare, this paper does not define the objectives as the goals of warfare, only as the primary goals of the operations. Accordingly, it does not include cyberattacks which aim for psychological warfare, based on the authors' determination that they did not reach the level of operations. Moreover, the objectives of operations may not be exclusive to each other, so an operation can fall in more than two categories. Therefore, this paper categorizes operations with only one primary goal for each operation.

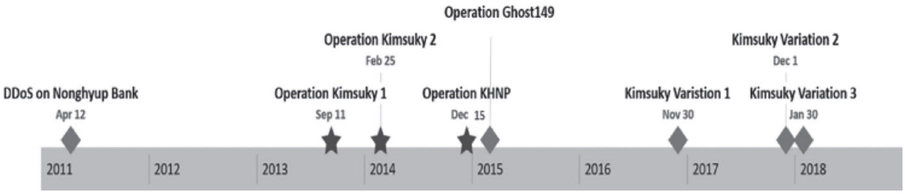
There are several recognizable state-sponsored actors in North Korea, such as: Lazarus, Bluenoroff, Hidden Cobra, Andariel, Bureau 121, APT37, ScarCruft, Reaper, Group123, DarkHotel, etc. These groups have been named by various security analysts to identify them as actors by the malware and tactics they used, which has resulted in multiple naming conventions used for specific actors. To avoid confusion arising from the various naming conventions, this paper has decided to identify the actors as a singular group executing orders from their country.

[2] NATO's definition of cyber terrorism is: "A cyberattack using or exploiting computer or communication networks to cause sufficient destruction or disruption to generate fear or to intimidate a society into an ideological goal." Center of Excellence, Defence Against Terror (2008, 119).

[3] TTP, in the context of cyber threat intelligence, is short for Tactics, Techniques, and Procedures, and also sometimes referred to as Tools, Techniques, Procedures. TTPs represent the behavior or *modus operandi* of cyber adversaries.

A. Information Espionage Operations

FIGURE 2. TIMELINE OF OPERATIONS FOR INFORMATION ESPIONAGE

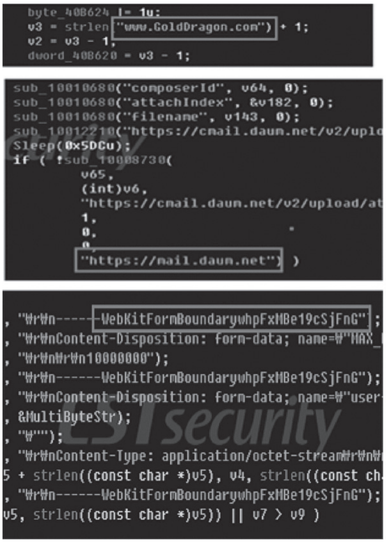


Campaign[4] Kimsuky: The first attack of Campaign Kimsuky started in Sep 11, 2013 when phishing emails that contained malicious files were sent through Belgian mail accounts. All the information collected by the malicious files was sent to two master mail accounts: iop110112@hotmail.com and rsh1213@hotmail.com, which were registered with the names "kimsukyang" and "kim asdfa," leading to the campaign being dubbed "kimsuky" (Tarakanov 2013). The second attack of Campaign Kimsuky started on Feb 25, 2014 and more attacks were conducted on March 11, 12, 17, and 19 (AhnLab 2014).

Operation KHNP: From Dec 15, 2014, anti-hacktivists attributed to North Korean hackers calling themselves "Who am I = No Nuclear Power" started releasing information about Korea Hydro and Nuclear Power (KHNP) employees and confidential technical documents on nuclear power plants after launching cyberattacks (Security News Special Coverage Team 2014). The South Korean government concluded that Operation KHNP was a hacking incident caused by North Korean hackers with the purpose of creating social unrest in South Korea by targeting the critical national infrastructure of nuclear power plants (Seoul Central District Public Prosecutors' Office 2015).

Links: There are ongoing espionage investigations into the affiliations of the Kimsuky malware. On Nov. 30, 2016, Dec. 1, 2017, and Jan 30, 2018, a substantial number of e-mails with malicious hwp files attached that contained variants of the Kimsuky malware were sent to specific universities and public organizations in South Korea. Some of the malicious files had an exact copy of the HwpSummaryInformation code in their shellcodes and the same creator account "MOFA," which stands for Ministry of Foreign Affairs, an acronym strongly associated with South Korea (Alyac 2018). This means that the same attackers used the same metadata for more than a year while conducting these attacks. There is another link indicating that the same group conducted the attacks in February 2015 and on January 30, 2018 by using similar names for its C&C Server and using the same HTTP parameter to communicate with the C&C Server (Gil 2018). Figures 3 and 4 show the similar C&C Server hostnames, “mail.daum.net” and “mail-daum-net.atwebpages.com,” respectively, and the same HTTP parameter “WebKitFormBoundarywhpFxMBe19cSjFnG.”

FIGURE 3. MALWARE IN THE 2015 APT ATTACK (ALYAC 2018)

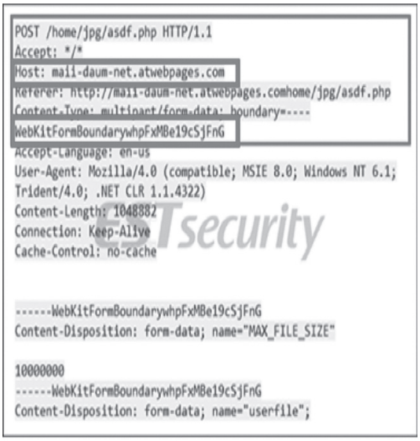


```
    byte_40B624 |= 1u;
    v3 = strlen("www.GoldDragon.com") + 1;
    v2 = v3 - 1;
    dword_40B620 = v3 - 1;
```

```
  sub_10010680("composerId", v64, 0);
  sub_10010680("attachIndex", &v182, 0);
  sub_10010680("filename", v143, 0);
  sub_10012210("https://cmail.daum.net/v2/uplo
  Sleep(0x5DCu);
  if ( !sub_10008730(
          v65,
          (int)v6,
          "https://cmail.daum.net/v2/upload/at
          1,
          0,
          0,
          "https://mail.daum.net") )
```

```
, "₩r₩n------WebKitFormBoundarywhpFxMBe19cSjFnG");
, "₩r₩nContent-Disposition: form-data; name=₩"MAX_
, "₩r₩n₩r₩n10000000");
, "₩r₩n------WebKitFormBoundarywhpFxMBe19cSjFnG");
, "₩r₩nContent-Disposition: form-data; name=₩"user
, &MultiByteStr);
, "₩"");
, "₩r₩nContent-Type: application/octet-stream₩r₩n₩
5 + strlen((const char *)v5), v4, strlen((const ch
, "₩r₩n------WebKitFormBoundarywhpFxMBe19cSjFnG");
v5, strlen((const char *)v5)) || v7 > v9 )
```

FIGURE 4. MALWARE IN THE 2018 APT ATTACK (GIL 2018)



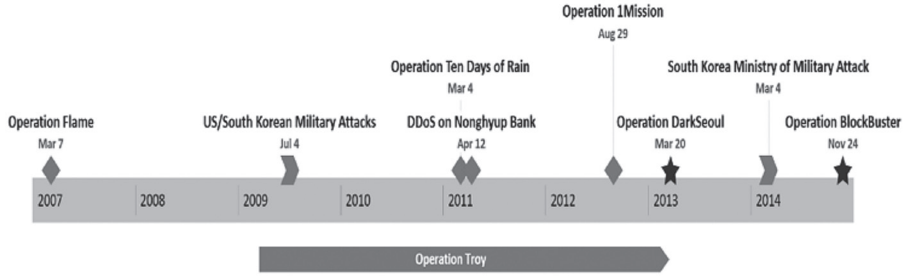
```
POST /home/jpg/asdf.php HTTP/1.1
Accept: */*
Host: mail-daum-net.atwebpages.com
Referer: http://mail-daum-net.atwebpages.comhome/jpg/asdf.php
Content-Type: multipart/form-data; boundary=----
WebKitFormBoundarywhpFxMBe19cSjFnG
Accept-Language: en-us
User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1;
Trident/4.0; .NET CLR 1.1.4322)
Content-Length: 1048882
Connection: Keep-Alive
Cache-Control: no-cache


------WebKitFormBoundarywhpFxMBe19cSjFnG
Content-Disposition: form-data; name="MAX_FILE_SIZE"

10000000
------WebKitFormBoundarywhpFxMBe19cSjFnG
Content-Disposition: form-data; name="userfile";
```

B. Cyber Terrorism Operations

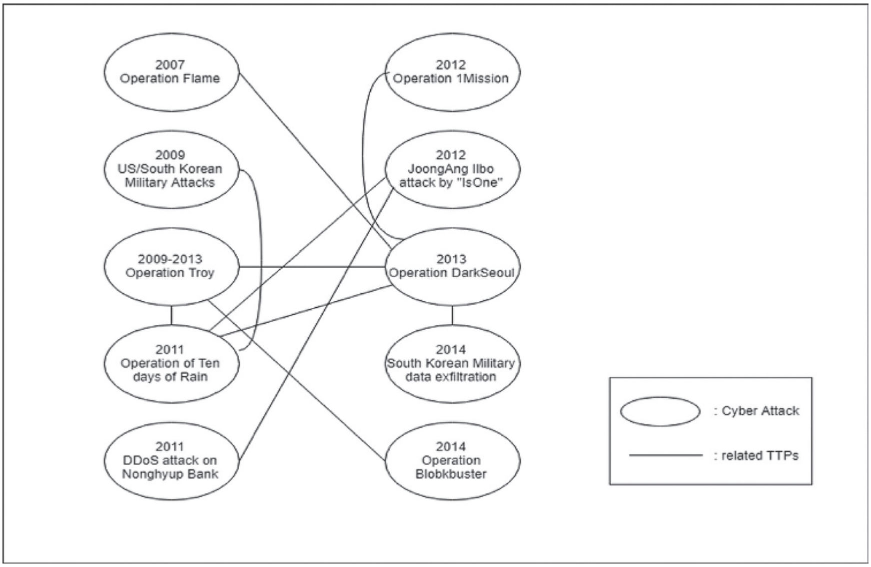
FIGURE 5. TIMELINE OF OPERATIONS FOR SYSTEM DESTRUCTION



Operation DarkSeoul: On March 20, 2013, a cyberattack paralyzed the network services of South Korean media and financial companies. The South Korean government officially announced that the attack was conducted by North Korean hackers for the following three reasons: first, it had discovered the logs that had targeted the victims, which indicated that the attackers had prepared for the operation for a long time; second, well-known IP addresses used by North Korean attackers were found in the South Korean C&C servers; third, the attackers had re-used malware from past operations, specific paths and strings used for creating malware (Kim 2013).

Operation Blockbuster: On November 24, 2014, employees of Sony Pictures Entertainment arrived at work to find their computer screens taken over by a picture of a red skeleton with a message signed "Guardians of Peace." The malware erased data stored on 3,262 of the 6,797 company's personal computers and 837 of 1,555 servers (Elkind 2015). U.S. officials believed that Operation Blockbuster was retribution for the upcoming Sony movie, *The Interview*, a comedy film that involves a plot to assassinate North Korea's leader, Kim (Bing and Lynch 2018). According to the FBI Director, James Comey, due to "mistakes" made by the North Koreans while conducting the operation, the FBI were able to find IP addresses "exclusively used by the North Koreans…several times." (Sanger, Kirkpatrick and Perlroth 2017)

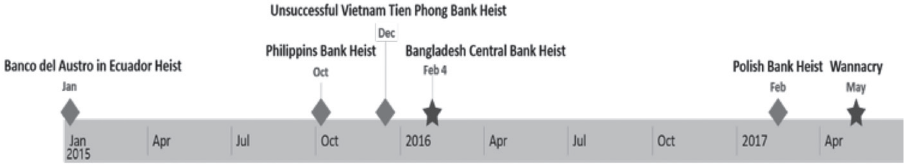
FIGURE 6. THE RELATIONSHIPS BETWEEN CYBERATTACKS IN 2007–2012 BASED ON SHARED TACTICS, TECHNIQUES, AND PROCEDURES (NOVETTA 2015; SYMANTEC SECURITY RESPONSE 2013)



Links: Figure 6 lists various security analyses, revealing the relationships between each cyberattack conducted in 2007–2012 based on shared Tactics, Techniques and Procedures (TTPs), which may indicate the attackers' proximity. It is notable that three attacks – Operation Troy, Ten days of Rain, and DarkSeoul – have the most shared TTPs in their malware, and all ten attacks are linked to the others with at least one shared relationship. TTPs cannot define whether attackers in the 10 attacks were from the same group or had code exchanges, but reveal the possibility of an attacking group's development as follows: Operation Flame → Operation 1Mission → Operation of Ten days of Rain → Operation Troy → Operation DarkSeoul.

C. Operations for Financial Warfare

FIGURE 7. TIMELINE OF OPERATIONS FOR FOREIGN EXCHANGE EARNING



Bangladesh Central Bank Heist: In February 2016, a cyberattack hit Bangladesh Central Bank by exploiting weaknesses in its security to infiltrate its network and steal its SWIFT credentials. The attackers used the stolen SWIFT credentials to make several fraudulent transactions – requests to the Federal Reserve Bank of New York to transfer a total of $101m of the Bangladesh bank's money to locations in the Philippines and Sri Lanka. Four requests to transfer $81m to the Philippines succeeded, but one request to transfer $20m to Sri Lanka was denied because the attackers misspelled the word "foundation" as "fandation." (Volkov 2017)

Operation WannaCry: In May 2017, WannaCry ransomware spread throughout the world via Jaku, a tool for targeted tracking and data exfiltration disguised as botnet malware (Ilascu 2018). The ransomware demanded $300 in Bitcoin per victim; however, according to London-based Elliptic Enterprises, an organization tracking illicit Bitcoin activities, very few victims of the WannaCry attack paid up: only $91,000 had been deposited in the three Bitcoin wallet accounts associated with the ransomware as of May 19, 2017 (Talmadge 2017).

Links: A series of bank heists in the timeline show TTPs related to the Bangladesh Central Bank heist. According to Symantec, researchers uncovered shared TTPs among the three bank heists at the Bangladesh Central Bank, the Philippines Bank, and the Vietnam Tien Phong Bank (Pham, Nguyen and Finkle 2016), meaning that different banks were targeted by the same group (Symantec Security Response 2016). The attackers used stolen credentials to send what looked like legitimate transfer requests to the SWIFT network and used malware after the attack to cover up the evidence of fraudulent transfers (Carter 2017).

D. Recent Attacks and Expected Future Attacks

In 2016–2018: Through human intelligence, the authors were able to gather a list of 21 cyberattacks attributed to North Korean hackers in 2016–2018. The attribution could be proved by the access logs of definitive IP addresses of North Korean attackers in Korean C&C Servers. The targets of these attacks were all located in South Korea or were North Korean defectors abroad. The data will be visualized for legibility.

FIGURE 8. CYBERATTACKS OF NORTH KOREA IN 2016–2018 CATEGORIZED BY TARGET

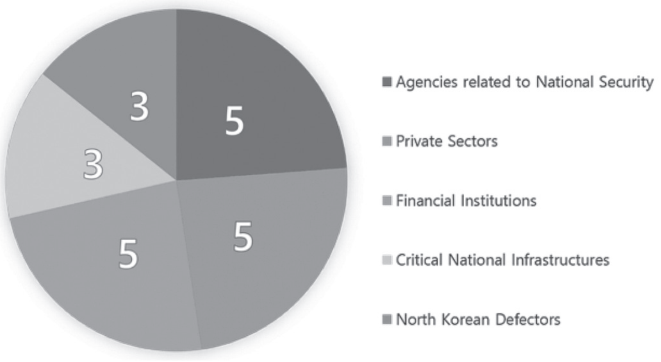
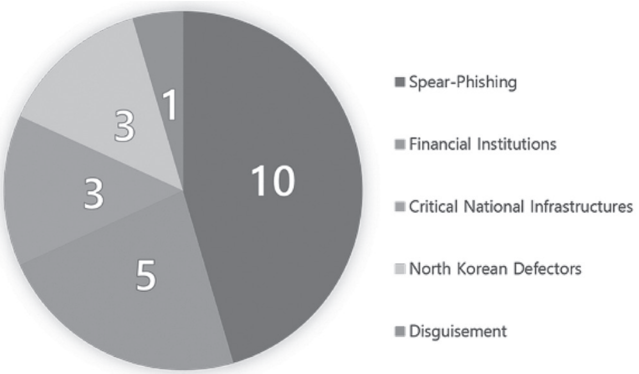


FIGURE 9. CYBERATTACKS OF NORTH KOREA IN 2016–2018 CATEGORIZED BY ATTACK VECTOR



It is noticeable in Figure 8 that a high proportion of targets were related to South Korean national security. The specific agencies related to South Korean national security are the ROK Ministry of National Defense, the National Police Agency, and Defense Industries. Critical National Infrastructure in South Korea includes airports, airplane companies, and telecommunications companies. Figure 9 shows that a spear-phishing attack is the most common attack vector used by North Korea.

Expected Future Attacks: The 2018 Defense White Paper stated that North Korea's military strategy is as follows: "During contingencies, there is a strong possibility that North Korean forces will launch surprise attacks using their asymmetric capabilities mainly to set favorable conditions to terminate the war early." North Korea will likely develop cyber operations in more strategic forms by choosing targets with careful consideration and creativity such as the recent attack on Automated Teller Machines, a blind spot that the US and South Korea have had to address directly (Ha 2018).

4. STRATEGIES

A. Overview: TTPs of North Korean Cyber Forces

Analyzing TTPs helps to highlight credential information in cyber operations and to define attackers' attributes such as attack vectors, scenarios, and identities. The followings are the features of TTPs elicited by North Korea's cyber operations: mainly attacks that targeted South Korea in 2007–2018.

Tactics: The tactics of conventional North Korean forces are analogous to the blitzkrieg military strategy: launching attacks quickly and with massive force, without giving victims time to counter it. However, the tactics of the cyber forces have become stealthy and long-term, since the cyberspace environment requires sufficient time for attackers to understand and invade their targeted information systems. (Jun, LaFoy, and Sohn 2015).

Techniques: The techniques used in North Korean cyber operations to target South Korea are comparatively sophisticated. The most common are one-day exploits and zero-day exploits of Adobe Flash, hwp files, and ActiveX programs (FireEye 2018). However, some analysts criticize the exaggerated media portrayal of these cyberattacks' technical sophistication. Ben Buchanan's report provides a rigorous framework with which to analyze the technical and operational factors of attacks; and highlights other important considerations, such as attackers' tendency to be cost-effective (so that they do not always perform technically sophisticated attacks), the choices that intruders make, tradeoffs between cost and effect, the timeliness of attacks, and barriers to entry for certain types of operation (Buchanan 2017).

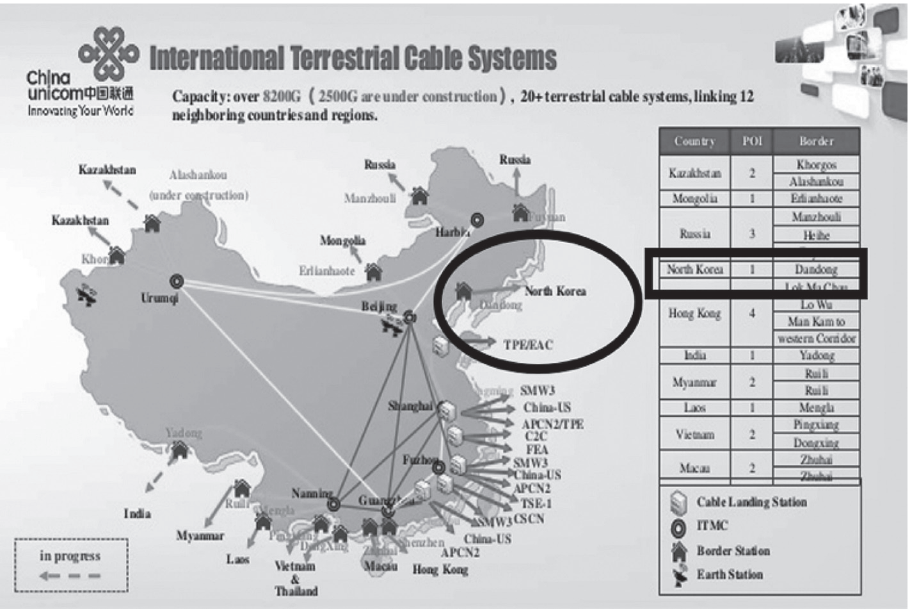
Procedures: North Korea spends a long time reconnoitering targets before attacking through highly profiled means, such as sending spear-phishing emails to infect targets' computers. Meanwhile, attackers hack websites to use them as watering-hole attack vectors or C&C servers. Once attackers have successfully penetrated the internal network of a target system by visiting the target router to the infected website, they initiate an investigation of valuable information. Obtained information is compressed and sent to the C&C server through a secure channel. Finally, the attackers destroy targets or leave bots for additional purposes (Meyers 2018).

B. The Future Military Strategy of North Korean Cyber Forces and Countermeasures

North Korea can mix its TTPs with those of other countries in two ways by sending North Korean cyber forces to third-party countries to conduct cyber operations, with or without the cooperation and consent of the host countries. By sending cyber forces to other countries, North Korea can overcome its limitations, gaining access to a continuous and stable electricity supply and avoiding any need to use North Korea-assigned IP addresses for conducting cyber operations (Sanger, Kirkpatrick and Perlroth 2017).

North Korea with China: North Korea's cyber strategy is said to imitate Chinese military doctrine. The JomHul strategy means pursuing the best result by targeting the weakest part of the enemy's information system to paralyze the whole (Lim, Kwon, Jang, and Baek 2013).

FIGURE 10. THE FIBER-OPTIC CABLE OF CHINA UNICOM LINKS DANDONG IN CHINA TO NORTH KOREA (CHINA UNICOM 2016)



International Terrestrial Cable Systems

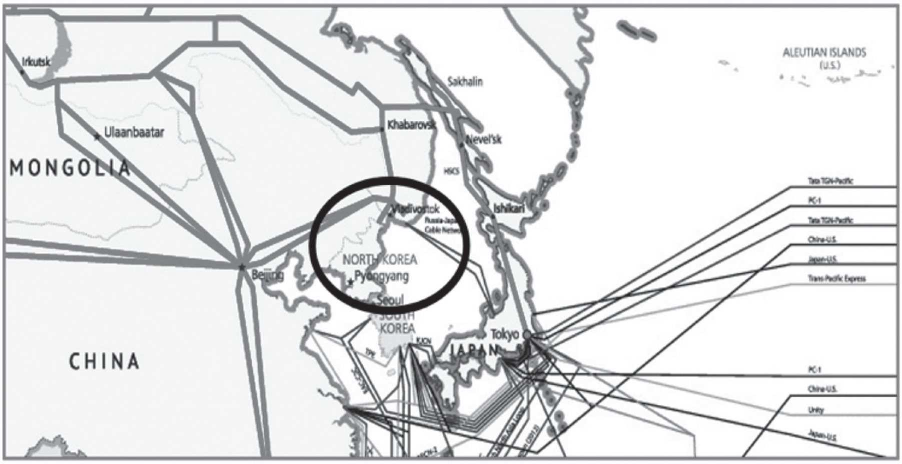
Capacity: over 8200G（2500G are under construction）, 20+ terrestrial cable systems, linking 12 neighboring countries and regions.

Country	POI	Border
Kazakhstan	2	Khorgos
		Alashankou
Mongolia	1	Erlianhaote
Russia	3	Manzhouli
		Heihe
		[illegible]
North Korea	1	Dandong
Hong Kong	4	Lok Ma Chau
		Lo Wu
		Man Kam to
		western Corridor
India	1	Yadong
Myanmar	2	Ruili
		Ruili
Laos	1	Mengla
Vietnam	2	Pingxiang
		Dongxing
Macau	2	Zhuhai
		Zhuhai

It has been consistently reported that North Korean cyber units are active in China. Defectors have verified that North Korea dispatches teams of hackers to carry out offensive cyber operations in Shenyang, China (Horowitz 2017). The U.S. Department of Justice revealed that North Korean cyber operatives such as Park Jin Hyok operate from China (U.S. Department of Justice 2018). Conducting operations under the shadow of China provides North Korean hackers with benefits in terms of attributing attackers; and even when attributed, North Korea can avoid diplomatic issues due to the jurisdiction problem. In fact, the Chinese Embassy in Washington D.C. refused to answer questions regarding whether China had supported North Korean cyber operations (Clayton 2013). This relationship is not explained by a historically trusted partnership; but rather, that the attacks conducted by North Korea do not harm China but instead help to balance power with the U.S. in Asia (Sin 2009).

North Korea with Russia: North Korea decided to expand its Internet connection to Russia after its network was paralyzed twice. The first paralysis of the North Korean network was conducted by the U.S. after Operation Blockbuster. The second network paralysis occurred for nine hours after North Korea launched an ICBM on July 29, 2017, according to BGPM (Mok 2017). As a result, TransTeleCom (TTK), one of Russia's largest telecommunications companies, started to provide the Internet to North Korea on October 1, 2017 (Williams 2017).

FIGURE 11. THE FIBER-OPTIC CABLE OF TTK LINKS VLADIVOSTOK IN RUSSIA TO THE NORTH KOREAN BORDER (WILLIAMS 2017)



The wireless network system failure during the opening ceremony of the 2018 PyeongChang Winter Olympics revealed that Russia tried to cause confusion when tracing cyberattacks by mixing its TTPs with those of North Korea (Ellen 2018; GReAT 2018).

North Korea with Iran: North Korea and Iran have had a technology-sharing treaty focused on the cyber sphere since 2012 (Stevenson 2012). Moreover, there are remarks which show that they have cooperated in sharing their TTPs and experiences on cyber warfare to prepare and conduct cyber operations. Several security analyses have indicated that Operation Shamoon, said to be Iran-backed, which hit Saudi Aramco and other oil company networks in August 2012, shared attack techniques and used the same commercially available EldoS RawDisk driver files as Operation Blockbuster (Kaspersky Lab 2014). In addition, Iran may have shared information about the uncovered Stuxnet with North Korea. According to Reuters, a U.S. intelligence official said there was a Stuxnet variation made for North Korea under the condition of only activating when it encountered Korean-language settings on an infected machine. However, due to North Korea's extremely isolated network, the malware could not successfully access the core machines that ran the North's nuclear weapons program (Noyes 2015).

North Korea with India: A report by Recorded Future concluded with confidence that there was a physical presence of North Korean cyber forces in India, by analyzing what significant cyber activity they had conducted. Their cyber activity showed that North Korean students in at least seven universities around the country might be working with several research institutes and government departments (Insikt Group 2017).

North Korea with other countries: North Korean cyber forces can be dispatched to third-party countries to conduct cyber operations without the consent of their governments. According to Recorded Future, which analyzed data on the Internet usage of North Korean cyber forces between April and July 2017, eight nations were identified in which North Koreans maintained an active physical and virtual presence: India, Malaysia, New Zealand, Nepal, Kenya, Mozambique, Indonesia, and China (Insikt Group 2017). A follow-up report by Recorded Future analyzed data for December 2017–March 2018: and added Thailand and Bangladesh to states where North Koreans were likely living and conducting illicit revenue-generation activities (Moriuchi 2018).

Countermeasures: Overall, the above cases indicate the possibility of North Korea conducting cyber operations in cooperation with other countries to make attribution more difficult. If this is true, it may be a great potential threat to other countries around the world, as it means that a targeted country will need to prepare and counteract the cooperating countries. Moreover, even if the attack's attribution is assumed with strong intelligence, collaborating countries can deny their involvement by denying their cooperation and publicly shirking their responsibility to the other country.

To confront the threats imposed by these possible movements in cyberwarfare, a new collective defense coalition model is suggested. The model starts from countries realizing these growing cyber threats as common threats, then gathering states to develop a collective defense against them, in anticipation of this deterrence power being consolidated. NATO is an example of this, as its members built a defensive coalition against common threats.

5. CONCLUSIONS

North Korea develops military strategies by monitoring other wars. By imitating NATO's utilization of C4I Surveillance and Reconnaissance in the Kosovo War, North Korea prepared for its networked military, allowing it to garner attention worldwide for its cyber capabilities and rise as a big player in cyberspace. Despite its relatively weak infrastructure environment, North Korea realized the importance of cyber warfare within asymmetric capabilities and has gradually developed its cyber power in sequential phases.

Through its analysis of cyber units in North Korea's military command structure, this paper stresses the importance of cyber warfare by making direct connections between North Korea's upper leadership and cyber units. The cyber units are largely divided into the RGB and the GSD; the former reports directly to the upper leadership, while the latter conducts cyber operations within its conventional military capabilities. Then, by arranging cyber operations conducted by actual cyber forces, this paper analyzes various operation objectives. This confirms the North Korean leader's resolve to utilize cyber capabilities and illustrates the relationship between operations presumed to be conducted by North Korea through TTP confirmation.

To avoid being traced as an aggressor by TTPs, this paper suggests that North Korea's future strategic direction will involve mixing TTPs with other countries. Considering its limited infrastructure, it is very likely that TTPs will move to and operate in third-party countries. There are two methods through which this can be fulfilled. One involves conducting operations based on preset coordination through diplomatic, political, and military channels, while the other involves dispatching cyber forces and conducting operations without the target country's acknowledgment. Expected future cyber threats from North Korea will be harder to identify; they will grow more sophisticated by continuing and expanding their efforts with third-party countries. As the attacks are apparently conducted by collaborating nations, the target country will face limitations in its ability to protect itself. This paper therefore proposes a defensive coalition model to respond to the growing common cyber threats imposed by North Korea and those countries it cooperates with.

REFERENCES

Ahn, Yonghyun. 2011. “North Korea’s Electonic Warfare Capability.” *ChosunIlbo*, March 7, 2011. http://news.chosun.com/site/data/html_dir/2011/03/07/2011030702345.html.

AhnLab. 2014. “APT Attack - New ‘Kimsuky’ malware emerged.” *ASEC Threat Research & Responding Blog*, March 19, 2014. http://asec.ahnlab.com/993.

Alyac. 2018. “Operation Kimsuky’s secret activities, Korea customized APT attack is currently in progress.” *ESTsecurity Alyac Blog*, Feb 12, 2018. http://blog.alyac.co.kr/1536.

Bechtol Jr., Bruce E. 2018. *North Korean Military Proliferation in The Middle East and Africa*. Kentucky: The University Press of Kentucky.

Buchanan, Ben. 2017. *The Legend of Sophistication in Cyber Operations*. Cambridge: Harvard Kennedy School, Belfer Center for Science and International Affairs.

Bing, Christopher and Lynch, Sarah. 2018. “U.S. charges North Korean hacker in Sony, WannaCry cyberattacks.” *Reuters*, Sep 6, 2018. https://www.reuters.com/article/us-cyber-northkorea-sony/u-s-charges-north-korean-hacker-in-sony-wannacry-cyberattacks-idUSKCN1LM20W.

Carter, William. 2017. “Forces Shaping the Cyber Threat Landscape for Financial Institutions.” *SWIFT Institute Working Paper*, No.2016-004 (October).

Center of Excellence Defence Against Terror. 2008. *Responses to Cyber Terrorism (NATO Science for Peace and Security)*, 199. Texas: IOS Press.

China Unicom. 2016. “China Unicom Global Brief Introduction.” Published on Sep 21, 2016 at *Slideshare*, Slide 9. https://www.slideshare.net/AbhijitDatey/china-unicom-global-profile.

Chung, Kuyoun and Lee, Kitae. 2017. *Advancement of Science and Technology and North Korea’s Asymmetric Threat: Rise of cyberwarfare and unmanned aerial vehicle*. Seoul: Korea Institute for National Unification.

Clayton, Mark. 2013. “In cyber arms race, North Korea emerging as a power, not a pushover.” *The Christian Science Monitor*, Oct 19, 2013. https://www.csmonitor.com/World/Security-Watch/2013/1019 /In-cyberarms-race-North-Korea-emerging-as-a-power-not-a-pushover.

Elkind, Peter. 2015. “Part 1: Who was manning the ramparts at Sony Pictures?” *Fortune*, Jun 25, 2015. http://fortune.com/sony-hack-part-1.

Ellen, Nakashima. 2018. “Russian spies hacked the Olympics and tried to make it look like North Korea did it, U.S. officials say.” *The Washington Post*, Feb 24, 2018. https://www.washingtonpost.com/world/ national-security/russian-spies-hacked-the-olympics-and-tried-to-make-it-look-like-north-korea-did-it-us-officials-say/2018/02/24/44b5468e-18f2-11e8-92c9-376b4fe57ff7_story.html?utm_term=.e4c57176 51a3.

FireEye. 2018. “APT37 (Reaper): The overlooked North Korean actor.” *FireEye*, Feb 20, 2018. https://www.fireeye.com/blog/threat-research/2018/02/apt37-overlooked-north-korean-actor.html.

Gil, Mingwon. 2018. “Korea Customized APT Attack of Kim Soo-ki Hacking Organization... Still.” *Dailysecu*, Feb 13, 2018. https://www.dailysecu.com/?mod=news&act=articleView&idxno=30007.

GReAT. 2018. “OlympicDestroyer is here to trick the industry.” *Kaspersky Lab*, March 8, 2018. https://securelist.com/olympicdestroyer-is-here-to-trick-the-industry/84295.

Ha, Mathew. 2018. “North Korea’s cyber threats are serious, the network of RGB should be disabled.” Interview by No Jungmin. Radio Free Asia, Sep 17, 2018. Audio, 5:47. https://www.rfa.org/korean/ in_focus/news_indepth/ne-jn-11162018162726.html.

Ha, Mathew and Maxwell, David. 2018. *Kim Jong Un's 'All-Purpose Sword' North Korean Cyber-Enabled Economic Warfare*. Washington, DC: FDD Press.

Han, Sangmi. 2016. "North Korea sends 50 to 60 talented students to study abroad to train as cyber agents." *Voice of America*, June 14, 2016. https://www.voakorea.com/a/3375411.html.

Horowitz, Josh. 2017. "Researchers have found an unexpected axis of North Korea's cyber activity: India." *Quartz*, Oct 22, 2017. https://qz.com/1105149/india-is-an-unexpected-axis-of-north-koreas-suspect-cyber-activity.

Ilascu, Ionut. 2018. "A First Look at the North Korean Malware Family Tree." *Bleepingcomputer*, August 9, 2018. https://www.bleepingcomputer.com/news/security/a-first-look-at-the-north-korean-malware-fam ily-tree.

Insikt Group. 2017. "North Korea Cyber Activity." *Recorded Future*, July 25, 2017. https://go.Recordedfutu re.com/hubfs/reports/north-korea-activity.pdf.

Jun, LaFoy, and Sohn. 2015. *North Korea's Cyber Operations: Strategy and responses*. Maryland: Rowman & Littlefield. CSIS Reports.

Kaspersky Lab. 2014. "Sony Sony/Destover: Mystery North Korean actor's destructive and past network activity." *Kaspersky Lab*, Dec 4, 2014. https://securelist.com/destover/67985.

Kim, Heungkwang. 2010. "Responses and Strategies against North Korea's Cyber Information Warfare." *North Korea Intellectuals Solidarity*, July 2, 2010. http://www.nkis.kr/board.php?board=nkisb501&page =1&sort=hit&command=body&no=3.

Kim, Kyungae. 2013. "Detailed explanation of government announcement of 3.20 cyber terrorism case." *Boan News*, April 12, 2013. https://www.boannews.com/media/view.asp?idx=35649.

Kim, Seungju. 2014. "North Korea's cyber-attack, and our response." *Monthly North Korea*, no. 516 (December): 66-71.

Lee, Yongsu. 2013. "Kim Jong-un, 'with brave cyber warriors, we can penetrate any sanctions.'" *ChosunIlbo*, April 8, 2013. http://news.chosun.com/site/data/html_dir/2013/04/08/2013040800165.html.

Lim, Kwon, Jang, and Baek. 2013. "North Korea's Cyber War Capability and South Korea's National Counterstrategy." *The Quarterly Journal of Defense Policy Studies*, 29th Issue 4 Winter 2013(Article 102)

Madde, Michael. 2018. "Kim Yong Chol, A Biography." *38 North*, May 29, 2018. https://www.38north.org/ 2018/05/mmadden052918.

Meyers, Adam. 2018. "Negotiations with North Korea may have Cyber Consequences." *38 North*, MARCH 13, 2018. https://www.38north.org/2018/03/ameyers031318.

Mok, Yongjae. 2017. "After ICBM provocation, North Korea Internet 9 hours paralysis." *RFA*, Jul 31, 2017. https://www.rfa.org/korean/in_focus/nk_nuclear_talks/internetdown-07312017092147.html.

Mok, Yongjae. 2017b. "6 Cyber Units were built after Kim Jong-un regime." *RFA*, Nov 22, 2017. https:// www. rfa.org/korean/in_focus/news_indepth/ne-jn-11162018162726.html.

Moriuchi, Priscilla. 2018. "North Korea's Ruling Elite Adapt Internet Behavior to Foreign Scrutiny." *Recorded Future*, April 25, 2018. https://www.recordedfuture.com/north-korea-internet-behavior.

Novetta. 2015. "Operation Blockbuster: Unraveling the Long Thread of the Sony Attack." *Novetta*, Feb 5, 2015. https://www.operationblockbuster.com/wp-content/uploads/2016/02/Operation-Blockbuster-Rep ort.pdf.

Noyes, Katherine. 2015. "The NSA reportedly tried - but failed - in Stuxnet strike against North Korea." *IDC News Service*, Jun 1, 2015. https://www.computerworlduk.com/security/nsa-reportedly-tried-failed-use-stuxnet-variant-against-north-korea-3613758.

Oxford Dictionary. n.d. “Campaign.” Accessed March 26, 2019. https://en.oxforddictionaries.com/definition/campaign.

Park, Moonbeom. 2018. “Let’s learn about enemy through various IoCs of real APT cases.” In *DragonCon 2018*, Dec 8, 2018. Dragon Threat Labs.

Pham, Nguyen and Finkle. 2016. “Vietnam bank says interrupted cyber heist using SWIFT messaging.” *Reuters*, May 15, 2016. https://www.reuters.com/article/us-vietnam-cybercrime/vietnam-bank-says-interrupted-cyber-heist-using-swift-messaging-idUSKCN0Y60EN.

ROK Ministry of National Defense. 2018. “2018 Defense White Paper.” Seoul: ROK Ministry of National Defense.

Sanger, Kirkpatrick and Perlroth. 2017. “The World Once Laughed at North Korean Cyberpower. No More.” *The New York Times*, Oct 15, 2017, https://www.nytimes.com/2017/10/15/world/asia/north-korea-hacking-cyber-sony.html.

Security News Special Coverage Team. 2014. “KHNP, staff information leaked to technical data.” *BoanNews*, Dec 14, 2014. https://www.boannews.com/media/view.asp?idx=44734.

Seoul Central District Public Prosecutors’ Office. 2015. “Intermediate investigation result of KHNP cyber terrorism case.” *Supreme Prosecutors’ Office*, March 16, 2015. http://www.spo.go.kr/spo/notice/press/press.jsp?mode=view&board_no=2&article_no=593028.

Sin, Steve S. 2009. “Cyber Threat posed by North Korea and China to South Korea and US Forces Korea.” *Defense and Technology*, 364 (2009): 28-33.

Stevenson, Alastair. 2012. “Iran and North Korea sign technology treaty to combat hostile malware.” *V3*, Sep 3, 2012. https://www.v3.co.uk/v3-uk/news/2202493/iran-and-north-korea-sign-technology-treaty-to-co mbat-hostile-malware.

Symantec Security Response. 2013. “Four Years of DarkSeoul Cyberattacks Against South Korea Continue on Anniversary of Korean War.” *Symantec Official Blog*, Jun 26, 2013. https://www.symantec.com/ connect/blogs/four-years-darkseoul-cyberattacks-against-south-kore a-continue-anniversary-korean-war.

Symantec Security Response. 2016. “SWIFT attackers’ malware linked to more financial attacks.” *Symantec*, May 26, 2016. https://www.symantec.com/connect/blogs/swift-attackers-malware-linked-more-finan cial-attacks.

Talmadge, Eric. 2017. “Experts question North Korea role in WannaCry cyberattack.” *AP News*, May 20, 2017. https://www.apnews.com/ed3298eaaff84e8ebb091cbbd4bc4ab6.

Tarakanov, Dmitry. 2013. “The ‘Kimsuky’ Operation: A North Korean APT?” *Securelist*, Sep 11, 2013. https://securelist.com/the-kimsuky-operation-a-north-korean-apt/57915.

U.S. Department of Justice. 2018. “North Korean Regime-Backed Programmer Charged with Conspiracy to Conduct Multiple Cyberattacks and Intrusions.” Press Release, Sep 6, 2018. https://www.justice.gov/usa o-cdca/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyberattacks.

U.S. Special Operations Command. 2015. “White Paper: The Gray Zone.” *USSOCOM*, Sep 9, 2015. https://info.publicintelligence.net/USSOCOM-GrayZones.pdf.

Volkov, Dmitry. 2017. “Lazarus Arisen: Architecture, Techniques and Attribution.” *Group IB*, May 30, 2017. https://www.group-ib.com/blog/lazarus.

Williams, Martyn. 2017. “Russia Provides New Internet Connection to North Korea.” *38 North*, Oct 1, 2017. https://www.38north.org/2017/10/mwilliams100117.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

NATO Members' Organizational Path Towards Conducting Offensive Cyber Operations: A Framework for Analysis

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

Abstract: NATO member states are starting to talk more openly about the incentives and opportunities to conduct offensive cyber operations for military purposes. This growing interest in 'offensive cyber' is most clearly expressed in the creation of cyber commands, branches or services within the armed forces. Little research, however, has analyzed these organizational developments. This article provides a conceptual framework to facilitate empirical analysis across military cyber organizations (MCOs). The framework distinguishes between five stages of organizational development: i) seed, ii) startup, iii) growth, iv) expansion, and v) maturity. Our empirical analysis reveals that a significant number of NATO members started to carefully consider establishing MCOs from 2008 onwards, and some states had already started significant organizational efforts in the 1990s. However, I also reveal that the MCOs of most NATO members are still at the early stages of organizational development, and even those at the growth stage still have limited budgets to address the different workforce, capability, strategic and other requirements.

Keywords: *NATO, military cyber organization, offensive cyber operations, development life cycle*

1. INTRODUCTION

Over the years, NATO members have presented and rolled out several plans for improving cyber-defense governance. Official commitments made at NATO summits on cyber security have become increasingly granular.[1] One topic that government leaders have long avoided talking about, however, is their *own* willingness and capacity to conduct military cyber operations.

Times are changing. As one senior official put it at a military cyber conference: "Speaking at NATO about offensive cyber was blasphemy a few years ago. We have advanced".[2] Last year the Alliance reached a landmark that went largely unnoticed: there are now more member states which have publicly declared they are seeking to establish an offensive cyber capability than there are member states which have remained publicly silent on this issue.[3] In late 2018, it was also announced that five countries would contribute national cyber forces to NATO missions and operations. This group consists of the United States, the United Kingdom, Denmark, Estonia, and the Netherlands.[4]

The growing interest in offensive cyber operations for military purposes is most clearly expressed in the creation of cyber commands, branches or services within the armed forces. These military cyber organizations (MCOs), as Piret Pernik from ICDS noticed in her study, are often predicated "by the need to centralise, consolidate, and streamline formerly fragmented capabilities and organisations, while eliminating overlapping roles and responsibilities" to effectively operate in this new "operational domain".[5]

Academic scholarship and policy research is still lagging behind in analyzing these developments. We still lack a comprehensive overview of where NATO member

[1] These include: Prague (2002), Riga (2006), Bucharest (2008), Strasbourg (2009), Lisbon (2010), Chicago (2012), Wales (2014), Warsaw (2016) and Brussels (2018). For a good overview on early thinking see: John B. Sheldon, "NATO and Cyber Defense: Hanging Together or Hanging Separately?", Presentation, (year unknown), hxxp://www.unidir.ch/files/conferences/pdfs/nato-and-cyber-defence-hanging-together-or-hanging-separatelyen-1-608.pdf.

[2] See: Dutch Ministry of Defense, Third International Cyber Operations Symposium, (2017, October); also see: Sophie Arts, "Offense as the New Defense: New Life for NATO's Cyber Policy," The German Marshall Fund of the United States, Policy Brief, 39, (2018):1-9.

[3] 'Offensive cyber capability' refers in this context to a broad set of capabilities referred to by states, including 'cyberwarfare capabilities'. 'military cyber arsenal', 'Computer Network Attack capabilities', and 'military cyber offense'. Section IV provides a more detailed overview on the use of different terminology and developments within each country.

[4] US Department of Defense, "News Conference By Secretary Mattis at NATO Headquarters, Brussels, Belgium," US Department of DEcen, (2018, October 4), retrieved from: dod.defense.gov; Also see the Brussel Summit Declaration for a reaffirmation of NATO mandate and cyber efforts: NATO, "Brussels Summit Declaration, (2018, July 11-12), retrieved from: https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2018_07/20180713_180711-summit-declaration-eng.pdf.

[5] Piret Pernik, "Preparing for Cyber Conflict: Case Studies of Cyber Command," (2018, December), retrieved from: https://icds.ee/wp-content/uploads/2018/12/ICDS_Report_Preparing_for_Cyber_Conflict_Piret_Pernik_December_2018-1.pdf.

states stand in terms of organizational development. The purpose of this article is therefore to provide a conceptual framework to facilitate analysis and comparison between different MCOs.[6]

This paper's argument is developed in five parts. Section II provides a framework which distinguishes between five stages of organizational development: i) seed and development, ii) startup, iii) growth, iv) expansion, and v) maturity. Section III offers an empirical perspective, providing an historical overview of member states' organizational achievements. The section indicates that a significant number of NATO members started to think about military cyber operations from 2008 onwards, and some states had already started significant organizational efforts in the 1990s. Yet, I also reveal that the MCOs of most states are still at the early stages of organizational development, and even those at the growth stage still have limited budgets. Section IV provides additional considerations about NATO members' MCO development. Section V concludes and identifies avenues for future research.

2. A CONCEPTUAL FRAMEWORK FOR MCOS

A military cyber organization is defined as a command, service, branch or unit within a government's armed forces which has the authority and mission to conduct offensive cyber operations to disrupt, deny, degrade and/or destroy (d4 effects).

MCOs come in all shapes and forms. Countries have different strategic objectives and approaches and base their decisions on different legal and organizational prerequisites. In some countries, MCOs can be authorized to direct and control the full spectrum of cyber operations. Other MCOs only have the narrow authority (following a mandate) to execute a small set of offensive missions. Some states' MCOs are small: their workforce could easily fit into a few school buses; for others, you would need a fleet of Boeing 747s to transport its workforce. Finally, some MCOs are expected to play a role in defense and resiliency efforts such as assisting civilian authorities in protecting critical infrastructure. Others are not.

Given this variation in MCOs, we must use a framework that balances two considerations: on the one hand, the framework needs to be sufficiently general to incorporate significant variation in missions and organizational structures; whilst, on the other hand, the framework's categorical distinctions need to be specific enough to capture empirical progress in a meaningful manner.

In finance and business management literature, the concept of the 'business life cycle'

[6] The goal of this article is not to provide an in-depth case study of a specific country's organizational development. Nor is the purpose of this research to explain why states seek to establish MCOs. For an analysis of this question see: Max Smeets, "Going cyber : the dynamics of cyber proliferation and international security," DPhil Dissertation in International Relations, University of Oxford, 2017.

is widely used to help with the strategic planning and operations of a company.[7] The idea is that the progression of a company can be divided into several stages, each with its own opportunities and challenges. For example, a small business will initially have to focus on market acceptance and determining a profitable business structure. It subsequently needs to think more carefully about how it can establish a customer base and manage financial issues such as funding and cash reserves. In later stages, different issues will have to be considered, such as dealing with (increased) market competition and expanding into new markets and distribution channels. Solutions which may have worked for one stage may not work in another. This means that businesses have to adjust operations accordingly.[8]

MCOs are not corporations, yet we can deploy a similar framework for this type of institutional development.[9] An overview of the stages and their associated challenges is provided in *Table I: The Life Cycle of an MCO*.

TABLE I: THE LIFE CYCLE OF AN MCO

Stage	Description
Seed & Development	A government recognizes the importance of investing in offensive cyber, and talks about the need to establish an MCO.
Startup / Launch	There is the political authorization to establish an MCO
Growth	The MCO has moved towards an actual operational capacity.
Expansion	The MCO has repeatedly conducted offensive cyber operations and is potentially assessing new options for further development.
Maturity	The MCO is able to conduct full spectrum operations against a wide range of targets, embedded in a strategy that has proven to be effective.

Each NATO member state begins at the *seed and development* phase: this is when senior officials within the government start to discuss the importance of establishing an MCO to conduct offensive cyber operations.[10] A government moves to the *startup* phase when the political authorization to establish an MCO is issued.[11] During the

[7] Neil Petch, “The Five Stages Of Your Business Lifecycle: Which Phase Are You In?,” *Entrepreneur*, (2016, February 29), retrieved from: https://www.entrepreneur.com/article/271290 For an alternative overview see: Neil C. Churchill and Virginia L. Lewis, “The Five Stages of Small Business Growth,” *Harvard Business Review*, (1983, May), retrieved from: https://hbr.org/1983/05/the-five-stages-of-small-business-growth.

[8] In the same vein, unresolved challenges from an earlier stage may also come to haunt a business at later stages. For example, missing a lack of accounting management initially might hinder to have an accurate reflection of the later business finances.

[9] Whilst using the same categories, I do not mean to suggest that the dynamics underlying each stage of organizational development are the same for MCOs as a business.

[10] This generally accumulates into a national security strategy which indicates that the government should start to invest in ‘offensive cyber capabilities’. There could be multiple reasons why the governments starts to talk about the need to establish an MCO – one could be the strategic landscape.

[11] Political authorization may come from various authorities, such as the parliament, government, president or minister of defense. It is normally part of the defense planning.

growth stage the MCO begins developing an actual operational capability. When an MCO has started to conduct offensive cyber operations it enters the *expansion* phase. The final stage of the MCO life cycle is called *maturity*: an MCO at this stage is able to conduct full spectrum operations against a wide range of targets, as part of a deliberate strategy embedded in the structural dynamics of cyberspace.[12]

The MCO life-cycle is non-deterministic. Each MCO follows its own path; they can progress and regress over time, and take more or less time to transition between stages. For instance, an MCO may lose its initial operational capability or forever be stuck in the *startup* phase and never actually conduct offensive cyber operations. And if a state has a well-established signal intelligence unit, it may rely on those knowledge-structures to quickly move from the *launch* to the *growth* stage.

3. AN OVERVIEW OF NATO MEMBERS

This section offers an historical perspective of the institutional progress across members of the NATO alliance. Table II provides a baseline overview for where NATO member states currently stand in terms of MCO development, based on publicly available information.

It is hardly surprising that the United States was among the first countries in the NATO alliance that sought to conduct offensive cyber operations to achieve d4 effects. Since the 1980s there had been a growing awareness in the US of the military potential of computer attacks, according to Michael Warner, U.S. Cyber Command historian.[13] It was Operation Desert Storm, in 1991, which is said to have given further impetus to the importance of conducting military cyber operations as part of modern warfare.[14] In the US, information warfare centers were officially created by the Air Force in 1993 and a year later by the Navy and Army. In the same period, the NSA set up the Tailored Access Operations (TAO) unit.[15]

In mid-2009, Secretary of Defense Robert Gates directed the commander of U.S. Strategic Command to establish a sub-unified command, Cyber Command

[12] For an overview of potential strategic use see: Max Smeets and Herbert Lin, "Offensive Cyber Capabilities: To What Ends?" *2018 10th International Conference on Cyber Conflict CyCon X: Maximising Effects*, T. Minárik, R. Jakschis, L. Lindström (Eds.) (Tallinn: NATO CCD COE Publications: 2018); Max Smeets,"The Strategic Promise of Offensive Cyber Operations," Strategic Studies Quarterly, (2018, Fall):90-113.

[13] Rid provides a similar statement: "Defense intellectuals slowly began to discern an offensive and a defensive logic in what Post called 'cybernetic war' in 1979. This development took some time". Michael Warner, "Cybersecurity: A Pre-history", *Intelligence and National Security*, Intelligence and National Security, 27:5 (2012)781-799.

[14] For a more in-depth discussion on this, see: Ronald J. Deibert, "Black Code: Censorship, Surveillance, and the Militarisation of Cyberspace," *Millennium: Journal of International Studies* (2003).

[15] For a more detailed history, see: Joint Task Force Global Network Operations, "A Legacy of Excellence: December 30, 1998- September 7, 2010", retrieved from: https://assets.documentcloud.org/documents/2849764/Document-05.pdf.

(USCYBERCOM).[16] The creation of this organization "marked the culmination of more than a decade's worth of institutional change. DoD defensive and offensive capabilities were now firmly linked, and, moreover, tied closely, with the nation's cryptologic system and premier information assurance entity, the NSA".[17] USCYBERCOM has grown significantly ever since – achieving full operational capability (133 teams) in May 2018.[18] In the same month, the Department of Defense (DoD) also elevated USCYBERCOM to a unified combatant command.[19]

Another early case – often overlooked – is that of Greece, where the government officially established an Office of Computer Warfare in 1999.[20] Five years later, in 2004, the Department of Cyber Defense was established, which was subsequently elevated to the Directorate of Cyber Defense in 2011.[21] Even though the Greek institution's development might look significant on paper, as John Nomikos writes, there is currently a lack of funding due to austerity measures, making it difficult to operate.[22] In that sense, it is unclear if the country ever passed the launch phase and actually started to conduct military cyber operations.[23]

[16] For a pre-institutional history of the U.S. Cyber Command, see United States Strategic Command, "JFT-CND/JTC-CNO/JTF-GNO: A Legacy of Excellence" (1998, December 30/ 2010, September 7), retrieved from: https://nsarchive2.gwu.edu//dc.html?doc=2849764-Document-05).

[17] Michael Warner, "U.S. Cyber Command's Road to Full Operational Capability," in *Stand Up and Fight: The Creation of U.S. Security Organizations, 1942–2005*, edited by Ty Seidule and Jacqueline E. Whitt (Carlisle, Penn.: Strategic Studies Institute and U.S. Army War College Press, 2015), chap. 7.

[18] Max Smeets and Herbert Lin, "4 A Strategic Assessment of the U.S. Cyber Command Vision," in *Bytes, Bombs, and Spies: The Strategic Dimensions of Offensive Cyber Operations*, (Washington DC: Brookings Institution Press: 2018), pp. 81-104.

[19] Jim Garamone and Lisa Ferdinando, "DoD Initiates Process to Elevate U.S. Cyber Command to Unified Combatant Command," Department of Defense News, (2017, August 18), retrieved from: www.defense.gov/News/Article/Article/1283326/dod-initiates-process-to-elevate-us-cyber-command-to-unified-combatant-command/.

[20] United Nations Institute for Disarmament Research (UNIDIR), "The Cyber Index International Security Trends and Realities," (2013), retrieved from: http://www.unidir.org/files/publications/pdfs/cyber-index-2013-en-463.pdf.

[21] John M. Nomikos, "Intelligence Requirements for Cyber Defense, Critical Infrastructure and Energy Security in Greece," *National Security and National Future*, 1-2:17 (2016).

[22] Ibid.

[23] There are no cases of CNA publicly attributed to the Greek government.

TABLE II: OVERVIEW OF MCO DEVELOPMENT IN NATO MEMBER STATES

	Seed & Development	Start Up / Launch	Growth	Expansion	Maturity
United States	1980s	2010	2011	Present	
Estonia	2017	2017	Present		
France	2008	2011	2016	Present	
Germany	2009	2016	Present		
Italy	2014	2015	Present		
The Netherlands	2012	2015	2018		
Spain	2012	2014	Present		
Turkey	2011	2012	Present		
Belgium	2015	2019	Present		
Canada	2011	2011	2015	Present	
Denmark	2013	2017	Present		
Greece	1999	2004	Present		
Norway	2012	2012			
Poland	2008	Unknown	Unknown		
United Kingdom	2012	2012			
Portugal	2015	Unknown			

* In 2008, Poland proposed to develop an independent information force. As yet, it is unclear to what degree it is operational and focuses on OCO to achieve d4 effects; In the 2011 National Strategic Framework, Italy mentions various cyber initiatives but leaves out the development of military offensive capability.
** It remains unclear to what degree Norway's Cyber Defense branch can actually be defined as an MCO, given its narrow mission.[24]
*** There are no known MCO developments in the following NATO countries: Albania, Bulgaria, Croatia, Czech Republic, Hungary, Iceland, Latvia, Lithuania, Luxembourg, Montenegro, Slovakia and Slovenia.

Most Alliance members started to talk publicly about the need for 'cyberwarfare capabilities' in the mid-2000s. The majority are now in either the *launch* or *growth* stages. For example, the Netherlands passed the *seed* phase about eight years ago, when the Dutch government mentioned in several government publications and news articles the need to develop an offensive cyber capability to effectively 'defend and deter' other actors.[25] The government developed its political and military priorities in cyberspace through a number of official publications, including the first National Cybersecurity Strategy (2011), the Defense Cyber Strategy (2012), the second

[24] Also see: Lilly Pijnenburg Muller, "Military Offensive Cyber-Capabilities: Small-State Perspectives", Norwegian institute of International Affairs, Policy Brief, 1, (2019), retrieved from: https://brage.bibsys.no/xmlui/bitstream/handle/11250/2583385/NUPI_Policy_Brief_1_2019_Muller.pdf?sequence=1&isAllowed=y.

[25] Strategic documents followed several parliamentary inquiries by two members of parliament (Raymond Knops and Marcial Hernandez) in 2010 and 2011.

National Cybersecurity Strategy (2013), and the Defense Cyber Strategy (2015).[26] The *startup* phase, commenced in the June 2015 when the Defense Cyber Command (DCC) was officially established. The DCC incorporates the Taskforce Cyber (TFC), established in 2012, under the Army.[27] Last year it reached the *growth* stage when it became operational, though it is known to struggle operationally.[28]

The Danish government writes in its Defense Agreement 2013-2017 that the country's "defense must have the capability for military operations in cyberspace, including the ability to protect own network infrastructure, and also to affect opponents' use of cyberspace".[29] It also explicitly states that the government should develop a "capacity that can execute defensive and offensive military operations in cyberspace".[30] In its 2012 National Cyber Security Strategy, Spain writes that one "line of action" is to "boost military and intelligence capabilities to deliver a timely, legitimate and proportionate response in cyberspace to threats or aggressions that can affect National Defence".[31] In 2011, Turkey revealed plans to establish a Cyber Command, which was officially established a year later (called the General Staff Warfare and Cyber Defense Command). At a conference in 2014, the commander of the military General Staff's Division for Electronic Systems and Cyber Defense said that Turkey considers "cyber" to be the "fifth military domain".[32]

For a long time the British government was coy in public about the offensive operations it sought to conduct and the doctrine it was following. Since 2012 this has started to change.[33] The National Cyber Security Strategy 2016-2021 is unequivocal about Britain's ambitions in this new domain. It states that: "Offensive cyber forms part of the full spectrum of capabilities we will develop to deter adversaries and to deny them opportunities to attack us, in both cyberspace and the physical sphere".[34] The UK aims to become "a world leader in offensive cyber capability; and […] to establish "a

[26] For an excellent overview see: Paul Ducheine, "Defensie in het digitale domein," *Militaire Spectator*, 186:4 (2017)152-168.

[27] One could potentially argue that the startup phase already started in 2012 with the establishment of the Taskforce Cyber.

[28] Liza van Lonkhuyzen and Kees Versteegh, "Het cyberleger kan en mag nog weinig," *NRC* (2018, December 18), retrieved from: https://www.nrc.nl/nieuws/2018/12/18/het-cyberleger-is-er-wel-maar-mag-weinig-a3099254.

[29] The report also mentions that: "Focus on transverse planning and deployment of capabilities, challenges in the Arctic and in cyberspace, as well as the adaptation of not least the army, will dominate the development of the defense". See: The Danish Ministry of Defense, "Danish Defense Agreement 2013-2017", (2012, November 30), retrieved from: http://www.fmn.dk/eng/allabout/Documents/TheDanishDefenceAgrement2013-2017english-version.pdf, p. 4 and p.8.

[30] Ibid, p. 16.

[31] Rajoy Brey, "National Cyber Security Strategy of Spain," (2013), retrieved from: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/NCSS_ESen.pdf, p.32.

[32] United Nations Institute for Disarmament Research (UNIDIR), "The Cyber Index International Security Trends and Realities".

[33] Since 2012, the Joint Forces Command has taken the lead in integrating and conducting offensive cyber operations.

[34] UK government, "Britain's cyber security bolstered by world-class strategy," (2016, November 1), retrieved from: https://www.gov.uk/government/news/britains-cyber-security-bolstered-by-world-class-strategy.

pipeline of skills and expertise to develop and deploy our sovereign offensive cyber capabilities".[35]

It is not always easy to delineate the *seed* and *development* from the *startup* phase. In some countries, overlapping organizations were created or reorganized over the course of several years. For example, in 2011, Canada set up the Directorate of Cybernetics to "build cyberwarfare capabilities" for the armed forces.[36] But, as James Lewis notes, earlier "[t]he Canadian Armed Forces Information Management Group [was] responsible for the protection of the armed forces' computer and communications networks" with subsidiary organizations including "the Canadian Forces Network Operation Centre as well as a centre for electronic warfare and signals intelligence".[37]

Another potentially ambiguous case is France. Bernard Barbier, the former director of France's external intelligence agency (Directorate-General for External Security), said at a university lecture that the country had already explored conducting espionage operations in the early 1990s and quickly moved on to also think about warfare applications.[38] Yet the French only publicly talked about the potential conduct of military cyber operations in 2008, in a White Paper under then President Nicolas Sarközy.[39] However, as Arthur Laudrain notes, France has from 2016-2019 "conceptualized and adopted a comprehensive cybersecurity and cyber defense model".[40] In late 2016, the then-Minister of Defense Jean-Yves Le Drian announced the creation of a new cyber defense command (COMCYBER) predicted to employ 2,500 personnel by 2019 and receiving an initial commitment of €2.1 billion in funding,.[41] In 2017, the Strategic Review for Defense and National Security was published recognizing cybersecurity and 'digital sovereignty' as top priority.[42] In February 2018, France published its first National Strategy for Cyber Defense clarifying how cyber operations are organizationally integrated as well as the legal framework surrounding their use. And

[35] Ibid.

[36] Matteo Gramaglia, Emmet Tuohy, Piret Pernik. "Military Cyber Defense Structures of NATO Members," The Star, (2016, January 9), retrieved from: https://www.thestar.com/news/canada/2016/09/01/former-electronic- spy-chief- urges-ottawa-to- prepare- for-cyber- war.html ; Alex Boutilier, "Canada developing arsenal of cyber-weapons," The Star, (2017, March 16), retrieved from: https://www.thestar.com/news/canada/2017/03/16/canada-developing-arsenal-of-cyber-weapons.html;.

[37] United Nations Institute for Disarmament Research (UNIDIR), "The Cyber Index International Security Trends and Realities".

[38] Henri Chain, "Espionnage et cybersécurité, Bernard Barbier reçu par Symposium CentraleSupélec," (2016, September 5), retrieved from: https://www.youtube.com/watch?v=s8gCaySejr4.

[39] Nicolas Sarközy, "The French White Paper on Defence and National Security", (New York: Odile Jacob Publishing Corporation: 2008), retrieved from: http://www.mocr.army.cz/images/Bilakniha/ZSD/French%20White%20Paper%20on%20Defence%20and%20National%20Security%202008.pdf;.

[40] Arthur Laudrain, "France's New Offensive Cyber Doctrine," (2019, February 26), *Lawfare*, retrieved from: https://www.lawfareblog.com/frances-new-offensive-cyber-doctrine.

[41] France has previously developed espionage platform Animal Farm. As yet, there is no public report indicating the country is conducting CNA operations. Tom Reeve, "France unveils cyber command in response to 'new era in warfare' ," *SC magazine*, (2016, December 16), retrieved from: https://www.scmagazineuk.com/france-unveils-cyber-command-in-response-to-new-era-in-warfare/article/579671/.

[42] République Française "Strategic Review of Defense and National Security: Key Points," (2017) retrieved from: https://www.defense.gouv.fr/content/download/514686/8664672/file/2017-RS-PointsClesEN.pdf.

in January 2019, France unveiled its first offensive cyber doctrine, marking another crucial milestone.[43]

This means that France and Germany stand out for the extent of resources allocated to their MCOs. In 2016, Germany outlined a plan for a cyber command said to have 13,500 personnel.[44]

Estonia is renowned for its active cybersecurity policy. For a long time, the government did not seem interested in conducting offensive cyber operations.[45] In October 2018, however, the Estonian government announced that they had established a military cyber command.[46] We can expect several other newcomers in the near future. For example, according to the Belgian media, "the Belgian military forces are to get a new [cyber] component as from 2019".[47]

4. A CLOSER LOOK AT MCO DEVELOPMENT

The above section provided a general overview of NATO member states' paths towards conducting offensive cyber operations. The purpose of this section is to highlight several additional observations.

First, MCOs do not emerge in a political and organizational vacuum. Indeed, they are often established based on rebranding, restructuring, or combining existing institutions. This means that MCO development in theory (and how it is presented in official documents) and in practice do not always closely match.

Second, it is unclear if *any* MCO is at the *maturity* stage. USCYBERCOM is undoubtedly the main candidate. Whilst its organizational structure is no longer embryonic, it cannot be described as mature. As said, USCYBERCOM only recently became a unified combatant command and achieved full operational capability. It is

[43] COMCYBER & Ministère des Armées, "Éléments publics de doctrine militaire de lutte informatique offensive," (2019), retrieved from: https://www.defense.gouv.fr/content/download/551555/9394645/Eléments%20publics%20de%20doctrine%20militaire%20de%20lutte%20informatique%20OFFENSIVE.pdf
Also see: William Moray, "France bolsters cyber capabilities and commitment through new doctrine," *Jane's Intelligence Review* (2019, February 26).

[44] Germany has a strategic reconnaissance unit in the Department of Information and Computer Networks Operations since 2009; John Goetz, Marcel Rosenbach, and Alexander Szandar, "War of the future: national defense in cyberspace", Spiegel Online, (2009, February 11), retrieved from: http://www.spiegel.de/international/ germany/war-of-the-future-national-defense-in-cyberspace-a-606987.html; The Federal Government of Germany, "White Paper on German Security Policy and the Future of the Bundeswehr," (2016), retrieved from: http://www.new-york-un.diplo.de/contentblob/4847754/Daten/6718448/160713weibuchEN.pdf; Nina Werkhäuser, "German army launches new cyber command," DW, (2017, April 1) retrieved from: https://www.dw.com/en/german-army-launches-new-cyber-command/a-38246517.

[45] Also, the Estonian Cyber Defence Unit (volunteer group) does not conduct CNA.

[46] Its establishment was announced a year earlier.

[47] Michael Torfs, "Belgian army to get new component to tackle cyber crime," (2017, April 7), *Flanders News*.

also still trying to strategically navigate the threat landscape – striving to end its heavy reliance on the NSA and stand on its own two feet.[48] In other words, there is still progress to be made in aligning the Cyber Command's ends, ways and means.

Third, whilst some states have devoted substantial budgets to their MCOs (which might in part be due to the broader mission and functioning requirements of the organization), most aspiring NATO cyber powers still have a rather small budget at their disposal. According to the *Wall Street Journal*, the Danish government "allocates about $10 million a year for 'computer-network operations,' including defense and offense, since 2013".[49] Other media reports indicate that, in 2015, $75 million was allocated for offensive cyber capabilities through 2017. In 2014, Spain for the first time allocated a budget of €2.3 million to enhance its ability to conduct offensive cyber operations.[50] In the Netherlands, the initial budget was €50 million to establish the new cyber command, with an annual budget around €20 million for the following years.[51] Considering the size of these budgets, it is unclear if these MCOs could ever go past the initial *growth* stage.

Fourth, there is a widely-held notion that establishing an MCO and conducting offensive cyber operations is cheap or easy. This is not the case. As an MCO moves through the stages of the life cycle, it will have to address different problems. First, the determining factor of an MCO – at any stage of the life cycle – is, as one military commander put it, "people, people, people".[52] Second, MCOs also need to acquire (or develop) toolsets in order to gain, escalate and maintain access to targeted computer systems and networks.[53] Whilst much public attention is paid to states' stockpiling of zero-day exploits, known exploits (and social engineering techniques) are unlikely to be found gathering dust at the bottom of an MCO's toolbox – even for more well-established military organizations. Third, an MCO may have the best cyber force in the world, but it is bound to fail without strategic guidance and organizational coordination. One critical issue for an MCO is to ensure that offensive cyber operations can be deployed as an integral part of the overall mission. This means that organizational coordination across the life cycle is essential to ensure interoperability.[54] This may help to explain

[48] Sulmeyer, "Much Ado About Nothing?"

[49] Jennifer Valentino-Devries and Danny Yadron, "Cataloging the World's Cyberforces," *Wall Street Journal* (2015, October 11), retrieved from: http://graphics.wsj.com/world-catalogue-cyberwar-tools/.

[50] Brey, "National Cyber Security Strategy of Spain," (2013), p. 32.

[51] Max Smeets, "People, People, People: Vragen over het DDC en het inzetten van cyberactiviteiten," *Atlantische Commissie*, (2018, April), retrieved from: https://www.atlcom.nl/upload/AP_6_2018_Smeets.pdf; Also see: van Lonkhuyzen & Kees, "Het cyberleger kan en mag nog weinig".

[52] Senior Military Cyber Commander, "The Second International Cyber Symposium: Cyberspace and the Transformation of 21st Century Warfare," The Royal United Services Institute (RUSI) (Church House, Westminster: London), 19-20 October 2016.

[53] A common distinction made is between exploits and implants (tools).

[54] For a more detailed analysis of the organizational challenges related to integration see: Michael Sulmeyer, "Much Ado about Nothing? Cyber Command and the NSA," *War on the Rocks*, (2017, July 19), https://warontherocks.com/2017/07/much-ado-about-nothing-cyber-command-and-the-nsa/; Max Smeets, "Organisational Integration of Offensive Cyber Capabilities: A Primer on the Benefits and Risks," 2017 9th International Conference on Cyber Conflict: Defending the Core, H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.) (Tallinn: NATO CCD COE Publications: 2017).

why so many states are still only at the early stages of development; why reports have been published in a number of states about operational struggles; and why we have publicly observed CNA-activity by only a small group of NATO member states.

Fifth, this overview of organizational development across NATO Member States is only based on *publicly* available information. It is expected that there are more institutional developments hidden from the public eye. Several states recognize the cyber threat as a priority issue, but do not seem to promote the establishment of an MCO. For example, Lithuania considers cyberspace to be a new "environment of warfighting" and recognizes the cyber threat, yet there is no evidence to suggest that the government has established a program to conduct offensive cyber operations to achieve d4 effects. A similar discussion is provided in the 2015 Security Strategy of the Czech Republic and the cyber strategy of the Slovak Republic.[55] It would be hardly surprising if some of these states are in fact considering conducting military cyber operations.

5. CONCLUDING REMARKS

The aim of this paper was to provide a conceptual framework to facilitate analysis and comparison between different MCOs across NATO member states. The life cycle framework distinguishes between five stages of organizational development: i) seed, ii) startup, iii) growth, iv) expansion, and v) maturity.

It was shown that a large number of NATO Member States started to carefully consider establishing MCOs from 2008 onwards, and some had already started significant organizational efforts in the 1990s. However, I also reveal that the MCOs of most NATO members are still at the early stages of organizational development – and even those at the growth stage still have limited budgets to address the different workforce, capability, strategic and other requirements.

Future research can fruitfully expand this analysis on the MCO life cycle in a number of ways. As an MCO moves through the stages, a government faces different organizational challenges. I did not assess how different governments have sought to overcome these challenges. Also, it remains unclear to what degree governments can help each other in MCO development through international cooperation with other like-minded states – within or outside the NATO alliance. Also, more attention should

[55] National Security Authority, "National Cyber Security Strategy of the Czech Republic for the Period from 2015 to 2020," (2015), retrieved from, https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/CzechRepublic_Cyber_Security_Strategy.pdf; Peter Pellegrini and Robert Fico, "Cyber Security Concept of the Slovak Republic for 2015 - 2020," retrieved from: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/cyber-security-concept-of-the-slovak-republic-1; Also see: Tomáš Minárik, "National Cyber Security Organisation: Czech Republic," CCD COE Publications, 2nd version, (2016), retrieved from: https://ccdcoe.org/uploads/2018/10/CS_organisation_CZE_032016.pdf.

be paid to the benefits and limitations of bringing in private sector solutions. Finally, we could benefit from more case study research, process tracing organizational decisions and capturing other developments, and looking at countries' progress in more detail.

REFERENCES

Arts, Sophie "Offense as the New Defense: New Life for NATO's Cyber Policy," *The German Marshall Fund of the United States, Policy Brief*, 39, (2018):1-9.

Boutilier, Alex, "Canada developing arsenal of cyber-weapons," *The Star*, (2017, March 16), retrieved from: https://www.thestar.com/news/canada/2017/03/16/canada-developing-arsenal-of-cyber-weapons.html;

Brey, Rajoy, "National Cyber Security Strategy of Spain," (2013), retrieved from: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/NCSS_ESen.pdf

Chain, Henri, "Espionnage et cybersécurité, Bernard Barbier reçu par Symposium CentraleSupélec," (2016, September 5), retrieved from: https://www.youtube.com/watch?v=s8gCaySejr4

Churchill, Neil C., and Virginia L. Lewis, "The Five Stages of Small Business Growth," *Harvard Business Review*, (1983, May), retrieved from: https://hbr.org/1983/05/the-five-stages-of-small-business-growth

COMCYBER & Ministère des Armées, "Éléments publics de doctrine militaire de lutte informatique offensive," (2019), retrieved from: https://www.defense.gouv.fr/content/download/551555/9394645/Eléments%20publics%20de%20doctrine%20militaire%20de%20lutte%20informatique%20OFFENSIVE.pdf

Deibert, Ronald J.,"Black Code: Censorship, Surveillance, and the Militarisation of Cyberspace," *Millennium: Journal of International Studies* (2003).

Ducheine, Paul "Defensie in het digitale domein," *Militaire Spectator*, 186:4 (2017)152-168.

Dutch Ministry of Defense, Third International Cyber Operations Symposium, (2017, October);

Garamone, Jim, and Lisa Ferdinando, "DoD Initiates Process to Elevate U.S. Cyber Command to Unified Combatant Command," *Department of Defense News*, (2017, August 18), retrieved from: www.defense.gov/News/Article/Article/1283326/dod-initiates-process-to-elevate-us-cyber-command-to-unified-combatant-command/)

Goetz, John, Marcel Rosenbach, and Alexander Szandar, "War of the future: national defense in cyberspace", *Spiegel Online*, (2009, February 11), retrieved from: http://www.spiegel.de/international/ germany/war-of-the-future-national-defense-in-cyberspace-a-606987.html

Gramaglia, Matteo, Emmet Tuohy, Piret Pernik. "Military Cyber Defense Structures of NATO Members," *The Star*, (2016, January 9), retrieved from: https://www.thestar.com/news/canada/2016/09/01/former-electronic- spy-chief- urges-ottawa-to- prepare- for-cyber- war.html

Joint Task Force Global Network Operations, "A Legacy of Excellence: December 30, 1998- September 7, 2010", retrieved from: https://assets.documentcloud.org/documents/2849764/Document-05.pdf

Laudrain, Arthur, "France's New Offensive Cyber Doctrine," (2019, February 26), Lawfare, retrieved from: https://www.lawfareblog.com/frances-new-offensive-cyber-doctrine

Minárik, Tomáš , "National Cyber Security Organisation: Czech Republic," CCD COE Publications, 2nd version, (2016), retrieved from: https://ccdcoe.org/uploads/2018/10/CS_organisation_CZE_032016.pdf

Moray, William, “France bolsters cyber capabilities and commitment through new doctrine,” *Jane’s Intelligence Review* (2019, February 26).

National Security Authority, “National Cyber Security Strategy of the Czech Republic for the Period from 2015 to 2020,” (2015), retrieved from, https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/CzechRepublic_Cyber_Security_Strategy.pdf

NATO, “Brussels Summit Declaration, (2018, July 11-12), retrieved from: https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2018_07/20180713_180711-summit-declaration-eng.pdf

Nomikos, John M., “Intelligence Requirements for Cyber Defense, Critical Infrastructure and Energy Security in Greece,” *National Security and National Future*, 1-2:17 (2016).

Pellegrini, Peter and Robert Fico, “Cyber Security Concept of the Slovak Republic for 2015 – 2020,” retrieved from: https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/cyber-security-concept-of-the-slovak-republic-1

Pernik, Piret, “Preparing for Cyber Conflict: Case Studies of Cyber Command,” (2018, December), retrieved from: https://icds.ee/wp-content/uploads/2018/12/ICDS_Report_Preparing_for_Cyber_Conflict_Piret_Pernik_December_2018-1.pdf

Petch, Neil, “The Five Stages Of Your Business Lifecycle: Which Phase Are You In?” *Entrepreneur*, (2016, February 29), retrieved from: https://www.entrepreneur.com/article/271290

Pijnenburg Muller, Lilly, “Military Offensive Cyber-Capabilities: Small-State Perspectives, Norwegian Institute of International Affairs, Policy Brief, 1, (2019), retrieved from: https://brage.bibsys.no/xmlui/bitstream/handle/11250/2583385/NUPI_Policy_Brief_1_2019_Muller.pdf?sequence=1&isAllowed=y

Reeve, Tom, “France unveils cyber command in response to ‘new era in warfare’ ,” SC magazine, (2016, December 16), retrieved from: https://www.scmagazineuk.com/france-unveils-cyber-command-in-response-to-new-era-in-warfare/article/579671/

République Française “Strategic Review of Defense and National Security: Key Points,” (2017) retrieved from: https://www.defense.gouv.fr/content/download/514686/8664672/file/2017-RS-PointsClesEN.pdf

Sarközy, Nicolas, “The French White Paper on Defense and National Security”, (New York: Odile Jacob Publishing Corporation: 2008), retrieved from: http://www.mocr.army.cz/images/Bilakniha/ZSD/French%20White%20Paper%20on%20Defence%20and%20National%20Security%202008.pdf

Schulze, Matthias and Sven Herpig, “Germany Develops Offensive Cyber Capabilities Without a Coherent Strategy of What to Do With Them,” Council on Foreign Relations, (2018, December 3), retrieved from: https://www.cfr.org/blog/germany-develops-offensive-cyber-capabilities-without-coherent-strategy-what-do-them

Senior Military Cyber Commander, “The Second International Cyber Symposium: Cyberspace and the Transformation of 21st Century Warfare,” The Royal United Services Institute (RUSI) (Church House, Westminster: London), 19-20 October 2016.

Sheldon, John B., “NATO and Cyber Defense: Hanging Together or Hanging Separately?”, Presentation, (year unknown), hxxp://www.unidir.ch/files/conferences/pdfs/nato-and-cyber-defence-hanging-together-or-hanging-separatelyen-1-608.pdf

Smeets, Max, and Herbert Lin, “4 A Strategic Assessment of the U.S. Cyber Command Vision,” in Bytes, Bombs, and Spies: The Strategic Dimensions of Offensive Cyber Operations, (Washington DC: Brookings Institution Press: 2018), pp. 81-104.

Smeets, Max, “Going cyber : the dynamics of cyber proliferation and international security,” DPhil Dissertation in International Relations, University of Oxford, 2017.

Smeets, Max, “Organisational Integration of Offensive Cyber Capabilities: A Primer on the Benefits and Risks,” *2017 9th International Conference on Cyber Conflict: Defending the Core*, H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.) (Tallinn: NATO CCD COE Publications: 2017).

Smeets, Max, and Herbert Lin, “Offensive Cyber Capabilities: To What Ends?” *2018 10th International Conference on Cyber Conflict CyCon X: Maximising Effects*, T. Minárik, R. Jakschis, L. Lindström (Eds.) (Tallinn: NATO CCD COE Publications: 2018).

Smeets, Max, “The Strategic Promise of Offensive Cyber Operations,” *Strategic Studies Quarterly*, (2018, Fall):90-113.

Smeets, Max “People, People, People: Vragen over het DDC en het inzetten van cyberactiviteiten,” *Atlantische Commissie*, (2018, April), retrieved from: https://www.atlcom.nl/upload/AP_6_2018_Smeets.pdf

Sulmeyer, Michael, “Much Ado about Nothing? Cyber Command and the NSA,” *War on the Rocks*, (2017, July 19), https://warontherocks.com/2017/07/much-ado-about-nothing-cyber-command-and-the-nsa/

The Danish Ministry of Defense, “Danish Defense Agreement 2013-2017”, (2012, November 30), retrieved from: http://www.fmn.dk/eng/allabout/Documents/TheDanishDefenceAgrement2013-2017english-version.pdf

The Federal Government of Germany, “White Paper on German Security Policy and the Future of the Bundeswehr,” (2016), retrieved from: http://www.new-york-un.diplo.de/contentblob/4847754/Daten/6718448/160713 weibuchEN.pdf

Torfs, Michael, “Belgian army to get new component to tackle cyber crime,” (2017, April 7), *Flanders News*.

UK Government, “Britain’s cyber security bolstered by world-class strategy,” (2016, November 1), retrieved from: https://www.gov.uk/government/news/britains-cyber-security-bolstered-by-world-class-strategy

United Nations Institute for Disarmament Research (UNIDIR), “The Cyber Index International Security Trends and Realities,” (2013), retrieved from: http://www.unidir.org/files/publications/pdfs/cyber-index-2013-en-463.pdf

United States Strategic Command, “JFT-CND/JTC-CNO/JTF-GNO: A Legacy of Excellence” (1998, December 30/ 2010, September 7), retrieved from: https://nsarchive2.gwu.edu//dc.html?doc=2849764-Document-05)

US Department of Defense, “News Conference By Secretary Mattis at NATO Headquarters, Brussels, Belgium,” US Department of DEcen, (2018, October 4), retrieved from: dod.defense.gov

Valentino-Devries, Jennifer, and Danny Yadron, “Cataloging the World’s Cyberforces,” *Wall Street Journal* (2015, October 11), retrieved from: http://graphics.wsj.com/world-catalogue-cyberwar-tools/

Van Lonkhuyzen, Liza and Kees Versteegh, “Het cyberleger kan en mag nog weinig,” *NRC* (2018, December 18), retrieved from: https://www.nrc.nl/nieuws/2018/12/18/het-cyberleger-is-er-wel-maar-mag-weinig-a3099254

Warner, Michael, “Cybersecurity: A Pre-history”, Intelligence and National Security, *Intelligence and National Security*, 27:5 (2012)781-799.

Warner, Michael, “U.S. Cyber Command’s Road to Full Operational Capability,” in Stand Up and Fight: The Creation of U.S. Security Organizations, 1942–2005, edited by Ty Seidule and Jacqueline E. Whitt (Carlisle, Penn.: Strategic Studies Institute and U.S. Army War College Press, 2015), chap. 7.

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

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

What are Military Cyberspace Operations Other Than War?

Brad Bigelow[1]
Principal Technical Advisor
Deputy Chief of Staff Cyberspace
SHAPE
Mons, Belgium
brad.bigelow@shape.nato.int

Abstract: NATO has recognized cyberspace as a domain of military operations, with the Cyberspace Operations Centre as the focal point for coordinating and directing effects in cyberspace in the context of Alliance operations and missions. Yet many of the threats nations face in cyberspace deliver their effects below the level of conventional armed conflict, involve systems and capabilities outside the span of military control, and do not lend themselves to traditional military response options. As concerns over the defense of critical national infrastructures and other non-military targets such as election systems and social media increase, however, many are calling for the military to take on a greater role in cyberspace outside the context of armed conflicts. This paper looks at calls for greater military involvement in cyberspace below the level of conventional armed conflict, in the context of previous doctrinal work on military operations other than war. It attempts to derive a set of equivalent principles that could be applied to military cyberspace operations performed below the level of armed conflict; it then assesses these functions in terms of whether the military should take a leading or supporting role, and what kinds of tasks, relationships, and authorities might be involved. The aims of this paper are to identify the appropriate roles for the military in cyberspace operations below the level of conflict and to highlight the importance of cross-functional coordination with civil authorities in performing these roles.

Keywords: *Cyberspace, Cyberspace Operations, Military Operations Other Than War*

[1] The views and opinions expressed in this article are those of the author alone and do not necessarily reflect those of NATO.

1. INTRODUCTION

Cyberspace is now broadly recognized as an essential element of national security. As a consequence, many nations are developing the role the military plays as an instrument of national defense. And in the case of the North Atlantic Treaty Organisation, cyberspace has been recognized as an instrument of collective defense, a domain of military operations "… in which NATO must defend itself as effectively as it does in the air, on land, and at sea" (NATO, 2016).

Much of the effort involved in developing military capabilities in cyberspace is focused on those aspects mentioned in the Warsaw Summit declaration quoted above: the "ability to protect and conduct operations across these domains" and to integrate these capabilities "into operational planning and Alliance operations and missions" (NATO, 2016). This is, in part, analogous to the recognition of airspace as a domain of military operations and the development of military air power capabilities that began in the early 20th century (Bigelow, 2002). For many nations, including the members of NATO, there has also been an explicit commitment to the employment of such capabilities in compliance with *jus in bello*, the law of armed conflict or the law of war.

Traditionally, much of military doctrine has focused on large-scale, sustained combat operations aimed at achieving national objectives or protecting national interests. Yet many of the threats that nations face in cyberspace deliver their effects below the level of conventional armed conflict, affect systems and capabilities outside the span of military control, and do not lend themselves to military response options involving combat operations. As concerns increase over the defense of critical national infrastructures and other non-military targets such as election systems and social media, many are calling for the military to take on a greater role in cyberspace outside the context of armed conflicts.

These problems are less related to large-scale combat operations than they are to what U.S. military doctrine once referred to as "Military Operations Other than War" (MOOTW): "deterring war, resolving conflict, promoting peace, and supporting civil authorities in response to domestic crises" (Joint Chiefs of Staff, 1995). Although this term is no longer used in U.S. doctrine, the concept of military operations other than war offers a useful framework within which the development of military cyberspace capabilities can be assessed.

This paper looks at calls for greater military involvement in cyberspace below the level of armed conflict in the context of previous doctrinal work on military operations other than war, including civil-military cooperation, peace support operations, and special operations. It attempts to derive a set of equivalent principles for military cyberspace operations performed below the level of armed conflict in physical domains. It then assesses these functions in terms of whether the military should take a leading or supporting role and what kinds of tasks, relationships, and authorities might be involved. The aims are to identify the appropriate roles for the military in cyberspace operations below the level of conflict and to highlight the importance of cross-functional coordination with civil authorities in performing these roles.

2. CALLS FOR A GREATER MILITARY ROLE

The security challenges now being seen in cyberspace have two fundamental and very different consequences for those implementing cyberspace as a domain of military operations. One is that of establishing cyberspace effectively as an operational domain in the context of what one might call traditional military combat operations and missions—situations in which an Area of Responsibility is defined, forces assigned, objectives set and Rules of Engagement provided, together enabling a military commander to achieve Alliance objectives while complying with the Laws of Armed Conflict. The second consequence, however, is the much more difficult problem of defining the military role in cyberspace outside this context: in other words, the nature of military cyberspace operations other than war.

Some have argued that military operations in cyberspace outside the context of armed conflict should be limited to the protection of military networks and information systems. Miriam Dunn Cavelty has flatly stated that "Militaries cannot defend the cyberspace of their country – it is no space where troops and tanks can be deployed because the logic of national boundaries does not apply" (Dunn Cavelty, 2012). Stephen J. Anderson agrees, writing that traditional concepts of national defense cannot be applied in cyberspace: "The US Navy defends the littoral territorial boundaries; air defenses, either through missile defense initiatives or alert aircraft, define airspace boundaries. Those lines are not readily identifiable in cyberspace" (Anderson, 2016). Some go even further, arguing that an active military role in peacetime cyber security undermines investment in alternative mechanisms. In a 2013 post for the Lowy Institute, Ian Wallace wrote that such efforts disincentivized "other longer-term and more sustainable efforts to address the new challenges that cyber brings to security systems" (Wallace, 2013).

Yet this debate has evolved significantly in recent years, in large part thanks to increasing evidence of state-sponsored attacks on civilian cyberspace infrastructure. In a recent paper entitled *Rethinking Cyber Security*, James Lewis has stated that "The primary source of risk in cybersecurity comes from conflict between states"

(Lewis, 2018). This assessment is echoed by the Netherlands' National Cyber Security Centrum, which concluded in its 2018 assessment that "The most significant threats are sabotage and disruption by nation-states" (National Cyber Security Centrum, 2018). As consensus on the state actor threat in cyberspace has grown, so have calls for the military to take a more active role in the defense of cyberspace.

In the 2017 U.S. Senate deliberations on increasing the Secretary of Defense's authority to conduct clandestine military cyberspace operations, Senator John McCain asserted that the need for a strong military role in peacetime was self-explanatory: "It's the Department of Defense's job to defend this nation: that's why it's called the Department of Defense" (Pomerleau, 2017). This more active role— sometimes referred to as defending forward—is reflected in recent updates to military cyber strategies. The *2018 U.S. Defense Department Cyber Strategy*, for example, states explicitly: "We are engaged in a long-term strategic competition with China and Russia" and declares that this requires (and justifies) "action in cyberspace during day-to-day competition to preserve U.S. military advantages and to defend U.S. interests" (U.S. Department of Defense, 2018). Similarly, the Netherlands' *Defence Cyber Strategy 2018*, subtitled *Investing in cyber striking power for the Netherlands*, concludes that the current security environment demonstrates that "a more active contribution from Defence within the existing structures is required" (Netherlands Ministry of Defence, 2018). Jan Kallberg and Thomas S. Cook have gone even further, stating that nations should be prepared not only to use military cyberspace forces in peacetime but to actively foster these capabilities as an alternative to armed conflict: "Cyber is no longer a mere enabler of joint operations, but instead a viable strategic option for confronting adversarial societies" (Kallberg & Cook, 2017).

3. MILITARY OPERATIONS OTHER THAN WAR: DOCTRINE

It is useful to consider these calls for a more active military role in cyberspace outside of war in the context of doctrinal work on the role of military operations other than war in general. Although early discussion of the use of military force outside large-scale conflicts stems from counterinsurgency operations and the use of Special Forces in the early days of the Vietnam conflict, the term "Military Operations Other Than War" first appeared in U.S. military training publications in the early 1980s and was formally incorporated into U.S. doctrine in 1995 with *Joint Publication 3-07, Joint Doctrine for Military Operations Other Than War* (now deleted from the official library of U.S. joint military doctrine).

JP 3-07 divided military operations into two categories: combat and non-combat, the latter constituting military operations other than war. It identified fifteen types of non-combat operations, ranging from arms control and combatting terrorism to providing support to civil authorities and humanitarian assistance, and divided these operations into two categories based on whether the operation involved the use or threat of military force. In operations involving the use or threat of force, *jus ad bellum*, the international law governing use of force as an instrument of national policy, would apply. According to *JP 3-07*, in such operations, "force or threat of its use may be required to demonstrate U.S. resolve and capability, support the other instruments of national power, or terminate the situation on favorable terms" (Joint Chiefs of Staff, 1995).

In operations not involving the use of force, the military is often acting in support of, or in close coordination with, a civilian authority—for example, in response to a natural disaster or humanitarian crisis. Even operations such as a show of force or blockades are carried out in a larger context of diplomatic objectives. In support of disaster relief or a humanitarian crisis, the military's role involves providing the organic capabilities that it maintains for the primary purpose of supporting combat operations. Army field hospitals and kitchens, for example, can provide care and comfort to civilian populations injured and displaced by a hurricane, and Navy and Air Force sealift and airlift capabilities can deliver heavy equipment to locations devastated by an earthquake. However, the military can also take the lead, as in providing capacity-building support to the military forces of another nation. As *JP 3-07* notes, such peacetime uses of military forces "helps keep the day-to-day tensions between nations below the threshold of armed conflict or war and maintains U.S. influence in foreign lands". At the time when *JP 3-07* was written, it was assumed that such operations were "usually, but not always, conducted outside of the United States" (Joint Chiefs of Staff, 1995).

In hindsight, *JP 3-07* can be seen to suffer from covering too broad a spectrum of operations. Differences in legal authorities rooted in U.S. federal laws made military operations conducted on U.S. territory in support of civil authorities very different from, for example, humanitarian assistance operations conducted in support of the Department of State outside the U.S. Similarly, arms control operations, which are normally conducted overtly and under the conditions of treaties or other international agreements, are fundamentally different from "strikes and raids", which have usually involved the use of special operations forces working through covert means under Presidential authority in the U.S. and are termed "clandestine traditional military activities".

To better address the range of military operations other than war, the U.S. has replaced the 1995 *JP 3-07* with a number of discrete doctrine publications. Activities such as peace operations, which some nations such as the United Kingdom and Australia refer to as peace support operations, are now covered by *JP 3-07, Stability* (2016). *JP 3-24* (2018) covers counterinsurgency, *JP 3-26* (2014) counterterrorism, and *JP 3-28* (2018) support to civil authorities. These clarifications greatly aid in the application of doctrinal principles to real-world problems.

For the purposes of this paper, however, the most important lesson to be drawn from *JP 3-07* is that it may no longer be useful, for cyberspace operations doctrine at least, to draw a line between military operations in war and those "other than war". This seems to be particularly true for military operations in the cyberspace domain. Michael Sulmeyer echoes the sentiment of many commentators when he states, "Today's fight in cyberspace occurs in the gray zone between war and peace" (Sulmeyer, 2018). Indeed, argues Michael Fischerkeller, an offensive military cyberspace capability "would offer many opportunities, both when used on its own and in combination with other military capabilities, to influence an adversary's decision making in pre-crisis and crisis environments" (Fischerkeller, 2017). The more important distinction, particularly when it comes to military cyberspace capabilities, is whether or not a military operation involves the use or threat of force.

To illustrate, consider the latest update of U.S. Department of Defense (DOD) doctrine on cyberspace, *JP 3-12, Cyberspace Operations*, issued in June 2018. *JP 3-12* states that there are three cyberspace missions: operations of DOD networks (DODIN Ops); Defensive Cyberspace Operations (DCO); and Offensive Cyberspace Operations (OCO). It further divides DCO into three categories: Internal Defensive Measures (DCO-IDM), "where authorized defense actions occur within the defended network or portion of cyberspace"; Response Actions (DCO-RA), "where actions are taken external to the defended network or portion of cyberspace without the permission of the owner of the affected system"; and Defense of Non-DOD Cyberspace, in which the military carries out DCO-IDM and DCO-RA missions on "any U.S. or other blue cyberspace when ordered" (Joint Chiefs of Staff, 2018).

If one accepts the premise that the most important distinction between military operations is whether they involve the threat or use of force, however, *JP 3-12* adds, rather than reduces, confusion. It is hard to understand how DCO-RA actions taken external to the defended network and without the permission of the owner of the affected system do not constitute the use of force in cyberspace. Furthermore, the explanation of the Defense of Non-DOD Cyberspace is contradictory: if, by definition, Defense of Non-DOD Cyberspace missions are carried out in "blue"—friendly, willing, cooperative—cyberspace, then they will not include actions taken external to these networks.

This confusion mirrors discussions of the concept of "active defense," which is the term most often used outside the U.S. military for DCO-RA. Scott Berinato has written, "As active defense tactics gain popularity, the term's definition and tenets have become a muddy mess. Most notably, active defense has been conflated with 'hacking back'—attacking your attackers" (Berinato, 2018). Others state that active defense measures fall into two categories: "those that have effects on systems or networks inside the organizational span of control of the defender and those that have effects on systems or networks outside that span of control"—leaving it unclear whether "outside that span of control" includes systems owned by unwilling system owners (Kehler, Lin & Sulmeyer, 2017). Former U.S. Air Force cyberspace operator Robert M. Lee, on the other hand, defines active defense as "the process of security personnel taking an active and involved role in identifying and countering threats to the system," and attributes association of the term with "hacking back" to "poor translations of active defense theory in military strategies into the field of cyber security" (Lee, 2015).

Elsewhere in *JP 3-12*, however, one can see that DCO-RA and OCO tasks are, in fact, carried out by different forces from DCO-IDM and DODIN Ops tasks. (For the sake of this discussion, DODIN Ops will hereafter be referred to as Defense network ops). DCO-IDM tasks are performed by Cyber Protection Forces, teams "organized, trained, and equipped to defend assigned cyberspace in coordination with and in support of segment owners, cybersecurity service providers (CSSPs), and users." DCO-RA and OCO tasks, on the other hand, are carried out by National Mission Teams or, when supporting a Joint Force commander, Combat Mission Teams (Joint Chiefs of Staff, 2018). These teams, in other words, exist to operate in external networks and without the permission of the owner of the affected system.

Military cyberspace forces intending to apply force or the threat of force against adversary systems must work very closely, if not side-by-side, with the elements authorized to collect intelligence and conduct reconnaissance and surveillance of these adversaries. This intelligence is essential to support the development and testing of cyberspace weapons, techniques, and tactics, to support targeting and intelligence gain/loss assessment, and, in most cases, to gain access to the systems they intend to affect. According to Sergei Boeke and Dennis Broeders: "Cyber operations are tailor-made combinations of intelligence, intrusion, and attack, and it is seldom clear where one phase ends and another begins" (Boeke & Broeders, 2018). These forces must not only develop in-depth understanding of the technical details of targeting systems but some understanding of how the adversary uses these systems in day-to-day business or operations. This typically also requires these forces to be capable of conducting covert operations and their personnel to hold special security clearances.

Contrast these constraints with the forces and personnel engaged in Defense network ops or DCO, which do not involve the use or threat of force. Here, there is far less of a dependence upon intelligence (and essentially none when it comes to knowledge of intelligence means and sources). U.S. Cyber Command, for example, distinguishes between securing systems, which it considers "threat agnostic," protecting systems, which is "threat specific but passive," and defending systems, "a threat and capability-focused activity designed to counter adversary strategy and capability" (U.S. Cyber Command, 2018). Likewise, while attribution of cyber-attacks is of critical importance in guiding decisions to apply offensive cyberspace capabilities in a pre-emptive or reactive manner, attribution is far less important in the majority of decisions involved in DODIN Ops or DCO tasks.

To accurately identify the appropriate roles for the military in cyberspace operations other than war, therefore, perhaps the most important distinction to be made is between military cyberspace operations that involve the use or threat of force in cyberspace and those that do not, particularly in the context of operations below the level of conventional conflicts. This can be demonstrated by contrasting the characteristics and considerations of these two different efforts.

4. MILITARY CYBERSPACE OPERATIONS INVOLVING THE USE OR THREAT OF FORCE BELOW THE LEVEL OF CONFLICT

In recent testimony before the U.S. Senate Armed Services Committee, Michael Sulmeyer proposed "two necessary conditions of posture" for U.S. military cyber mission forces to be better prepared to defend the U.S. against foreign attempts to interfere with elections. First, "Our cyber mission forces should be constantly conducting reconnaissance missions abroad to discover election-related threats to the United States and provide indicators and warnings to our forces and decision-makers." Second, "Our cyber mission forces must be sufficiently ready to strike against targets abroad identified by reconnaissance as threats to our election" (Sulmeyer, 2018).

Although Sulmeyer's proposal was in the specific context of reactions to Russian meddling in U.S. elections in 2016, at a more general level these two conditions apply to any application of military OCO capabilities: first, they are highly dependent upon sustained reconnaissance of potential adversaries and their systems; and second, they need to be maintained at a high level of readiness because there may be little or no warning before they need to be engaged. If a nation intends to use offensive cyberspace capabilities to precede or pre-empt kinetic operations, then operational preparation of the cyber battlefield must become "as routine as reconnaissance or surveillance of potential adversary activity" (Kehler, Lin & Sulmeyer, 2017). What does "operational preparation of the cyber battlefield" involve? Robert Chesney spells it out clearly in his analysis of the *2018 DOD Cyber Strategy*: "Intrusions into the systems of potential adversaries in order to secure access of a kind that can be exploited for disruptive or destructive effect if and when the need later arises" (Chesney, 2018).

One can also argue that military OCO requires the same framework of command and control, rules of engagement, weapons release control, and damage assessment processes whether employed below the level of conflict or not. When *JP 3-12* states that "Clearly established command relationships are crucial for ensuring timely and effective employment of forces" in cyberspace operations, it does not stipulate at what level of conflict these forces are engaged (Joint Chiefs of Staff, 2018). If, as James Lewis has written, "The implicit threshold governing cyberattack is the line between force and coercion", then this line must apply to both those authorizing the attack and those affected by it (Lewis, 2018) This is why, as C. Robert Kehler and colleagues have written, standing rules of engagement for military cyberspace operations need to be in place to inhibit the unintended escalation of conflict (Kehler, Lin & Sulmeyer, 2017).

Recognizing the unique role of the military in conducting OCO—whether below the level of conflict or not—would also improve the ability of a nation to plan and organize how it deals with deterrence in cyberspace. Alex Wilner has written that the U.S. continues to struggle to understand which government agency or department is expected to engage in cyber deterrence: "To date, the division of labor remains uncertain" (Wilner, 2017). Of course, while some argue that a ready military OCO capability is essential to ensuring deterrence in cyberspace, others have suggested that deterrence in cyberspace is an impossible goal. But one good reason to clearly establish the unique military role of such a capability is to counter attempts to create OCO capabilities in the private sector. As Peter Singer testified before the U.S. House of Representatives in 2017, allowing companies to engage in OCO "is a very bad idea. It's is a bad idea for the same reason that vigilantism in general is a bad idea." Singer pointed out that such activities could raise significant risks at the international level because other nations could mistake private attempts to attack their systems for state-sponsored actions (US House of Representatives, 2017).

Establishing a military capability to conduct OCO below the level of conflict may be one key to realizing the unique benefits of cyberspace as an operational domain. Gregory Rattray and Jason Healey have argued that: "It may be that the future of cyber conflict is not equivalent to larger, theatre-level warfare but only to select covert attacks which could range across a wide set of goals and targets." In part, this argument draws upon the substantial base of experience showing that offensive operations between nations using conventional forces are relatively rare and usually condemned by other states (Rattray & Healey, 2010). But conventional offensive operations are also quite visible, are easy to attribute, and raise higher risks of escalation, which is why they have traditionally been seen as "a last resort and a temporary state" (Maurer, 2012).

OCO below the level of conflict, on the other hand, demonstrates the potential for states to exploit "grey zones"—areas where "international law principles and rules that are poorly demarcated or are subject to competing interpretations" (Schmitt, 2017). The willingness to operate in this "grey zone" is clearly demonstrated in the *2018 DOD Cyber Strategy*, which states that in the U.S. "the Department seeks to pre-empt, defeat, or deter malicious cyber activity targeting U.S. critical infrastructure that could cause a significant cyber incident regardless of whether that incident would impact DoD's warfighting readiness or capability." In the United Kingdom, Defence Minister Sir Michael Fallon called for "new doctrine to clarify our response within NATO to anonymous cyber activity which often takes place now in that grey zone below the previously understood threshold of war" (Fallon, 2017). A similar appetite is demonstrated in the Netherlands' *Defence Cyber Strategy 2018*, which states an intent to focus Defence support for civil authorities "on the vital infrastructure through closer collaboration with the responsible security partners" such as the National Cyber Security Centre (NCSC) (Netherlands Ministry of Defence, 2018). And in Germany, Defense Minister Ursula von der Leyen has stated that the Bundeswehr's cybersecurity forces are permitted to "offensively defend" their networks if attacked (Somaskanda, 2018).

NATO heads of state and governments have also recognized the value in leaving some amount of "greyness" in the "grey zone," as Jonatan Vseviov, the permanent secretary of the Estonian Ministry of Defence, explained in an interview: "there is a good level of what I would call 'constructive ambiguity' built into the wording of the Washington Treaty and also Article 5…. We don't want to give anybody a list of attacks that would trigger Article 5 because that would obviously mean that we automatically also create a list of potential attacks that would not trigger Article 5" (Mehta, 2018). The willingness of nations to consider use of OCO capabilities below the level of conflict is also a recognition that, as Michele Flournoy and Michael Sulmeyer have written, "for all the increasingly vehement warnings about a cyber Pearl Harbor, states have shown little appetite for using cyberattacks for large-scale destruction. The immediate threat is more corrosive than explosive" (Flournoy & Sulmeyer, 2018). All of which suggests that OCO can fulfil the vision proposed by Bernard Brodie at the dawn of the nuclear age: "Thus far the chief purpose of our military establishment has been to win wars. From now on its chief purpose must be to avert them" (Brodie, 1946).

From a doctrinal standpoint, however, the importance of recognizing OCO as a type of military operation that can be carried out not only in "war"—large-scale armed conflicts—but below the level of crisis, in the context of *jus ad bellum*, is that such capabilities cannot be employed in any context unless they are ready at the time of need. For conventional forces to be ready to act on short notice, they have to exist. They have to be equipped, armed, trained, sustained, able to move, informed about their potential adversaries, positioned to able to engage within their required readiness timelines—even though they may never need to move past that point of readiness and actually engage in battle. The same is true for cyberspace forces.

5. MILITARY CYBERSPACE OPERATIONS NOT INVOLVING THE USE OR THREAT OF FORCE

Readiness is just as critical for Defense network ops and DCO, if far less controversial. Today's militaries depend upon myriad networks, information systems, and communications transmission systems operating at different levels of classification and involving a wide variety of static, deployable, strategic, operational, tactical, and commercial systems and services. They also depend to a greater or lesser extent on the "littorals" of cyberspaces—the places where cyberspaces meet other environments, including physical infrastructure such as fences, buildings, gates, and transportation networks, the radio frequency spectrum, and critical infrastructures such as electrical power and water supplies (Withers, 2015). Many of these systems must be in constant operation to support standing tasks as well as to meet their readiness requirements, and consequently, must be protected against threats to their availability, confidentiality, and integrity.

This level of readiness raises the possibility that some of these capabilities can be employed below the level of conflict in support of some of the types of non-combat operations identified in *JP 3-07*, such as providing support to civil authorities and humanitarian assistance. In the case of a natural disaster, combat deployable communications and information systems could be used to restore or augment critical civil communications capabilities while the damaged infrastructure is being repaired. The U.S. Defense Information Systems Agency, for example, put its Transnational Information Sharing Cooperation network, which was still in preparation, into live operation in January 2010 to support U.S. Southern Command efforts to coordinate relief operations following a devastating earthquake in Haiti (Chossudovsky, 2010).

Effectively employing these capabilities in support of civil authorities, however, remains a relatively immature aspect of military cyberspace operations. For one thing, when the support takes place within the nation's borders, there can be complex legal and regulatory constraints, which stem in part from the aim of maintaining civil control over military affairs. This is illustrated by the use of the terms "secure" and “defend” in distinguishing whether the DOD or the Department of Homeland Security (DHS) is the lead agency. *JP-3-28, Defense Support of Civil Authorities*, states that the DOD “is the lead agency for homeland defense,” while *JP 3-12, Cyberspace Operations*, states that the DHS is the lead agency for homeland security, including the responsibility to “safeguard and secure cyberspace” (Joint Chiefs of Staff, 2013), (Joint Chiefs of Staff, 2018).

In addition, while there is general agreement that the military should play some role in responding to cyber incidents with national-level impacts, the precise nature of this role, what responsibilities and authorities are required to perform it, and how it relates to the roles performed by civil authorities are still unclear. In some nations, even the statutory foundation for such cooperation is lacking. Piret Pernik found that Finnish Defence Forces had not been assigned any responsibility to support civil authorities in the event of a “cyber emergency” (Pernik, 2018). A 2013 assessment by the U.K. House of Commons suggested that the role was similar to that associated with other military capabilities such as medical and logistical resources: in the event of a large-scale cyberattack, the military could be drawn upon to provide “additional staff, planning resources or technical expertise” (House of Commons Defence Committee, 2012). *JP 3-12* notes that the military may be called upon to perform DCO in support of civil authorities, but a 2016 study by the U.S. Government Accountability Office (GAO) found that the DOD’s basic doctrine publication on defense support of civil authorities (DSCA), *JP 3-28*, “does not provide specific details on how DOD will provide cyber support to civil authorities” (U.S. Government Accountability Office, 2016). A subsequent GAO report published in 2017 found that the DOD had not yet developed a plan for “collective training activities that are integrated with exercises conducted with other agencies and state and local governments” (U.S. Government Accountability Office, 2017).

Nations attempting to develop the military role in the defense of non-military domestic networks are running into “grey zone” challenges of their own. Although protection of critical infrastructures against cyber-attacks has been a topic at the national policy level since President Clinton established the President’s Commission on Critical Infrastructure Protection in 1996, views on the appropriate role for the military to play remain divided. Some argue that any such involvement would represent a militarization of cyberspace as a whole. Others suggest the role is limited to that of offering OCO as a response option. Alex Wilner, for example, has written “It is not clear, however, if Cyber Command has a role to play in protecting both military and civilian cyber infrastructure. It may chiefly respond to attacks on the former, despite the fact that civilian cyber infrastructure appears far more vulnerable than military infrastructure to cyber-attack” (Wilner, 2017).

There is some merit to this argument. The development of military cyberspace capabilities has, from the very beginning, suffered from the inappropriate use of analogies from conventional domains. The military can, for example, protect a power plant from ground and air attack by positioning land and ground-based air defense troops around it. In neither case is the military defense taking an active role in the operation of the infrastructures they are supporting. A military cyber defense unit positioned to protect the networks and information systems of the power plant, on the other hand, would be challenged *not* to interfere with the plant's operation. "The private sector knows its own systems better," Peter Singer has argued, "so it is going to be the one best equipped to defend itself, set aside all of the other kind of appropriate questions." Singer put the situation in well-recognized military terms: "I think the private sector should be the supported command, not the supporting command" (U.S. House of Representatives, 2017).

A number of nations are now building new mechanisms to enable the military to play an effective supporting role in the defense of critical national infrastructures against cyberattack. Estonia has established a volunteer Cyber Defence Unit of the Estonian Defence League (CDU), which can be deployed to assist civilian authorities with cyber security challenges in both crises and routine operations. Monica Ruiz proposes a similar approach for the U.S.: "state-level volunteer units … [for] the protection of critical U.S. infrastructure." These units would focus on "[i]mproving general readiness through trainings, exercises, and strengthening cooperation and synergy between public and private sectors through information sharing" and on providing support—particularly technical and analytical—in the event of major cyber incidents (Ruiz, 2018). Germany has launched a program of regular information exchange and job visits of members of its new Bundeswehr cyber service and Deutsche Telekom employees (Knirsch, 2018). Nina Kollars has suggested the need for the military to reach beyond established civil and commercial cyber defense organizations and establish better links with the "white hat" or ethical hacker community: "the work of the white hat defender community is largely unrecognized in the discourse surrounding national security and cyber strategy" (Kollars, 2018).

It is not surprising that nations are struggling with the military role in critical infrastructure defense. This is still very much work in progress. In the *John S. McCain National Defense Authorization Act for Fiscal Year 2019*, the U.S. Senate approved establishment of a "Cyberspace Solarium Commission" charged to "develop a consensus on a strategic approach to protecting the crucial advantages of the United States in cyberspace against the attempts of adversaries to erode such advantages." One particular task of the commission was to weigh "the options for defending the United States, to consider possible structures and authorities that need to be established, revised, or augmented within the Federal Government" (U.S. Senate, 2018). Michele

Flournoy and Michael Sulmeyer have already proposed a possible structure: “a new cyberdefense agency whose purpose would be not to share information or build criminal cases but to help agencies, companies, and communities prevent attacks” (Flournoy & Sulmeyer, 2018). The discussions demonstrate Jan Kallberg and Thomas S. Cook’s argument that “cyber as an area of conflict will require unorthodox approaches, innovation, and an ability to look beyond how we are used to organize defenses” (Kallberg & Cook, 2017).

6. CONCLUSION

Joint Publication 3-07, Joint Doctrine for Military Operations Other Than War was, in its time, an attempt to define the military’s role in a variety of unconventional situations. It was useful in moving the military mindset—in the U.S., at least—away from the view that fighting wars on a large scale was not only the military’s ultimate purpose but also its only proper role. The development of military cyberspace capabilities, however, has progressively revealed the need to move beyond thinking of military roles in the simplistic terms of “war” and “other than war” and to focus instead on the appropriate role for the military’s defensive and offensive cyberspace capabilities across a variety of situations, ranging from supporting civil authorities in disaster relief to responding to threats against critical infrastructure or the security of elections.

On the one hand, while the appropriate scenarios for nations to employ offensive cyberspace capabilities continue to be debated, the development of these capabilities cannot be deferred until there is an immediate need. Instead, like any conventional military capability, they need to be organized, equipped, trained, and sustained at a high level of readiness—and supported as necessary through intelligence preparation of potential cyberspace battlefields. On the other hand, it will be difficult to organize, train, and equip military cyber defenders to lead or support the defense of civil and commercial networks and information systems until the nation can decide on the appropriate structures by which to bring together military, intelligence, diplomatic, law enforcement, governmental, and commercial resources. In the meantime, however, *JP 3-07* still offers some value in reminding us that the primary role for the military in peacetime is to help “keep the day-to-day tensions between nations below the threshold of armed conflict or war” (Joint Chiefs of Staff, 1995).

REFERENCES

Anderson, S. J. (2016). *Airpower Lessons for an Air Force Cyber-Power Targeting Theory (Drew Paper No. 23)*. Maxwell Air Force Base, AL: Air University Press.

Berinato, S. (2018, May 21). *Active Defense and 'Hacking Back': A Primer.* Retrieved from *Harvard Business Review*: https://hbr.org/2018/05/active-defense-and-hacking-back-a-primer

Bigelow, B. (2002). Forces, Targets, and Effects: Militarising Information Warfare. *Journal of Information Warfare*, 2(1), 15-22.

Boeke, S., & Broeders, D. (2018). The Demilitarisation of Cyber Conflict. *Survival: Global Politics and Strategy*, 60(6), 73-90.

Brodie, B. (1946). "The Development of Nuclear Strategy". In B. Brodie, *The Absolute Weapon* (p. 76). New York: Harcourt Brace.

Chesney, R. (2018, September 25). *The 2018 DOD Cyber Strategy: Understanding 'Defense Forward' in Light of the NDAA and PPD-20 Changes*. Retrieved from Lawfare Blog: https://www.lawfareblog.com/2018-dod-cyber-strategy-understanding-defense-forward-light-ndaa-and-ppd-20-changes

Chossudovsky, M. (2010, January 21). *A Haiti Disaster Relief Scenario Tested by US Military One Day Before the Earthquake*. Retrieved from Global Research: https://www.globalresearch.ca/a-haiti-disaster-relief-scenario-was-envisaged-by-the-us-military-one-day-before-the-earthquake/17122

Dunn Cavelty, M. (2012). The Militarisation of Cyberspace: Why Less May Be Better. In C. Czosseck, R. Ottis, & K. Ziolkowski (Eds.), *2012 4th International Conference on Cyber Conflict* (pp. 141-153). Tallinn: NATO C.

Fallon, M. (2017, June 27). *Defence Secretary's speech at Cyber 2017 Chatham House Conference*. Retrieved from Gov.uk: https://www.gov.uk/government/speeches/defence-secretarys-speech-at-cyber-2017-chatham-house-conference

Fischerkeller, M. (2017). Incorporating Offensive Cyber Operations into Conventional Deterrence Strategies. *Survival: Global Politics and Strategy*, 59(1), 103-134.

Flournoy, M., & Sulmeyer, M. (2018, September/October 2018). Battlefield Internet: A Plan for Securing Cyberspace. *Foreign Affairs*, *97*(5), pp. 40-46.

House of Commons Defence Committee. (2012). *Defence and Cyber-Security: Sixth Report of Session 2012-13, Volume I*. London: The Stationery Office Limited.

Joint Chiefs of Staff. (1995). *Joint Publication 3-07, Joint Doctrine for Military Operations Other Than War*. U.S. Department of Defense.

Joint Chiefs of Staff. (2013). *Joint Publication 3-28, Defense Support of Civil Authorities*. U.S. Department of Defense.

Joint Chiefs of Staff. (2018). *Joint Publication 3-12, Cyberspace Operations*. U.S. Department of Defense.

Joint Chiefs of Staff. (2018). *Joint Publication 3-27, Homeland Defense*. U.S. Department of Defense.

Kallberg, J., & Cook, T. S. (2017). The Unfitness of Traditional Military Thinking in Cyber. *IEEE Access, 5*, 8126-8130.

Kehler, C. R., Lin, H., & Sulmeyer, M. (2017). Rules of Engagement for Cyberspace Operations: a View from the USA. *Journal of Cyber Security, 3*(1), 69-80.

Knirsch, R. (2018, September 25). *Deutsche Telekom and Bundeswehr (German Armed Forces) cooperate in cyber defense*. Retrieved from Deutsche Telekom: https://www.telekom.com/en/media/media-information/archive/dt-and-bundeswehr-cooperate-in-cyber-defense-542510

Kollars, N. (2018, September 6). *Beyond the Cyber Leviathan: White Hats and U.S. Cyber Defense*. Retrieved from War on the Rocks: https://warontherocks.com/2018/09/beyond-the-cyber-leviathan-white-hats-and-u-s-cyber-defense/

Lee, R. M. (2015, February 25). *The active cyber defense cycle*. Retrieved from Control Engineering: https://www.controleng.com/articles/the-active-cyber-defense-cycle-a-strategy-to-ensure-oil-and-gas-infrastructure-cyber-security/

Lewis, J. A. (2018). *Rethinking Cybersecurity*. Center for Strategic and International Studies.

Maurer, T. (2012, December 5). *Is it Legal for the Military to Patrol American Networks?* Retrieved from Foreign Policy: https://foreignpolicy.com/2012/12/05/is-it-legal-for-the-military-to-patrol-american-networks/

Mehta, A. (2018, June 26). *'We need to be impatient': Estonia's No. 2 defense official dives into NATO priorities*. Retrieved from Defense News: https://www.defensenews.com/smr/nato-priorities/2018/06/26/we-need-to-be-impatient-estonias-no-2-defense-official-dives-into-nato-priorities/

National Cyber Security Centrum. (2018, August 7). *Cyber Security Assessment Netherlands 2018*. Retrieved from National Cyber Security Centrum: https://www.ncsc.nl/english/current-topics/Cyber+Security+Assessment+Netherlands/cyber-security-assessment-netherlands-2018.html

NATO. (2016, July 9). *Warsaw Summit Communiqué*. Retrieved from NATO HQ: http://www.nato.int/cps/en/natohq/official_texts_133169.htm

Netherlands Ministry of Defence. (2018). *Defence Cyber Strategy 2018: Investing in cyber striking power for the Netherlands*. The Hague: Ministry of Defence.

Pernik, P. (2018, December 1). *Preparing for Cyber Conflict*. Retrieved from International Centre for Defence and Security: https://icds.ee/preparing-for-cyber-conflict-case-studies-of-cyber-command/

Pomerleau, M. (2017, October 19). *DoD says it shouldn't protect homeland from cyberthreats; McCain disagrees*. Retrieved from The Fifth Domain: https://www.fifthdomain.com/congress/capitol-hill/2017/10/19/dod-says-it-shouldnt-protect-homeland-from-cyberthreats-mccain-disagrees/

Rattray, G., & Healey, J. (2010). Categorizing and Understanding Offensive Cyber Capabilities and Their Use. *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy* (pp. 77-98). Washington, DC: The National Academy Press.

Ruiz, M. M. (2018, January 9). *Is Estonia's Approach to Cyber Defense Feasible in the United States?* Retrieved from War on the Rocks: https://warontherocks.com/2018/01/estonias-approach-cyber-defense-feasible-united-states/

Schmitt, M. N. (2017, August 8). *Grey Zones in the International Law of Cyberspace (2017 James Crawford Lecture on International Law)*. Retrieved from University of Adelaide: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjJsdjZornYAhUKbFAKHV7yD8QQFggpMAA&url=https%3A%2F%2Fore.exeter.ac.uk%2Frepository%2Fbitstream%2Fhandle%2F10871%2F27563%2FGrey%2520Zones%2520YJIL%2520-%2520Clean%2520

Somaskanda, S. (2018, June 4). *Cyberattacks Are 'Ticking Time Bombs' for Germany*. Retrieved from The Atlantic: https://www.theatlantic.com/international/archive/2018/06/germany-cyberattacks/561914/

Sulmeyer, M. (2018, February 13). *Department of Defense's Role in Protecting Democratic Elections: Testimony of Michael Sulmeyer before the Senate Armed Services Committee, Subcommittee on Cybersecurity*. Retrieved from U.S. Senate: https://s3.amazonaws.com/files.cnas.org/documents/SASC-Testimony-Feb-8.pdf

Sulmeyer, M. (2018, March 22). *How the U.S. Can Play Cyber-Offense*. Retrieved from Foreign Affairs: https://www.foreignaffairs.com/articles/world/2018-03-22/how-us-can-play-cyber-offense

U.S. Cyber Command. (2018, July 11). *2018 Cyberspace Strategy Symposium Proceedings*. Retrieved from U.S. Cyber Command: https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Cyberspace%20Strategy%20Symposium%20Proceedings%202018.pdf

U.S. Department of Defense. (2018, September 18). *Summary of the 2018 Department of Defense Cyber Strategy*. Retrieved from U.S. Department of Defense: https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF

U.S. Government Accountability Office. (2016). *GAO-16-332: DOD Needs to Clarify Its Roles and Responsibilities for Defense Support of Civil Authorities During Cyber Incidents*. Washington, DC: U. S. Government Accountability Office.

U.S. Government Accountability Office. (2017, November 30). *GAO-18-47: DOD Needs to Address Cyber Incident Training Requirements*. Retrieved from U.S. Government Accountability Office: https://www.gao.gov/products/GAO-18-47

U.S. Senate. (2018, 1 August). H.R.5515 - *John S. McCain National Defense Authorization Act for Fiscal Year 2019*. Retrieved from U.S. Congress: https://www.congress.gov/bill/115th-congress/house-bill/5515/text

U.S. House of Representatives. (2017). *H.A.S.C. No. 115-8, Cyber Warfare in the 21st Century: Threats, Challenges, and Opportunities*. Committee on Armed Services. Washington: U.S. Government Publishing Office.

Wallace, I. (2013, October 3). *Cyber security: Why military forces should take a back seat*. Retrieved from Lowy Institute: https://www.lowyinstitute.org/the-interpreter/cyber-security-why-military-forces-should-take-back-seat

Wilner, A. (2017). Cyber deterrence and critical-infrastructure protection: Expectation, application, and limitation. *Comparative Strategy, 36*(4), 309-318.

Withers, P. (2015, Spring). What is the Utility of the Fifth Domain? *Air Power Review*, 18(1), 126-150.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Covert or not Covert: National Strategies During Cyber Conflict

Gil Baram
School of Political Science
Tel Aviv University
Tel-Aviv, Israel
gilbaram@tauex.tau.ac.il

Udi Sommer
School of Political Science
Tel Aviv University
Tel-Aviv, Israel
http://people.socsci.tau.ac.il/mu/udis

Abstract: Anonymity is considered to be a key characteristic of cyber conflict. Indeed, existing accounts in the literature focus on the advantages of the non-disclosure of cyber attacks. Such focus inspires the expectation that countries would opt to maintain covertness. This hypothesis is rejected in an empirical investigation we conducted on victims' strategies during cyber conflict: in numerous cases, victim states choose to publicly reveal the fact that they had been attacked. These counterintuitive findings are important empirically, but even more so theoretically. They motivate an investigation into the decision to forsake covertness. What does actually motivate states to move into the international arena and publicly expose a cyber attack?

The goal of this paper is to understand why and under which geopolitical circumstances countries choose to give up the advantages of anonymity. Whether they wish to Name and Shame opponents for ignoring international norms or whether they try to avoid public humiliation, victims of cyber attacks occasionally reveal the fact that they had been attacked. There is tension between such motivations and the will to protect intelligence sources and the incentives to prevent escalation if an attack is revealed, even more so if the attacker is exposed. Indeed, we find that sunk costs, counter-escalation risks and the need to signal resolve—while critical in motivating victims to keep cyber attacks secret—may not suffice under such specific circumstances. By focusing on the victim's side, we draw inspiration from data on real-world cyber attacks in order to place cyber operations in the larger context of secrecy and covert actions in the international arena. In so doing, the aim is to advance the use of empirical data

[*] This research was supported by a research grant from the Blavatnik Interdisciplinary Cyber Research Center at Tel-Aviv University

for understanding the dynamics of cyber conflict and the decision-making process of states operating in this increasingly complex domain.

Keywords: *covert actions, cyber operations, national cyber strategies*

1. INTRODUCTION

In its 2019 Global Risks Report, the World Economic Forum ranked cyber attacks as one of the top ten risks, with respect to likelihood and impact (Myers and Whiting 2019). This concern is neither new nor surprising, given the anonymity that cyber attacks afford perpetrators and victims alike. By cyber attacks, which can be a part of an ongoing cyber operation, we mean both CNA (Computer Network Attack) and CNE (Computer Network Exploitation), as they cannot be fully separated (Siedler 2016).[1] Indeed, cyber technology enables countries to act covertly: the results of offensive actions in the cyber realm and their influence are not always exposed to the public eye. Furthermore, it is not always easy to identify who is behind a given attack. Even if the results of the attack are publicly observable—e.g., damage to a power grid leading to the severance of electricity supply—the victim can still dismiss these effects, arguing that they were the result of a technical fault. To date, our understanding of those strategic interactions between attacker and victim—and their decisions about whether or not to keep attacks covert—is theoretically and empirically limited.

Recent work regarding covert actions in the international arena offers three mechanisms that make the use of covert actions preferable for countries: sunk costs, counter-escalation risks and signaling resolve (Carson 2016; Carson and Yarhi-Milo 2017). These mechanisms, to be discussed in detail in Section 3, suggest that countries have strong incentives to engage in covert actions and keep those actions away from the public eye, domestically as well as internationally.

Yet an empirical investigation conducted on states' strategies in the wake of cyber attacks reveals a different picture. Notwithstanding the advantages of maintaining secrecy, it is not uncommon for victims to reveal the fact that they have been attacked. What causes victims of cyber attacks to "abandon" the covert space and move to the public arena in the aftermath of an attack? Existing literature does not offer satisfying answers (for exceptions see Edwards, Furnas, Forrest and Axelrod 2017; Poznansky and Perkoski 2018). To understand the puzzling strategic choice to abandon the advantages of ambiguity in favor of a public strategy, we need to understand the tradeoffs between the strategies. As not all countries choose to either publicly reveal

[1] As Libicki concluded, "as long as the methods of cyber espionage look like the methods of cyberattack the discovery of one will raise fears about the imminence of the other." (Libicki 2018, 121)

the attack or to hide it, we recognized that the strategies of the victims vary between four possible approaches:

(1) "Pointing a finger" – publicly disclosing that an attack occurred (revealing vulnerability) and publicly putting the responsibility on a specific attacker;

(2) Admitting injury – publicly disclosing that an attack took place, while failing to identify an attacker;

(3) Revealing damage – disclosing damage but denying that it had been caused by a deliberate hostile attack (claiming technical malfunctions, system "glitches" etc.);

(4) Maintaining ambiguity – denying or downplaying any damage, thus reducing the chances that the attack would ever be divulged.

Table 1 summarizes those four strategies with illustrations from cyber attacks in recent years.

TABLE 1: VARIANCE IN VICTIM'S STRATEGIES DURING CYBER ATTACKS, WITH REAL-LIFE EXAMPLES

	Publicized		**Concealed**	
Victim's Strategy:	(1) Publicizing the attack and blaming the attacker (Public Strategy #1)	(2) Publicizing the attack and not blaming the attacker (Public Strategy #2)	(3) Partial concealment (claim of fault)	(4) Full concealment of the attack
Real-life Example:	DNC hack 2016	SingHealth hack 2018	USS John S McCain collision 2017	---

Our discussion focuses on the first two options, where the victim decided to make the attack public and sometimes also to reveal the attacker's identity. The third option (partial concealment) deals with cases where the alleged victim claims that a certain event was the result of a technical problem and not due to a cyber attack. To illustrate this option in a nutshell—since we do not delve into its details in the paper—let us look at the summer 2017 case of the *USS John S McCain*, which collided with a merchant ship in the Straits of Malacca, resulting in the death of 10 sailors (Werner 2018). The Chief of Naval Operations argued that there was no evidence suggesting the accident was the result of a cyber attack. However, according to experts, since the destroyer had a large navigational team as well as another team in charge of radar, it was impossible that human error had led to the accident. In addition, both the destroyers *USS McCain* and the *USS Fitzgerald*, which had been hit in a similar incident in June 2017, belong to the Seventh Fleet. Experts believed these attacks may have been related to Chinese or Russian intervention (Mass 2017).[2]

[2] The Navy's investigation found no evidence of a cyber attack (Tritten 2017; Navy Releases Collision Report 2017).

After discussing the place of attribution and secrecy in cyber operations and their impact on states' strategic calculations, we develop our theoretical framework and examine two cases – hacking into the Democratic National Committee in 2016 and the SingHealth hack in 2018. It is particularly in the analysis of those two well-studied cases that our theoretical framework helps to shed new light on the national and international considerations leading countries to give up secrecy. We highlight the taxonomy of the different prototypes of these strategies and help to identify when countries might choose each strategy.

2. ATTRIBUTION AND SECRECY – AN INHERENT COMPONENT OF CYBER OPERATIONS?

The covertness of cyber attacks can be expressed in two ways. First, the attack itself is covert. Its technological characteristics enable an attacker to carry out the operation in a clandestine way, without revealing how it was carried out. The second aspect concerns the attackers themselves, who can maintain covertness.[3] It is often difficult to point out the source of an attack and to attribute it to a particular attacker. This problem is known as the Attribution Problem.

The Attribution Problem arises when the victim identified the attack, but has yet to identify the attacker. The immediate effect of this lack of certainty raises questions concerning the feasibility of retaliation, and the desire for it. Such a situation creates uncertainty as to the attacker's demands. It can be difficult to determine by technical means the motivation for an attack (Wheeler and Larsen 2003, 1). So, as Rid and Buchanan argue, "attribution is what states make of it" (Rid and Buchanan 2015, 7).

When an attribution process is conducted using intelligence sources and methods, it is difficult to expose it without endangering these sources. But if the domestic public—especially in a democratic polity—perceives the attribution as unreliable, the state may lose the legitimacy to retaliate (Lindsay 2015). An important part of the attribution process is its political implications. Indeed, "communicating attribution is part of attributing" (Rid and Buchanan 2015, 26). When an attack is executed, security researchers attempt to find out who is behind it. In order to do so, they examine the code, techniques and protocols that the attacker used. However, this is not considered legitimate proof in court and is seen, especially today, more as playing a "blame game" (Berghel 2017, 86).

Faith-based attribution happens when actors blame other actors for an attack if they believe the former carried it out. This also happens in modern politics, where politicians knowingly make incorrect statements, simply because no one checks their

[3] On the distinction between clandestine and covert operations, see Poznansky and Perkoski, 2018, 403.

validity (Berghel 2017; Carr 2016). Healey (2013) also argues that scholarship should move forward from dealing with the attribution problem. Instead of asking “who is behind the attack?” the question should be “who is to blame for it?” (Healey 2013, 55) and what are the political consequences of blaming?

This study adopts Healey’s approach in the sense that the technical attribution problem is not as crucial for our framework. In practice, countries routinely accuse each other even without disclosing the full technical process that led them to attribute the attack to a particular attacker. This was the case in the Sony hack (2014) and the “WannaCry” attack (2017) when the US blamed North Korea without fully disclosing technical evidence.

Despite the inherent overlap between cyber operations and covert actions, the scholarship has not fully explored this connection and has studied these fields separately for the most part. On the one hand, the cyberwarfare scholarship in International Relations and Security Studies hardly deals with the different aspects of secrecy in cyber operations, and mainly accepts the assumption that anonymity is an immutable feature of cyberspace rather than something actors select into and which they can therefore forfeit (for exceptions see Lupovici 2016; Poznansky and Perkoski 2018). On the other hand, scholars dealing with covert operations largely tend not to include cyber operations in their analyses (for exceptions see Brecher 2012). This study is an important step towards merging these bodies of literature.

Recent work regarding secrecy in cyberspace tends to study the considerations before the attack (Edwards, Furnas, Forrest and Axelrod 2017), the perpetrators’ calculations (Poznansky and Perkoski 2018), and the effect of cyber attacks on democratic states’ accountability to their citizens (Schulzk 2018). While these studies are an important step in combining the two literatures, more research is needed in order to understand cyber operations as covert actions and to investigate to what extent countries choose to use the advantages of this covertness or to give it up. In the following sections these considerations are examined from the victim’s point of view. We focus on the victim, since in most circumstances the victim is the first to make a choice about whether to use covertness or forsake it.

3. GIVING UP SECRECY AS A NATIONAL STRATEGY

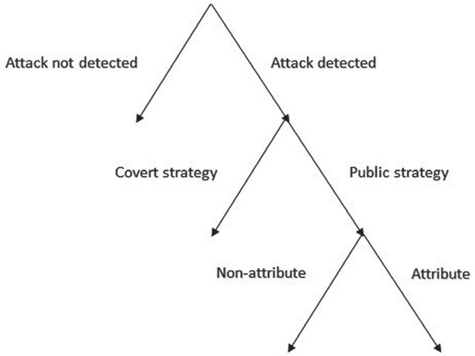
Three mechanisms are offered in the literature for making the use of covert actions preferable for countries. First are sunk costs, which refer to situations where states decide to take covert action because of non-recoverable resources: by choosing to use covert actions, leaders can employ a more “creative” way to address security threats

(Carson and Yarhi-Milo 2017, 135). Second are counter-escalation risks: using covert action can appear credible because of its impact on the risk of crisis escalation, since leaders using covert signaling tools can be free to engage in more aggressive behavior. This explanation is based mainly on the audience costs literature, which identifies a link between the type of action that the state takes and the costs the leader will have to bear as a result (see Fearon 1994; Tomz 2007). The last mechanism is signaling resolve: under certain conditions, the use of covert operations allows states to convey the desired message to their rivals, and therefore they do not have to act in the public arena (Carson 2016; Carson and Yarhi-Milo 2017, 134-135).

But it seems that during cyber attacks, that might be a part of an on-going operation or a one-time attack, the options available to the victim are different, and revealing the attack has its benefits. Generally, there are cases where the incentives to remain covert are not enough and decision-makers have other incentives—such as avoiding public humiliation, warning the attacker from taking future actions and more—that lead them to decide to publicly reveal the attack.

Once the victim has identified the attack and decides to use a public strategy, it has two major options as mentioned earlier: (1) reveal the attack and point a finger towards the attacker, or (2) reveal only the fact that the attack has occurred, without disclosing the identity of the alleged attacker. Figure 1 summarizes the strategies at earlier stages and as they lead up to the strategies at this stage.

FIGURE 1: VICTIM’S STRATEGIES DURING A CYBER ATTACK

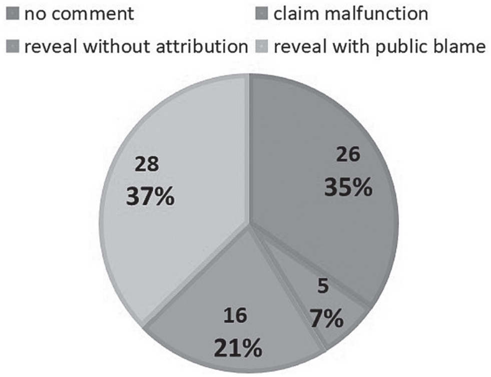


To assess the conditions under which countries that have suffered a cyber attack choose to reveal the attack and go public, we examined all known cyber attacks between rival states from 2015 to mid-2018. The framework of the Dyadic Cyber Incident Dataset (DCID) v1.1 (Maness, Valeriano and Jensen 2017) was the basis for the coding, and new attacks from the Council on Foreign Relation Cyber Operations Tracker were added (Segal 2017). The unit of analysis in both datasets is state-sponsored cyber attacks.[4] We focus on state-sponsored actors because our purpose is to identify when states and their proxies conduct cyber operations in pursuit of their foreign policy interests. New variables originally collected by us were added in order to examine the victims' strategies. All data collected are open source.[5]

Our data indicate that there is wide variation in the victims' strategies: Between 2015 and mid-2018, 75 cyber attacks were conducted between rival states. In 44, the victims chose to address the attack publicly. Of those, in 16 the victim revealed the attack and did not attribute it. In the remaining 28, the victim revealed the attack and publicly attributed it to a specific attacker (Figure 2). Out of the 28 cases where victims chose to publicly reveal the attack and the attacker, only three states were not democratic.[6]

The data suggest that states frequently choose public strategies. Although at first glance, revealing the attack might be perceived as exposing a country's weakness, there are several considerations with positive implications, which could lead the country to decide to reveal the attack. The question is: why do states act that way, and in the pursuit of which advantages?

FIGURE 2: VARIANCE IN VICTIMS' STRATEGIES BETWEEN 2015 AND MID-2018



[4] This paper focuses only on state-sponsored cyber attacks. Doing that allowed us to achieve in-depth insights regarding the ways countries operate during cyber conflict. Keeping out of the analysis other kinds of cyber attacks, such us multi-victim attacks and attacks against NGOs, might pose a methodological challenge. Due to the limited scope of this paper we do not treat these kinds of cyber attacks here, and will deal with them in future projects.

[5] The "unknown-unknowns" cyber attacks are the ones that are not known to the public. This paper deals only with cyber attacks that have been publicly revealed and that had sufficient data on them in order to code it in our dataset.

[6] According to Freedom House.

Reasons to publicly reveal the attack

In most cyber attacks the victim does not have full confidence regarding the identity of the attacker. Furthermore, there are questions around to what extent a victim that chooses to accuse the attacker is certain of the accuracy of its identification. If it possesses technical evidence that can be exposed, the attacker will have more difficulty denying the charges. However, more often than not this is not the case. It is common for a victim to point a finger at a particular attacker even without disclosing the full technical evidence that led to that attribution.

In the political and technical landscapes of our time, it is important to consider cyber attacks in the broader geostrategic context. In many cases there is an ongoing political tension that means it is in the victim's interests to reveal the aggressive actions of its adversary, a strategy known as Naming and Shaming. A Naming and Shaming strategy means publicly identifying perpetrators that are "doing wrong" and undermining international law and the rules-based order. This might look like the victim is admitting to its weakness. Yet, in a long-term cost-benefit analysis, sometimes it is better to "call out" the aggressor as violating international norms than to remain silent. This might help the victim and its allies to improve their cybersecurity readiness, while also reaffirming the victim's commitment to law and norms (on publicizing states activities see Carnegie and Carson 2018).

An additional consideration in revealing attacks is the need to avoid public humiliation. The victim can decide to disclose the attack due to the desire to avoid humiliation and degradation, which will most likely accompany the publication of the said attack by the attacker or by a third party. In a post-Snowden reality, remaining covert is hard. The general public is more aware of state activities and has the means to publicize them via social media as well as in various other ways. As a result, the political costs of transparency may be less than those associated with hiding an attack. This minimizes the victim's reputational damage and helps to improve overall cybersecurity of both victim and international allies alike.

Another goal may be showing strength in front of an international audience by warning the attacker against taking future actions. By disclosing the attack and accusing the attacker, the victim conveys a message that it has identified the attack and may intend to retaliate; plus, it has the technical know-how to identify the attack and point out the entity behind it. If the victim can say to the presumed attackers that it knows what they are up to, it implies that it also knows a lot more about the attackers' operations and capabilities. This may introduce uncertainty into the decision-making process and induce a strategic effect. Such was the case with the Obama-Xi agreement from 2015 that reduced Chinese industrial cyber espionage for a limited period of time (Spetalnick and Martina 2015). A country that exposes the attack and points a finger at the attacker, while showing its methods of coping and the ways in which it operates to strengthen its defense capabilities, is portrayed as a leader in the international arena in dealing with cyber attacks.[7] Other countries will observe and learn from it, as was the case with the Democratic National Committee hack, which is discussed in detail later on.

Motivations not to reveal the attacker

Assuming that the victim identified the attacker, there are at least two main reasons why the victim would not want to reveal the attacker's identity in public:

(1) Safety of intelligence sources. The desire to avoid exposing intelligence and sources is an important reason not to make the identity of the attacker public. This is even more acute in cyberspace, because it is difficult to identify the attacker only using technical tools. Therefore, it is often necessary to use intelligence of various kinds, such as advanced technological and even human resources to obtain the necessary information. These sources are considered highly important and valuable for the country's intelligence services, and therefore it is essential to protect their safety and not to expose them.

(2) Preventing escalation. There may be differences in the existing technological capabilities and power of the victim and the attacker. If this is the case, the victim might choose not to publicize the attack in order to avoid the chance that the exposure would lead to open confrontation. An aggressive public intervention by one country in another's affairs poses a political-strategic challenge to the victim in the eyes of the domestic public and the international community, who are watching and waiting to see how it responds (Carson 2016). Not revealing the identity of the attacker allows the victim to refrain from the obligation to respond, contain the attack and prevent undesirable escalation.

We expect victims to choose to reveal the attack publicly and attribute it when (a) they want to expose the aggressor and blame them for violating international norms; (b) avoid international and domestic humiliation; (c) warn the attacker. However, by revealing the attack and not attributing it, the victim can also avoid humiliation and there are covert ways to convey a deterrent message. Therefore, we hypothesize that in this case key reasons for not attributing the attack are (a) the safety of intelligence sources; and (b) preventing escalation. The two cases tested in the next section will help examine these expectations.

[7] We are aware that there are other considerations for countries to reveal the attack, such as creating a false attribution for political reasons or faking non-existent capabilities by revealing; using the publicized attack for political reasons such as increase allied support; cases when there is a public leak and the victim is being forced to reveal the attack; internal political considerations and more. The scope of this paper will not allow us to deal with all these considerations but they will be taken into account in our larger research agenda.

4. GIVING UP SECRECY IN CYBER OPERATIONS – REAL-LIFE CASES

Two major cyber attacks that occurred in the past three years are examined. They allow us to illustrate the public strategies identified and described theoretically above.

Democratic National Committee Hack 2016

In April 2016, hackers gained access into the Democratic National Committee (DNC) network, stealing several gigabytes of data. From June-November 2016, WikiLeaks published 20,000 emails of DNC members, and in July 2016 the FBI began an investigation of the hack. The investigation revealed that in the months prior to the WikiLeaks releases, two groups of hackers operating under the auspices of the Russian government broke into the computers of the DNC and leaked the emails. This action was part of a broader Russian operation in the months before the presidential election in 2016, intended to influence the election results and to jeopardize the integrity of the democratic processes (Bump 2018).

On December 2016, President Obama publicly accused Russia of carrying out these attacks, warned that it must stop and said that the US had offensive cyber capabilities and it might respond. At the end of that month, President Obama ordered the expulsion of 35 Russian diplomats from the US, as well as the closure of sites which were used by the Russians to gather intelligence (Landler and Sanger 2016; Ryan, Nakashima and De Young 2016). The Department of Homeland Security (DHS) and the FBI published a joint statement describing the process of the Russian cyber attack, directly accusing military and civilian Russian intelligence agencies. According to the statement, "The US Intelligence Community is confident that the Russian Government directed the recent compromises of emails from US persons and institutions […] only Russia's senior-most officials could have authorized these activities." (Department of Homeland Security 2016). The operations of Russian intelligence agencies included "spear phishing" attacks of entities in government agencies, critical infrastructure, think tanks, universities, political organizations, and more, in order to steal information (Masters 2018).

The fact that the US chose to publicly accuse Russia of the attack helped strengthen its international standing by calling out Russia's undermining of the international order in trying to manipulate and sabotage democratic procedures. Such attempts to influence election results are perceived by Western democracies as damaging their political and institutional integrity. Other countries also saw and learned from the American experience. Following the exposure of the attack, the US became the focus of interest for other democratic countries—such as France and Germany—which were about to hold their own elections and feared Russian intervention. For example, the NSA warned French officials that Russian hackers had compromised some elements of the election (Greenberg 2017).

The experience gained by the US in dealing with Russian activity enabled it to share information and assist other countries. The US became a role model for confronting Russian influence attempts and protecting election campaigns (Graham, 2017). This case demonstrates the value of our theoretical framework: by publicly revealing the attack, the US avoided public humiliation that could have happened if a third party or Russia itself had revealed the attack instead. Also, by conveying a deterrent message to the Russians, the US made a coercive threat and demonstrated resolve. It showed its will to spend valuable resources in order to make Russia pay a price for its offensive actions.

SingHealth Hack 2018

On 4 July 2018, data administrators detected unusual activity on one of SingHealth's IT databases. With more than two million patients, SingHealth is the largest health provider in Singapore. The security team immediately investigated the suspicious activity to determine its nature and whether it was malicious. On July 10th, after forensic investigations confirming it was a cyber attack, SingHealth, the Ministry of Health and the Cyber Security Agency (CSA) were informed (Tham 2018). The cyber attack resulted in the personal details of 1.5m SingHealth patients being accessed and copied; this included names, identification numbers, address, gender, race and date of birth, including the personal data of Singapore's Prime Minister. On July 20th, even while investigations were still under way, SingHealth and investigating authorities assessed that the situation had been stabilized and informed the public of the cyber attack, (Singapore Ministry of Health 2018).

Following the attack, a public Committee of Inquiry was established. A senior counsel in the Ministry of Justice summarized in front of the committee how advanced, determined and disciplined the attackers were: "The skill and sophistication used in the SingHealth attack highlights the challenges that cyber defenders face" (Tham and Baharudin, 2018). Speaking at a press conference on July 20th 2018, the Chief Executive of the CSA, David Koh, confirmed that: "We have determined that this is a deliberate, targeted and well-planned cyber attack, not the work of casual hackers [...] we are not able to reveal more because of operational security reasons" (Koh 2018). From Koh's words it seems that for national security reasons the CSA wanted to keep its intelligence sources safe and did not reveal any information that could risk them.

Although the head of the CSA estimated that a nation state was behind the attack, and many security analysts even estimated it was China, Singapore was careful not to reveal the identity of the attacker in public. The decision to make the attack public was based on two main considerations. The first derived from the theft of personal information that is critical for the daily life of citizens. As most activities that are essential to the daily lives of Singapore's citizens take place online, there was a concern that the attacker might want to use the data to gain access to additional personal details (Tham and Baharudin, 2018).

Another consideration for exposing the attack, but keeping the identity of the attacker undisclosed, was concern about public humiliation. If the attacker or a third party exposed the attack before the Singaporean authorities did, it could damage the reputation of the administration. In such circumstances, the administration would appear to have failed to protect its citizens and to have made an attempt to conceal it.

While experts pointed fingers at China (Lee 2018), authorities remain tight-lipped. One explanation for that is the need to avoid escalation. China and Singapore have a close relationship, but differences have been experienced during numerous high-profile events, including Singapore's stance against China regarding the South China Sea dispute. The power differential between the two, and the will of Singapore not to take any steps that could risk this relationship and escalate the situation, seem to be among the main reasons why Singapore chose not to reveal the identity of the attacker. Further support for the decision not to reveal the identity of the attacker was given by the Minister-in-Charge of cybersecurity. In January 2019, the Minister stated that: "Revealing the identity of the perpetrator would not be in the Republic's national interest […] We've got nothing to hide here […] the only part that's been held back are those that pertain to sensitive national security matters and also patient confidentiality" (Nair 2019; Yufeng 2019).

5. CONCLUSIONS

Reasons ranging from attempts to Name and Shame or avoid public humiliation, to incomplete confidence about the identity of the attacker may lead victims of cyber attacks to reveal the fact they had been attacked. There is tension, however, between such reasons and the motivation to protect the safety of intelligence sources and the will to prevent escalation if an attack is revealed and even more so if the attacker is exposed. The preliminary results and analyses presented here demonstrate that despite a range of reasons to remain covert, countries that suffered cyber attacks have sufficiently strong incentives to reveal the fact they had been attacked. The three mechanisms presented in the literature as motivating decision-makers to keep the attack covert—sunk costs, counter-escalation risks and signaling resolve—do not always suffice in the cyber reality. Not only would victims make the attack public, but under certain circumstances, they would even expose the attacker. This finding is both unintuitive and largely undocumented in the literature.

The will to avoid domestic and international humiliation if the attack will be exposed by a third party leads countries to give up the advantages of secrecy in cyberspace and reveal the fact that they had been attacked. Furthermore, attributing the attack to a specific attacker helps the victim to warn the attacker from taking future actions and be model for other countries who deal with similar attacks. Such was the case in the DNC hack where the US not only set the standard for other countries in the West but also aided them in preventing potential threats to the integrity of their democratic process.

National security considerations such as keeping intelligence sources safe and avoiding escalation play an important part in the decision to reveal the attack without attributing it to a specific attacker. Such was the case in the SingHealth hack. To protect citizens' online identity and e-government business, the attack was made public by the government in Singapore. Yet, its source remained undisclosed, possibly to avoid causing a geostrategic threat of escalation.

Future research is essential. In particular, in this paper we limited the theoretical discussion and empirical work to public strategies exclusively. We did not deal with the other two options from Table 1 – partial concealment and full concealment of the attack and did not analyze the attackers' strategies and the utility of the interaction between both sides.

Acknowledgments

We thank Dana Litman, Adi Lotan and Eden Poch for excellent research assistance.

REFERENCES

Berghel, Hal. "On the Problem of (Cyber) Attribution." *Computer* 3, no. 50 (2017): 84-89. https://www.computer.org/csdl/mags/co/2017/03/mco2017030084.pdf

Brecher, Aaron. "Cyberattacks and the Covert Action Statute: Toward a Domestic Legal Framework for Offensive Cyberoperations." *Michigan Law Review* (2012): 423-452. https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/mlr111&id=452&men_tab=srchresults

Bump, Philip. 2018. "Timeline: How Russian agents allegedly hacked the DNC and Clinton's campaign." *Washington Post*, July 13, 2018. https://www.washingtonpost.com/news/politics/wp/2018/07/13/timeline-how-russian-agents-allegedly-hacked-the-dnc-and-clintons-campaign/?utm_term=.f7d3b8b7fe50

Carnegie, Allison and Austin Carson. "The Spotlight's Harsh Glare: Rethinking Publicity and International Order." *International Organization* 72, no. 3 (2018): 627-657. https://doi.org/10.1017/S0020818318000176

Carr, Jeffrey. 2016. “Faith-based Attribution?” *Medium*, July 10, 2016. https://medium.com/@jeffreyscarr/faith-based-attribution-30f4a658eabc

Carson, Austin. “Facing Off and Saving Face: Covert Intervention and Escalation Management in the Korean War.” *International Organization* 70, no. 1 (2016): 103-131. https://doi.org/10.1017/S0020818315000284

Carson, Austin and Keren Yarhi-Milo. “Covert Communication: The Intelligibility and Credibility of Signaling in Secret.” *Security Studies* 26, no. 1 (2017): 124-156. https://doi.org/10.1080/09636412.2017.1243921

Edwards, Benjamin, Alexander Furnas, Stephanie Forrest and Robert Axelrod. “Strategic Aspects of Cyberattack and Blame,” *Proceedings of the National Academy of Sciences* 114, no. 11 (March, 2017): 2825-2850. https://doi.org/10.1073/pnas.1700442114

Fearon, James. “Domestic Political Audiences and the Escalation of International Disputes.” *American Political Science Review* 88, no. 3 (1994): 577-592. https://doi.org/10.2307/2944796

Graham, Chris. 2017. “French Election: Are Russian Hackers to Blame for Emmanuel Macron’s Leaked Emails - and Could They Target UK Election?” *The Telegraph*, May 6, 2017. https://www.telegraph.co.uk/news/2017/05/06/russian-hackers-blame-emmanuel-macrons-leaked-emails-could/

Greenberg. Andy. 2017. “The NSA Confirms it: Russia Hacked French Election ‘Infrastructure’”, *Wired*, September 5, 2017. https://www.wired.com/2017/05/nsa-director-confirms-russia-hacked-french-election-infrastructure/

Healey, Jason. “The Spectrum of National Responsibility for Cyberattacks.” *Brown Journal of World Affairs* 18, no. 1 (2013): 57-70. https://www.jstor.org/stable/24590776

Koh, David. 2018. “*CSA on Investigations regarding the Deliberate Cyber Attack*,” YouTube video, 0:31, July 20, 2018. https://www.youtube.com/watch?v=toM_WXImOBc&index=3&list=PLH2CR4s1lqyiZZ1n6wVvW_uMMR4fFXrW4

Landler, Mark and David Sanger. 2016. “Obama Says He Told Putin: ‘Cut it Out’ on Hacking.” *New York Times*, December 16, 2016. https://www.nytimes.com/2016/12/16/us/politics/obama-putin-hacking-news-conference.html

Lee, Justina. 2018. “Suspected China cyberhack on Singapore is a wake-up call for Asia,” *Nikkei Asian Review*, August 21, 2018. https://asia.nikkei.com/Spotlight/Asia-Insight/Suspected-China-cyberhack-on-Singapore-is-a-wake-up-call-for-Asia

Libicki, Martin. 2018. “Drawing Inferences from Cyber Espionage,” *2018 10th International Conference on Cyber Conflict*, NATO CCD COE Publications, Tallinn. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8405013

Lindsay, Jon. “Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack.” *Journal of Cyber Security* 1, no 1 (2015): 53-67. https://doi.org/10.1093/cybsec/tyv003

Lupovici, Amir. “The ‘Attribution Problem’ and the Social Construction of ‘Violence’: Taking Cyber Deterrence Literature a Step Forward.” *International Studies Perspectives* 17, no. 3 (2016): 322-342. https://doi.org/10.1111/insp.12082

Maness, Rayn, Brandon Valeriano, and Benjamin Jensen. “Codebook for the Dyadic Cyber Incident and Dispute Dataset Version 1.1.” (2017). https://drryanmaness.wixsite.com/cyberconflcit/cyber-conflict-dataset

Mass, Warren. 2017. “Cyber Experts Believe Hacking may have Caused Collision of USS John S. McCain.” *The New American*, August 22, 2017. https://www.thenewamerican.com/tech/computers/item/26753-cyber-experts-believe-hacking-may-have-caused-collision-of-uss-john-s-mccain

Masters, Jonathan. 2018. “Russia, Trump, and the 2016 U.S. Election.” *Council on Foreign Relations*. February 26, 2018.https://www.cfr.org/backgrounder/russia-trump-and-2016-us-election.

Myers Joe and Kate Whiting. 2019. “These are the biggest risks facing our world in 2019.” *World Economic Forum*, January 16, 2019. https://www.weforum.org/agenda/2019/01/these-are-the-biggest-risks-facing-our-world-in-2019/

Nair, Suresh. 2019. “Singapore Healthcare cyberattack: Not revealing hacker still a ‘puzzler’.” *The Independent*, January 16, 2019. http://theindependent.sg/singapore-healthcare-cyberattack-not-revealing-hacker-still-a-puzzler/

Navy Office of Information. *Navy Releases Collision Report for USS Fitzgerald and USS John S McCain Collisions*, 2017. https://www.navy.mil/submit/display.asp?story_id=103130

Poznansky, Michael and Evan Perkoski. “Rethinking Secrecy in Cyberspace: The Politics of Voluntary Attribution.” *Journal of Global Security Studies* 3, no. 4 (2018): 402-416. https://doi.org/10.1093/jogss/ogy022

Rid, Thomas and Ben Buchanan. “Attributing Cyber Attacks.” *Journal of Strategic Studies* 38, no. 1-2 (2015): 4-37. https://doi.org/10.1080/01402390.2014.977382

Ryan, Missy, Ellen Nakashima and Karen De Young. 2016. “Obama administration announces measures to punish Russia for 2016 election interference.” *The Washington Post*, December 29, 2016. https://www.washingtonpost.com/world/national-security/obama-administration-announces-measures-to-punish-russia-for-2016-election-interference/2016/12/29/311db9d6-cdde-11e6-a87f-b917067331bb_story.html?utm_term=.e4a51ff3e57d

Schulzke, Marcus. “The Politics of Attributing Blame for Cyberattacks and the Costs of Uncertainty.” *Perspectives on Politics* 16, no. 4 (2018): 954-968. https://doi.org/10.1017/S153759271800110X

Segal, Adam. 2017. *Tracking State-Sponsored Cyber Operations*, Council on Foreign Relations, November 6, 2017. https://www.cfr.org/blog/tracking-state-sponsored-cyber-operations

Siedler, Endresen. “Hard power in cyberspace: CNA as a political means.” *2016 8th International Conference on Cyber Conflict (CyCon)*. IEEE, (2016). https://ieeexplore.ieee.org/abstract/document/7529424

Singapore. Ministry of Health. *Cyberattack on SingHealth’s IT System*. August 6, 2018. Accessed December 31, 2018. https://www.moh.gov.sg/news-highlights/details/cyberattack-on-singhealth’s-it-system.

Spetalnick, Matt and Michael Martina. 2015. “Obama announces ‘understanding’ with China’s Xi on cyber theft but remains wary.” *Reuters*, September 26, 2015. https://www.reuters.com/article/us-usa-china/obama-announces-understanding-with-chinas-xi-on-cyber-theft-but-remains-wary-idUSKCN0RO2HQ20150926

Tham, Irene. 2018. “Personal info of 1.5m SingHealth patients, including PM Lee, stolen in Singapore’s worst cyberattack.” *The Straits Times*, July 20, 2018. https://www.straitstimes.com/singapore/personal-info-of-15m-singhealth-patients-including-pm-lee-stolen-in-singapores-most

Tham, Irene and and Hariz Baharudin. 2018. “Attempt on July 19 was detected and cut off on same day, thanks to heightened monitoring.” *The Straits Times*, October 6, 2018. https://www.straitstimes.com/tech/hackers-made-another-intrusion-attempt-as-probe-was-under-way

Tomz, Michael. “Domestic Audience Costs in International Relations: An Experimental Approach.” *International Organization* 61, no. 4 (2007): 821-840. https://doi.org/10.1017/S0020818307070282

Tritten, Travis. 2017. “Navy Chief: There’s no evidence recent collisions were caused by hacking.” *Business Insider*, August 30, 2017. https://www.businessinsider.com/navy-chief-no-evidence-recent-collisions-were-caused-by-hacking-2017-8

United States. Department of Homeland Security. *Joint Statement from the Department of Homeland Security and Office of the Director of National Intelligence on Election Security*: DHS Press Office, October 7, 2016. https://www.dhs.gov/news/2016/10/07/joint-statement-department-homeland-security-and-office-director-national

Werner, Ben. 2018. “USS John S. McCain Collision, A Year Later.” *USNI News*, August 21, 2018. https://news.usni.org/2018/08/21/35947.

Wheeler, David and Gregory Larsen. *Techniques for Cyber Attack Attribution*. Alexandria, VA, 2003. https://apps.dtic.mil/dtic/tr/fulltext/u2/a468859.pdf

Yufeng, Kok. 2019. “SingHealth attacker known: Iswaran.” *The New Paper Singapore*, January 16, 2019. https://www.tnp.sg/news/singapore/singhealth-attacker-known-iswaran

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

The Cyber-ASAT: On the Impact of Cyber Weapons in Outer Space

James Pavur
DPhil Researcher
Cybersecurity Centre for Doctoral Training
Oxford University
Oxford, United Kingdom
james.pavur@cybersecurity.ox.ac.uk

Ivan Martinovic
Professor of Computer Science
Department of Computer Science
Oxford University
Oxford, United Kingdom
ivan.martinovic@cs.ox.ac.uk

Abstract: Satellites have revolutionized military strategy and the dynamics of national power. However, satellites themselves are fragile and can be destroyed by even miniscule projectiles. Anti-Satellite Weapons (ASATs) which exploit this weakness have long been prophesied as the Achilles heel of space power; yet orbit has remained relatively peaceful for more than sixty years.

As the threat of cyber attacks against space assets looms, the impact that cyberspace will have on stability in outer space is not well understood. This paper presents a strategic analysis of the impact of cyber weapons on three key stabilizing factors which have thus far contributed to peace in space. Based on this analysis, it contends that cyber-ASATs threaten the foundations of space's longstanding stability due to their high accessibility, low attributability, and low risk of collateral damage.

This conjecture is tested experimentally though the development of a simulated cyber-ASAT capability targeting one small component of satellite operations: space situational awareness data. By leveraging orbital simulations and genetic algorithms, we demonstrate the ability to artificially alter debris collision forecasts and cause direct harm to critical space systems without firing a single rocket. The attack method is tested in realistic simulations and shown to have a high success rate against real-world satellites of vital strategic importance.

Our interdisciplinary approach unifies strategic analysis with technical experimentation to present the case that cyber-ASATs are not merely a distant theoretical threat, but a real and present danger to the balance of power in space.

Keywords: *satellites, ssa, cyber-ASAT, space, space power, ASAT*

1. INTRODUCTION

In 1958, then US Senator Lyndon Johnson predicted that ‘control of space means control of the world’ [1, p. 287]. 33 years later, Operation Desert Storm, widely referred to as ‘the first space war’, validated this prophecy [2]. Overwhelming US dominance during the 100-hour ground war was directly attributable to the support of over 60 positioning, communications and reconnaissance satellites [3]–[5].

Modern space power has created a world in which ‘no enemy can withstand a frontal assault upon U.S. forces due to the American ability to sense, move, and strike with precision’ [6, p. 236]. As the world becomes increasingly multipolar, many other states are expected to seek the same prestige and military power, associated with membership of the ‘space club’ [7], [8]. Over the past half-century, space has become the ‘ultimate high ground’ for information age warfare [9], [10, p. 714].

This strategic vitality stands at odds with critical vulnerability. Satellites are lightweight and fragile devices moving at incredible speeds. A marble-sized projectile or debris particle in orbit could strike a satellite with the force of a one-ton object falling from a height of five storeys [11]. In the seminal days of space strategy, this physical weakness was thought to undermine the strategic utility of space itself [12, Ch. 5]. The rise of anti-satellite weapons (ASATs), which exploit this weakness, has long been prophesied to bring about the end of space power. However, somehow, orbit has remained remarkably peaceful.

As space systems become increasingly interconnected and computationally complex, new concerns about the threat of cyber-attacks have been raised [13]. However, the strategic implications and technical feasibility of cyber-ASATs are not well understood. This paper seeks to unite strategic and technical perspectives on cyber attacks in space as a starting point for policymakers and technicians to address these threats.

2. CONTRIBUTIONS

The core motivator for our research was to credibly assess if cyber-ASAT capabilities pose a fundamental challenge to the dynamics of orbital peace, or if the structural factors which have stabilized space for the past half-century will continue to endure.

To this end, this paper begins with a brief overview of three widely recognized stabilizing forces: limited accessibility, attributable norms and environmental interdependence. We then contribute what we believe to be the first high-level strategic consideration of cyber-ASATs with regard to each of these factors. We predict that cyber-ASATs can undermine all three, due to their widespread accessibility, weak norms and attribution, and environmental indifference.

To bolster these theoretical claims, this paper adopts an interdisciplinary approach, leveraging an experimental case study to verify the technical feasibility of the cyber-ASATs that it predicts will emerge. This case study revolves around the creation and simulation of a cyber-ASAT capability, targeting space situational awareness (SSA) data. Our attack method combines orbital simulations and genetic algorithms to artificially alter debris collision projections and induce harmful satellite manoeuvres. The attack is verified through experimental simulations against more than 100 major communications satellites; we demonstrate a greater than 90% success rate against all targets.

Together, our experimental findings and strategic assessment suggest that cyber-ASATs are not merely another tool in the anti-satellite arsenal, but a real and present danger to the very foundations of stability in orbit.

3. STABILITY IN SPACE

Given the uncomfortable combination of high dependency and low survivability, one might expect to observe frequent attacks against critical military assets in orbit. However, despite decades of recurring prophesies of impending space war, no such conflict has broken out [14]–[18]. It is true that a handful of space security crises have occurred; most notably, the 2007 Chinese anti-satellite weapon (ASAT) test and the 2008 US ASAT demonstration in response [19]. Moreover, a recent Centre for Strategic and International Studies report suggests increasing interest in attacking US space assets, particularly among the Chinese, Russian, North Korean and Iranian militaries [20]. Overall, however, the space domain has remained puzzlingly peaceful. In this section, we outline three major contributors to this enduring stability: limited accessibility, attributable norms, and environmental interdependence.

A. Limited Accessibility

Space is difficult. Over 60 years have passed since the first Sputnik launch and only nine countries (ten including the EU) have orbital launch capabilities. Moreover, a launch programme alone does not guarantee the resources and precision required to operate a meaningful ASAT capability. Given this, one possible reason why space wars have not broken out is simply because only the US has ever had the ability to fight one [21, p. 402], [22, pp. 419–420].

Although launch technology may become cheaper and easier, it is unclear to what extent these advances will be distributed among presently non-spacefaring nations. Limited access to orbit necessarily reduces the scenarios which could plausibly escalate to ASAT usage. Only major conflicts between the handful of states with 'space club' membership could be considered possible flashpoints. Even then, the fragility of an attacker's own space assets creates de-escalatory pressures due to the deterrent effect of retaliation. Since the earliest days of the space race, dominant powers have recognized this dynamic and demonstrated an inclination towards de-escalatory space strategies [23].

B. Attributable Norms

There also exists a long-standing normative framework favouring the peaceful use of space. The effectiveness of this regime, centred around the Outer Space Treaty (OST), is highly contentious and many have pointed out its serious legal and political shortcomings [24]–[26]. Nevertheless, this *status quo* framework has somehow supported over six decades of relative peace in orbit.

Over these six decades, norms have become deeply ingrained into the way states describe and perceive space weaponization. This *de facto* codification was dramatically demonstrated in 2005 when the US found itself on the short end of a 160-1 UN vote after opposing a non-binding resolution on space weaponization. Although states have occasionally pushed the boundaries of these norms, this has typically occurred through incremental legal re-interpretation rather than outright opposition [27]. Even the most notable incidents, such as the 2007-2008 US and Chinese ASAT demonstrations, were couched in rhetoric from both the norm violators and defenders, depicting space as a peaceful global commons [27, p. 56]. Altogether, this suggests that states perceive real costs to breaking this normative tradition and may even moderate their behaviours accordingly.

One further factor supporting this norms regime is the high degree of attributability surrounding ASAT weapons. For kinetic ASAT technology, plausible deniability and stealth are essentially impossible. The literally explosive act of launching a rocket cannot evade detection and, if used offensively, retaliation. This imposes high diplomatic costs on ASAT usage and testing, particularly during peacetime.

C. Environmental Interdependence

A third stabilizing force relates to the orbital debris consequences of ASATs. China's 2007 ASAT demonstration was the largest debris-generating event in history, as the targeted satellite dissipated into thousands of dangerous debris particles [28, p. 4]. Since debris particles are indiscriminate and unpredictable, they often threaten the attacker's own space assets [22, p. 420]. This is compounded by Kessler syndrome, a phenomenon whereby orbital debris 'breeds' as large pieces of debris collide and disintegrate. As space debris remains in orbit for hundreds of years, the cascade effect of an ASAT attack can constrain the attacker's long-term use of space [29, pp. 295–296]. Any state with kinetic ASAT capabilities will likely also operate satellites of its own, and they are necessarily exposed to this collateral damage threat. Space debris thus acts as a strong strategic deterrent to ASAT usage.

4. THE APPEAL OF THE CYBER-ASAT

The overall effect of cyber-attacks vis-à-vis this strategic stability in space is not well understood. The general need to incorporate cyber risk into satellite mission planning and various legal parallels between the cyber and space commons have attracted some attention [13], [30]. However, cyber weapons in space are often thought of as just one tool among many in the growing ASAT arsenal [31], [32]. In this section, we argue that cyber weapons pose unique strategic threats by undermining the stabilizing dynamics of the *status quo*. Specifically, we contend that cyber-ASATs are accessible, difficult to deter, and environmentally indifferent.

A. Widespread Accessibility

Cyber-attack capabilities are far more widespread than orbital launch technology. In 2017, a former deputy director of the National Security Agency estimated that 'well over 100' countries could harm the US with offensive cyber capabilities [33]. This is over ten times the number of independent spacefaring nations and 50 times the number with proven ASAT technology. Of course, mere possession of cyber capabilities does not guarantee that these can be used against satellites. Nevertheless, this suggests that, for many actors, digital attacks are far more feasible than the creation of national space weapons programmes.

This calculus is further bolstered by the fact that cyber attack capacities which could threaten satellites may apply to other unrelated systems. Thus, even if space is not the primary motivator for cyber-weapons development, one can expect states to cultivate offensive cyber capabilities which can be repurposed for ASAT attacks [34]. Moreover, while the idea of terrorist cells developing orbital spaceflight programmes appears almost comically absurd, even non-state actors have demonstrated sophisticated cyber capabilities [20], [35].

B. Detterence Challenges

International norms influencing cyber combat are both younger and weaker than their space parallels. Scepticism has emerged as to the possibility of ever developing meaningful normative backstops against cyber attacks [36]. Nevertheless, much of the cyber policy community remains optimistic about the eventual cultivation of global norms – a debate which is well beyond the scope of this paper. At present, however, the cyber norms regime has an indisputably worse track record than even the oft-maligned OST.

Moreover, unlike kinetic ASATs, cyber attacks have low risk of attribution and, by extension, low risk of retaliation (and its associated deterrent effect). There has been a great deal of recent debate over the ultimate attributability and deterrability of sophisticated cyber operations [37]–[39]. However, few on either side would contend that cyber attacks are as attributable as the launch of an orbital rocket from sovereign territory. A kinetic ASAT would be noticed and credibly attributed within minutes, but the average data breach evades detection for 200 days, even for critical systems [40]. A cyber-ASAT could lie dormant on target systems for years before triggering at a critical moment. Moreover, this stealth and deniability provides cover for states which publicly encourage the peaceful use of space while they covertly develop ASAT capabilities.

C. Environmental Indifference

Finally, cyber-ASATs undermine the ecological dynamics constraining space weaponization. Actors with cyber-ASAT capabilities may have significantly less strategic dependence on the space environment than the major spacefaring powers. As such, the deterrent effect of collateral damage through space debris would be reduced. Although debris in space can have negative commercial effects on almost all countries, in times of war, this may be an acceptable cost for smaller nations with asymmetric weaknesses. Cyber-ASATs also raise the new spectre of non-destructive ASATs. For example, an exploit which disables or reduces the lifetime of a targeted satellite (e.g. by wasting fuel) could prove environmentally palatable even to states with exposure to space debris.

D. Feasibility of a Cyber-ASAT

In short, cyber-ASATs appear to threaten the foundations of a half-century's stability in orbit. However, premature predictions of instability have become a long-standing tradition in the space policy world. Nearly every major advancement in space technology has been incorrectly heralded as the harbinger of space power's demise. Flawed assumptions about underlying technologies can easily snowball into hyperbolic political strategic theory.

To hold our claims to a higher standard, we have devised a practical case study on the development and use of a cyber-ASAT. In it, we target one aspect of space-flight operations: the collection and use of space situational awareness (SSA). We design and simulate a cyber-attack method that has all three attributes suggested by our strategic analysis. Specifically, our attack uses widely available technology, is stealthy, and minimizes collateral damage. This allows us not only to present the *theoretical* dangers of cyber-ASATs; but to assess their *practical* threat to the *status quo*.

5. SSA: TERRESTRIAL TARGET, CELESTIAL EFFECTS

A. Role of SSA Data

At present, more than 21,000 pieces of orbital debris measuring larger than 10cm in diameter are tracked by the US government [41]. Well over 100 million additional smaller objects are believed to exist but are too small to track reliably. These objects whizz overhead at velocities in excess of 8 km/s and collide at speeds exceeding 10 km/s, meaning that collisions with even miniscule objects can cause catastrophic satellite failures [41].

To safely navigate this ever-growing debris field, operators depend on reliable tracking of orbital hazards. This data is a core component of SSA, which is used by orbital simulation models to predict collisions and inform day-to-day flight control decisions.

Even with modern SSA technologies, collisions still take place. For example, in March 2013, a piece of debris from the 2007 Chinese ASAT test collided with a Russian nanosatellite [42]. Without accurate and reliable SSA data, such incidents would occur far more frequently. In 2017 alone, more than 300,000 potential collision events were identified in US government SSA, 655 of which crossed 'emergency' proximity thresholds for pass distances [43].

B. SSA Data Sources

Although mathematical modelling makes it possible to roughly project orbital motion, complex gravitational and environmental interactions quickly degrade estimates. Reliable SSA data therefore requires frequent observational measurements. The primary sensors employed are radar platforms used in missile defence [44]. This data is supplemented with optical telescopes, ground-based lasers and some space-based observation platforms [44], [45].

The principal constraint on SSA capabilities is often geographic rather than technological. SSA sensors cannot detect objects which do not cross their visible horizon. Large networks of sensors distributed across the planet are thus needed to maintain a complete SSA data repository. This geographic distribution requirement has caused heavy centralization of SSA data into a handful of large repositories.

The Space Surveillance Network (SSN), operated by the US military, is the most widely used and accurate repository. It is believed that only the SSN has global coverage for small objects (~10 cm) [45]. The next closest competitor is the Russian Space Surveillance System, which operates in many former Soviet states and has decent coverage over the northern hemisphere and for larger objects [46]. The Chinese government also operates a network, largely constrained by China's borders [46]. Other networks include the European Space Surveillance System and smaller systems operated by Japan, India, Korea, Canada, Kazakhstan, and Ukraine [44], [46]. Alone, these are unlikely to provide adequate SSA. Commercial SSA products have also begun to emerge, although none offer complete catalogues for objects even 20cm in diameter [45].

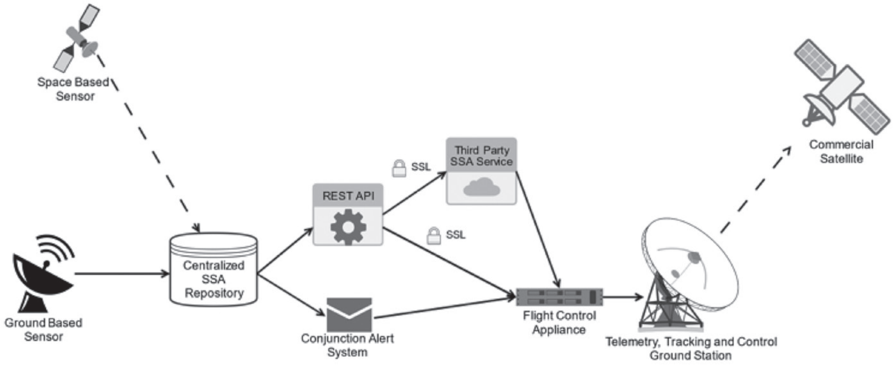
The US freely shares its SSA data through the Space-Track.org platform [47], [48]. Typically, a satellite operator will download SSA from Space-Track and use it to perform conjunction analysis for space missions. Space-Track provides opt-in conjunction alerts and collision avoidance services, but many operators still perform these tasks in-house [47]. Beyond Space-Track, Russia operates a similar scheme through the semi-governmental International Scientific Optical Network (ISON), but usage is far less common [45].

Game-theoretic studies of SSA have demonstrated that these sharing schemes benefit all stakeholders [49]. Intuitively, this makes sense, as the US gains little by concealing SSA data from Russian military operators and causing a collision which would threaten both countries. As a result, a trans-national trust dynamic has emerged around SSA.

C. Value of SSA as Cyber Target

Given that most actors lack the capability to independently verify SSA claims, this trust dynamic is essentially blind. As repositories are highly centralized and hard to verify, a small change to the integrity of the central repository could have massive effects.

FIGURE 1: A NOTIONAL OVERVIEW OF THE SSA DATA FLOW AND POTENTIAL TARGETS.



A cyber attacker might gain access to such repositories through Stuxnet-esque attacks against sensors, direct compromise of centralized databases, modification of data stored at the flight controller's operation centre, exploitation of third-party SSA aggregation services, or alteration of data in transit (Figure 1). Some components of this infrastructure (such as radar sensors or encrypted connections) might require high degrees of sophistication to attack; while others (such as SSA-sharing APIs) may be within the means of most cyber adversaries.

Using this access, an attacker may alter data to effect satellite operator behaviour. For example, an attacker might manipulate an SSA repository to make a near-miss between a debris object and a targeted satellite appear as a collision. This would cause the victim to undertake collision avoidance manoeuvres, shortening the satellite's lifetime through fuel wastage. The reverse attack could also be executed, where an attacker conceals a projected collision and destroys the targeted satellite, all without launching a single rocket.

In essence, SSA exploitation elevates simple integrity compromises into Cyber-ASAT capabilities. Furthermore, the fuel wastage attack scenario does not threaten collateral debris damage. As such, an attack against SSA data meets all three design objectives outlined in section 3.

6. CASE STUDY: SIMULATING ATTACKS AGAINST SSA

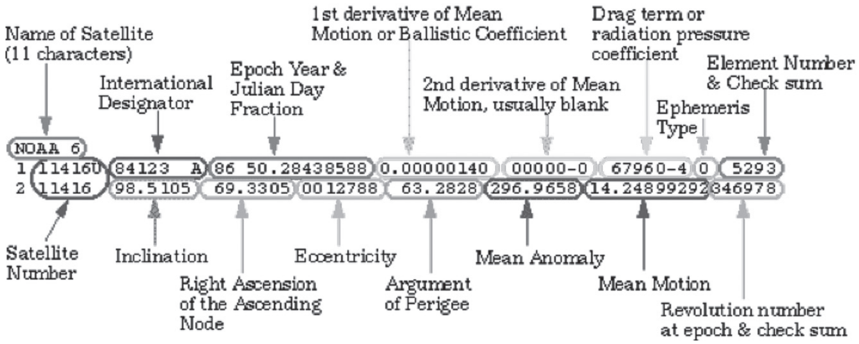
A. Experimental Design and Assumptions

We elected to assess the technical feasibility of attacks on SSA repositories through simulations with a commercial spaceflight planning tool [50].

The simulated attacker's overall objective was to cause an arbitrary satellite in Low Earth Orbit to take unnecessary collision-avoidance manoeuvres over the next 72 hours (the current SSN emergency notification threshold). We assumed that our attacker wished to be stealthy and that significant modification of SSA data (such as the creation of new debris objects) would be detected. Finally, we granted that the attacker had already obtained the ability to modify data through traditional cyber exploitation techniques (e.g. malware installed on the SSA web servers).

Target data was assumed to be in the widely used two-line element (TLE) format (Figure 2). This format is used to distribute projections from Space-Track.org. The format was originally designed to fit on two 80-column punch cards; no security features or significant revisions have been made since its adoption by NORAD in the 1970s [51].

FIGURE 2: THE TLE EPEHEMERIS DATA FORMAT [52]. STARRED PARAMETERS ARE TARGETED BY OUR ATTACK.


1st derivative of Mean Motion or Ballistic Coefficient
Drag term or radiation pressure coefficient
Name of Satellite (11 characters)
Element Number & Check sum
Epoch Year & Julian Day Fraction
International Designator
2nd derivative of Mean Motion, usually blank
Ephemeris Type
NOAA 6
1 11416U 84123 A 86 50.28438588 0.00000140 00000-0 67960-4 0 5293
2 11416 98.5105 69.3305 0012788 63.2828 296.9658 14.24899292346978
Satellite Number
Inclination
Eccentricity
Mean Anomaly
Right Ascension of the Ascending Node
Argument of Perigee
Mean Motion
Revolution number at epoch & check sum

The simulations themselves were built using real-world data from the US SSN. Projections were propagated with the SGP4 propagator provided by Air Force Space Command and recommended for usage with TLE data [53].

B. Attack Method

Our proposed attack consists of three stages: acquisition, perturbation, and generation. In the acquisition phase, five 'near-miss' debris objects are selected as candidates for potential tampering. In the perturbation phase, the SSA data describing these objects are strategically altered to artificially cause a collision projection. Finally, in the generation stage, these alterations are merged with authentic data to create a falsified TLE entry for insertion into the SSA repository.

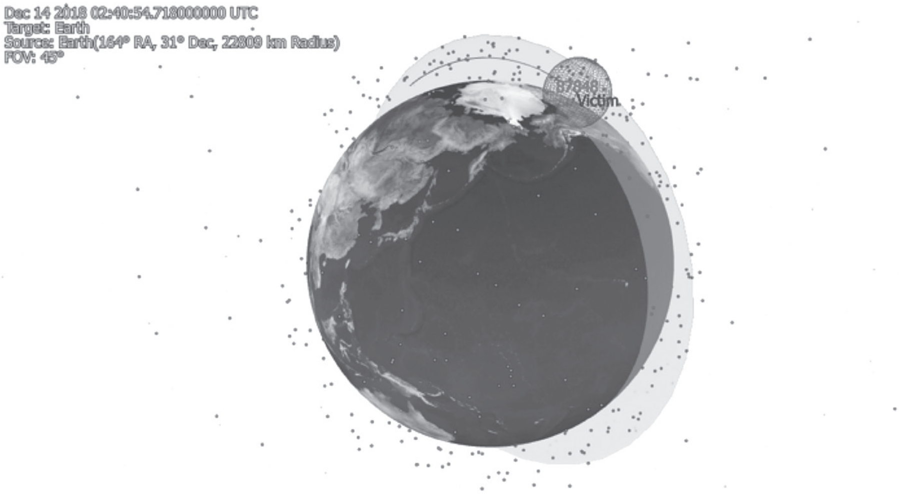
a) Acquisition stage

To begin, an attacker must provide accurate TLEs characterizing a victim satellite's orbit and any debris objects to be considered. This information is readily available online.

Our attack tool automatically synchronizes these TLEs to a common starting epoch. From this epoch, the debris objects and victim satellite are propagated to project their locations over a simulated 72-hour period, subdivided into 10-second intervals.

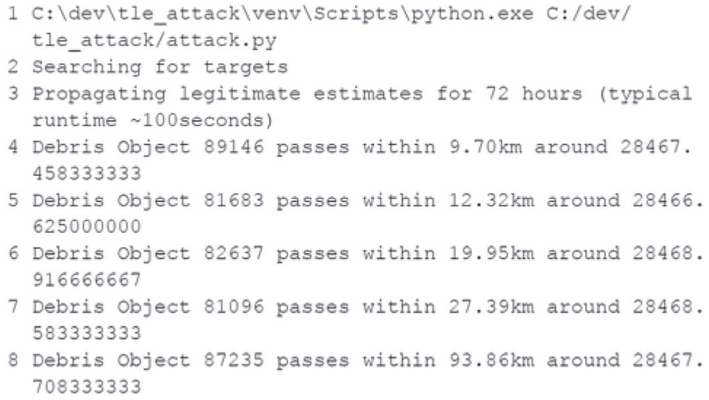
At each interval, a three-step filter is employed to remove irrelevant debris objects (Figure 3). First, we select only debris objects currently inside the victim satellite's orbit plane (represented by a 100km deep cylinder, centred at the Earth's core and oriented along the victim's orbit). Second, we remove debris with altitudes outside a range bounded by the victim satellite's perigee (lowest orbital altitude) and apogee (highest orbital altitude). Third, we remove debris objects more than 1000km away from the victim satellite in any direction.

FIGURE 3: THE THREE-STEP DEBRIS FILTER. DEBRIS OBJECT 87848 HAS JUST ENTERED A 1000KM SPHERE CENTERED ON THE VICTIM SATELLITE.


Dec 14 2018 02:40:54.718000000 UTC
Target: Earth
Source: Earth(164° RA, 31° Dec, 22809 km Radius)
FOV: 45°
87848
Victim

For any debris which survive this filtering, we calculate the time and distance of closest approach to the victim over a full orbital period. Ultimately, the five objects which pass closest over the whole 72-hour window are selected (as in Figure 4). TLE data for these objects is passed on to the perturbation stage along with times of their closest approaches.

FIGURE 4: TYPICAL ACQUISITION STAGE OUTPUT.



```
C:\dev\tle_attack\venv\Scripts\python.exe C:/dev/
tle_attack/attack.py
Searching for targets
Propagating legitimate estimates for 72 hours (typical
runtime ~100seconds)
Debris Object 89146 passes within 9.70km around 28467.
458333333
Debris Object 81683 passes within 12.32km around 28466.
625000000
Debris Object 82637 passes within 19.95km around 28468.
916666667
Debris Object 81096 passes within 27.39km around 28468.
583333333
Debris Object 87235 passes within 93.86km around 28467.
708333333
```

b) Perturbation Stage

In the perturbation stage, TLEs of the five selected debris objects are altered with the goal of reducing the projected nearest pass distance to the target to less than 1km. This is based on Air Force Space Command guidance that TLEs can be considered accurate to approximately 1km of precision. Any object which passes within this range could thus trigger an anticipated conjunction.

In order to reduce the risk of detection, two further constraints are imposed. First, only four TLE fields (along with the TLE checksum) are subject to modification. Moreover, these fields are altered within certain boundaries (detailed in Table 1). To our knowledge, no study has investigated to what extent, if any, satellite operators vet SSA data for anomalies. As such, these boundaries were selected arbitrarily based on the overall precision of the TLE format (also detailed in Table 1). Decreasing these bounds lowers the chance of detection but increases computational complexity.

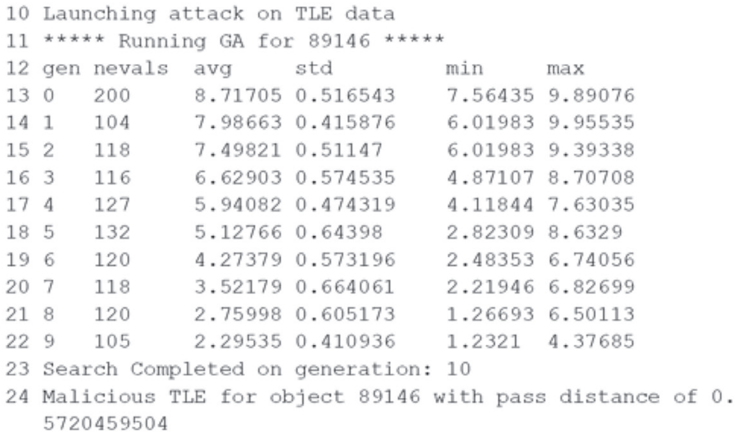
TABLE 1: MODIFIED TLE FIELDS AND BOUNDARIES

TLE Field	Maximum Alteration	TLE Precision
Orbital Inclination	± .1 degrees	.0001 degrees
Right Ascension of the Ascending Node	± .1 degrees	.0001 degrees
Eccentricity	± .01	.0000001
Argument of Perigee	± .1 degrees	.0001 degrees

SGP4, like most orbital projection models, is complex; the overall effect of any given modification over a 72-hour window is non-trivial. However, we can greatly reduce this complexity by recognizing that there is no need to find the *optimal* perturbation set, but rather only an *adequate* set to cause a collision.

This realization allows us to employ a rudimentary genetic algorithm, where we treat the TLE fields themselves as genetic features. Our model's fitness is simply the minimization of nearest pass distance; our initial population size is arbitrarily set to 200 individuals. Over a span of up to 40 generations, each individual is used to generate a fake TLE and propagated for the 3-hour period surrounding the debris object's closest approach (Figure 5). Once a sub-1km pass is found, this result is passed along to the generation stage.

FIGURE 5: TYPICAL PERTURBATION STAGE OUTPUT. IN THIS CASE, A SET OF MODIFICATIONS WAS DETECTED THAT CAUSED DEBRIS OBJECT 89146 TO PASS WITHIN 600M OF THE VICTIM SATELLITE.



```
Launching attack on TLE data
***** Running GA for 89146 *****
gen nevals  avg     std         min     max
0   200     8.71705 0.516543    7.56435 9.89076
1   104     7.98663 0.415876    6.01983 9.95535
2   118     7.49821 0.51147     6.01983 9.39338
3   116     6.62903 0.574535    4.87107 8.70708
4   127     5.94082 0.474319    4.11844 7.63035
5   132     5.12766 0.64398     2.82309 8.6329
6   120     4.27379 0.573196    2.48353 6.74056
7   118     3.52179 0.664061    2.21946 6.82699
8   120     2.75998 0.605173    1.26693 6.50113
9   105     2.29535 0.410936    1.2321  4.37685
Search Completed on generation: 10
Malicious TLE for object 89146 with pass distance of 0.
5720459504
```

Our naïve genetic algorithm may be further optimized. It is likely that a generalized approach, which does not rely on genetic algorithms at all, may be found. However, the operational benefit of finding a pass within 10m versus a pass within 900m is minimal, since both fall within the collision detection radius. Further, given that an attacker has hours, if not days, to calculate these modifications, computational efficiency is far from vital.

c) Generation Stage

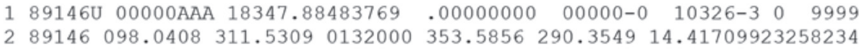
In the generation stage, the results of the five genetic algorithm runs may be compared using two further metrics:

- The proximity of the projected pass caused by a malicious TLE
- The overall magnitude of modifications introduced into a malicious TLE.

The first metric is useful for an attacker who wishes to have the highest likelihood of causing a satellite manoeuvre. The second metric would be more desirable for attackers seeking to minimize the risk of detection. An attacker can also ignore these metrics and simply select the first valid attack found to minimize search time.

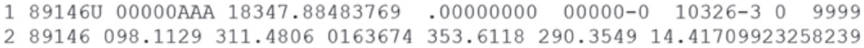
Once a malicious TLE parameter set has been found, its modifications are merged with data from the original debris TLE (as in Figure 6). The result of this process is a new TLE which can be inserted into the SSA database by an attacker as required, completing the attack (Figure 7).

FIGURE 6: A TYPICAL ORIGINAL TLE.



```
1 89146U 00000AAA 18347.88483769  .00000000  00000-0  10326-3 0  9999
2 89146 098.0408 311.5309 0132000 353.5856 290.3549 14.41709923258234
```

FIGURE 7: A TYPICAL MALICIOUS TLE.



```
1 89146U 00000AAA 18347.88483769  .00000000  00000-0  10326-3 0  9999
2 89146 098.1129 311.4806 0163674 353.6118 290.3549 14.41709923258239
```

C. Attack Simulation

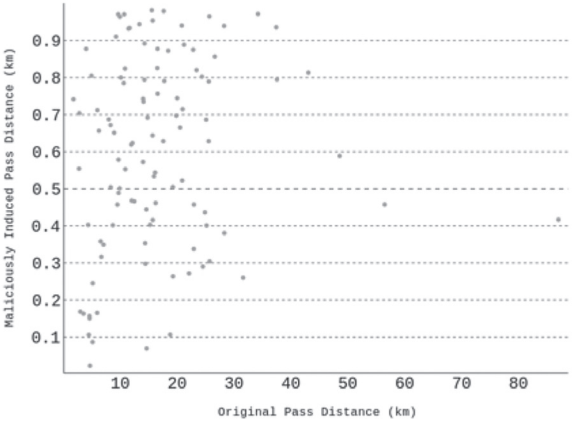
To test this approach experimentally, we simulated attacks against each of 111 satellites in the Iridium constellation. Iridium is a commercial communications service with over one million satellite customers [54]. The network's largest customer is the US Defense Information Systems Agency [55]. For our debris field, we selected 529 objects from Space-Track.org's 'Well-Tracked Analyst Objects of Unknown Origin' dataset [48]. Prior to launching our attack, none of the Iridium satellites were projected to pass within 1km of these objects over a 72-hour window.

In order to simulate attacks against many satellites quickly, we enforced no optimizations in the 'generation' phase. This means that our experiment represents the worst case scenario for our method in terms of pass distance and stealth.

Our technique successfully generated collision events for more than 93% of the Iridium constellation. On average, it took about 12 genetic generations to find a valid attack; the total attack runtime for each object averaged a little over 6 minutes on consumer grade hardware.

Although we accepted any pass under 1km, the mean pass distance of our attack parameters was around 600m and the minimum only 2m. No obvious correlation between original pass distance and malicious pass distance was observed (Figure 8). This suggests that more restrictive boundaries and more demanding proximity requirements are obtainable using this general approach.

FIGURE 8: ORIGINAL AND MALICIOUS PASS DISTANCES FROM THE IRIDIUM ATTACK SIMULATION.



Our findings demonstrate that, once an attacker has compromised the integrity of an SSA repository, elevating this to ASAT capability is quite feasible. With consumer grade hardware and a minimally optimized attack method, we falsified collision projections for over 100 real-world satellites used by the world's largest militaries.

7. CONCLUSION

In this paper, we have argued that the free pursuit of space power has been facilitated by structural features of the space domain. Specifically, we isolated three key features: limited accessibility, attributable norms, and environmental interdependence. We theorized that cyber-attacks can undermine all three of these dynamics and thus pose a structural threat to the long-standing peace in orbit.

To assess these theoretical claims, we designed a cyber-ASAT capability, targeting space situational awareness. Our cyber-ASAT was built using widely accessible technologies and minimized both the risk of attribution and collateral damage. This cyber-ASAT was tested in orbital spaceflight simulations and successfully attacked 93% of the strategically vital Iridium satellite constellation, all without firing a single rocket.

Our experimental findings suggest that the rise of cyber-ASATs is not merely a distant technological spectre, but rather a real and present danger. Satellite operators and the states who rely upon them must assess the risks of 'blind trust' information-sharing relationships and, more broadly, the overall cyber-security profile of these systems.

This paper considers only one demonstrative example among many plausible mechanisms for cyber-ASAT capabilities. Future work considering vectors such as on-board malware, compromise of satellite control telemetry, sensor injection, and signal hijacking may help to further characterize this emerging domain. Additionally, there is a clear need for research into defensive mechanisms which prevent such attacks. For example, a statistical approach to anomaly detection in SSA datasets may prove useful in this case. Such research to defend satellites from Cyber-ASATs will be a vital prerequisite for the continued exercise of space power.

Acknowledgments

We extend thanks to a.i. solutions, Inc. for providing research access to the FreeFlyer astrodynamics software used in our spaceflight simulations. J. P. also thanks the Rhodes Trust for funding in support of this research; and the Satellite Applications Catapult, for providing background information on spaceflight operations.

REFERENCES

[1] A. Wilson, *The Culture of Nature: North American landscape from Disney to the Exxon Valdez*. Between the Lines, 1991.

[2] S. P. Anson and D. Cummings, ‘The First Space War: The contribution of satellites to the Gulf War’, *RUSI J.*, vol. 136, no. 4, pp. 45–53, Dec. 1991.

[3] S. Lambakis, ‘Space Control in Desert Storm and Beyond’, *Orbis*, vol. 39, no. 3, pp. 417–433, Jun. 1995.

[4] Y. Fukushima, ‘Debates over the Military Value of Outer Space in the Past, Present and the Future: Drawing on Space Power Theory in the US’, *NIDS J. Def. Secur*., pp. 35–48, 2013.

[5] L. Greenemeier, ‘GPS and the World’s First “Space War”’, *Scientific American*, 8 Feb 2016. [Online]. Available: https://www.scientificamerican.com/article/gps-and-the-world-s-first-space-war/. [Accessed: 17-Dec-2018].

[6] T. Brown, ‘Space and the Sea: Strategic considerations for the commons’, *Astropolitics*, vol. 10, no. 3, pp. 234–247, Dec. 2012.

[7] B. E. Bowen, ‘British Strategy and Outer Space: A missing link?,” *Br. J. Polit. Int. Relat*., vol. 20, no. 2, pp. 323–340, May 2018.

[8] D. Paikowsky, *The Power of the Space Club*. Cambridge University Press, 2017.

[9] C. B. Halstead, ‘The Ultimate High Ground - U.S. intersector cooperation in outer space’, *J. Air Law Commer*., vol. 81, pp. 595–610, 2016.

[10] K. Pollpeter, ‘Space, the new domain: Space operations and Chinese military reforms’, *J. Strateg. Stud*., vol. 39, no. 5–6, pp. 709–727, Sep. 2016.

[11] D. Koplow, ‘ASAT-isfaction: Customary international law and the regulation of anti-satellite weapons’, *Georget. Law Fac. Publ. Works*, Jan. 2009.

[12] D. E. Lupton, *On Space Warfare: A space power doctrine*. PN, 1988.

[13] Chatham House, ‘Making the Connection: The future of cyber and space’, London, International Security Workshop Seminar, Jan. 2013.

[14] D. J. St. James, ‘The Legality of Antisatellites’ Recent Development’, *Boston Coll. Int. Comp. Law Rev.*, vol. 3, pp. 467–494, 1980 1979.

[15] B. Jasani and C. Lee, *Countdown to Space War*. Taylor & Francis, 1984.

[16] S. J. Bruger, “Not Ready for the ‘First Space War,’ What about the second?’ Naval War Coll Newport RI Dept of Operations, May 1993.

[17] J. E. Hyten, ‘A sea of peace or a theater of war? Dealing with the inevitable conflict in space’, *Air Space Power J*., vol. 16, no. 3, p. 78, 2002.

[18] V. Anantatmula, ‘U.S. Initiative to Place Weapons in Space: The catalyst for a space-based arms race with China and Russia’, *Astropolitics*, vol. 11, no. 3, pp. 132–155, Sep. 2013.

[19] M. A. Gubrud, ‘Chinese and US Kinetic Energy Space Weapons and Arms Control’, *Asian Perspect*., vol. 35, no. 4, pp. 617–641, 2011.

[20] T. Harrison, K. Johnson, and T. Roberts, ‘Space Threat Assessment 2018’, 2018.

[21] N. Tannenwald, ‘Law versus Power on the High Frontier: The case for a rule-based regime for outer space’, *Yale J. Int. Law*, vol. 29, pp. 363–422, 2004.

[22] R. Handberg, ‘Is Space War Imminent? Exploring the possibility’, *Comp. Strategy*, vol. 36, no. 5, pp. 413–425, Oct. 2017.

[23] P. Stares, ‘Space and US National Security’, *J. Strateg. Stud*., vol. 6, no. 4, pp. 31–48, Dec. 1983.

[24] S. Freeland, ‘Peaceful Purposes - Governing the military uses of outer space’, *Eur. J. Law Reform*, vol. 18, pp. 35–51, 2016.

[25] J. A. Urban, ‘Soft Law: The key to security in a globalized outer space’, *Transp. Law J*., vol. 43, pp. 33–50, 2016.

[26] P. Meyer, ‘Dark Forces Awaken: The prospects for cooperative space security’, *Nonproliferation Rev*., vol. 23, no. 3–4, pp. 495–503, Jul. 2016.

[27] F. Grimal and J. Sundaram, ‘The Incremental Militarization of Outer Space: A threshold analysis’, *Chin. J. Int. Law*, vol. 17, no. 1, pp. 45–72, Mar. 2018.

[28] B. Gill and M. Kleiber, ‘China’s Space Odyssey: What the antisatellite test reveals about decision-making in Beijing’, *Foreign Aff.*, vol. 86, no. 3, pp. 2–6, 2007.

[29] J. Moltz, *The Politics of Space Security: Strategic restraint and the pursuit of national interests, Second edition*. Redwood City, United States: Stanford University Press, 2014.

[30] C. Baylon, ‘Challenges at the Intersection of Cyber Security and Space Security’, *Int. Secur.*, 2014.

[31] Z. Shabbir and A. Sarosh, ‘Counterspace Operations and Nascent Space Powers’, *Astropolitics*, vol. 16, no. 2, pp. 119–140, Aug. 2018.

[32] B. L. Triezenberg, ‘Deterring Space War: An exploratory analysis incorporating prospect theory into a game theoretic model of space warfare’, Rand Corporation, Santa Monica, CA, Product Page, 2017.

[33] M. Levine, ‘Russia Tops List of Countries that could Launch Cyberattacks on US’, *ABC News*, 19-May-2017. [Online]. Available: https://abcnews.go.com/US/russia-tops-list-100-countries-launch-cyberattacks-us/story?id=47487188. [Accessed: 18-Dec-2018].

[34] M. Smeets, ‘The Strategic Promise of Offensive Cyber Operations’, *Strateg. Stud. Q*., vol. 12, no. 3, pp. 90–113, 2018.

[35] J. Sigholm, ‘Non-State Actors in Cyberspace Operations’, *J. Mil. Stud.*, vol. 4, no. 1, pp. 1–37, Dec. 2013.

[36] A. Grigsby, ‘The End of Cyber Norms’, *Survival*, vol. 59, no. 6, pp. 109–122, 2017.

[37] T. Rid and B. Buchanan, ‘Attributing Cyber Attacks’, *J. Strateg. Stud*., vol. 38, no. 1–2, pp. 4–37, Jan. 2015.

[38] J. R. Lindsay, ‘Tipping the Scales: The attribution problem and the feasibility of deterrence against cyberattack’, *J. Cybersecurity*, vol. 1, no. 1, pp. 53–67, Sep. 2015.

[39] N. Tsagourias, ‘Cyber attacks, self-defence and the problem of attribution’, *J. Confl. Secur. Law*, vol. 17, no. 2, pp. 229–244, Jul. 2012.

[40] B. I. Koerner, ‘Inside the OPM Hack, the cyberattack that shocked the US government’, *Wired*, 23-Oct-2016.

[41] NASA, ‘ARES: Orbital Debris Program Office Frequently Asked Questions’, 2018. [Online]. Available: https://orbitaldebris.jsc.nasa.gov/faq.html#3. [Accessed: 10-Dec-2018].

[42] L. David, ‘Russian Satellite Hit by Debris from Chinese Anti-Satellite Test’, *Space.com*, 08-Mar-2013. [Online]. Available: https://www.space.com/20138-russian-satellite-chinese-space-junk.html. [Accessed: 10-Dec-2018].

[43] D. Mosher, ‘The US Government Logged 308,984 Potential Space-Junk Collisions in 2017 — and the problem could get much worse’, *Business Insider*, 15-Apr-2018. [Online]. Available: https://www.businessinsider.com/space-junk-collision-statistics-government-tracking-2017-2018-4. [Accessed: 10-Dec-2018].

[44] B. Weeden, ‘Global Space Situational Awareness Sensors’, Sep. 2010.

[45] B. Lal, A. Balakrishnan, B. Caldwell, R. Buenconsejo, and S. Carioscia, ‘Global Trends in Space Situational Awareness and Space Traffic Management’, Apr. 2018.

[46] D. A. Vallado and J. D. Griesbach, ‘Simulating Space Surveillance Networks’, Paper AAS 11-580 presented at the AAS/AIAA Astrodynamics Specialist Conference. Jul. 2011.

[47] D. Bird, ‘Sharing Space Situational Awareness Data’, Strategic Command Offutt AFB NE, Sep. 2010.

[48] JFSCC, ‘SSA Sharing & Orbital Data Requests’, *Space-Track.org*, 2018. [Online]. Available: https://www.space-track.org/documentation#/odr. [Accessed: 10-Dec-2018].

[49] N. Shah, M. Richards, D. Broniatowski, J. Laracy, P. Springmann, and D. Hastings, ‘System of Systems Architecture: The case of space situational awareness’, in *AIAA Space 2007 Conference & Exposition*, 0 vols., American Institute of Aeronautics and Astronautics, 2007.

[50] ai-solutions, *FreeFlyer® Software*. 2018.

[51] C. Früh and T. Schildknecht, ‘Accuracy of Two-Line-Element Data for Geostationary and High-Eccentricity Orbits’, *J. Guid. Control Dyn*., vol. 35, no. 5, pp. 1483–1491, 2012.

[52] NASA, ‘Human Space Flight (HSF) - Realtime data’, *NASA SkyWatch*, 2011. [Online]. Available: https://spaceflight.nasa.gov/realdata/sightings/SSapplications/Post/JavaSSOP/SSOP_Help/tle_def.html. [Accessed: 20-Dec-2018].

[53] Air Force Space Command, ‘Astrodynamic Standards Software’, *Air Force Space Command*, 22 Mar 2017. [Online]. Available: https://www.afspc.af.mil/About-Us/Fact-Sheets/Display/Article/249006/astrodynamic-standards-software/. [Accessed: 11-Dec-2018].

[54] F. Johnson, ‘1 Million Subscribers Connected: Iridium helps prevent shark attacks while protecting local ecosystems’, *Iridium Satellite Communications*, 18 Jun. 2018.

[55] P. Selding, U.S. Defense Agency Encourages Allied Nations to Join Unlimited-Use Iridium Program’, *SpaceNews.com*, 11 Nov. 2016. [Online]. Available: https://spacenews.com/u-s-defense-agency-encourages-allied-nations-to-join-unlimited-use-iridium-program/. [Accessed: 14-Dec-2018].

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Challenges and Opportunities to Counter Information Operations Through Social Network Analysis and Theory

Alicia Bargar, MS
Data Scientist
Asymmetric Operations
Johns Hopkins Applied Physics Laboratory
Laurel, Maryland, United States
alicia.bargar@jhuapl.edu

Janis Butkevics, MS
Data Scientist
Asymmetric Operations
Johns Hopkins Applied Physics Laboratory
Laurel, Maryland, United States
janis.butkevics@jhuapl.edu

Stephanie Pitts, PhD
Social Scientist
Asymmetric Operations
Johns Hopkins Applied Physics Laboratory
Laurel, Maryland, United States
stephanie.pitts@jhuapl.edu

Ian McCulloh, PhD
Chief Data Scientist
Accenture Federal Services
Accenture
Arlington, Virginia, United States
ian.mcculloh@accenturefederal.com

Abstract: Information operations on social media have recently attracted the attention of media outlets, research organizations and governments, given the proliferation of high-profile cases such as the alleged foreign interference in the 2016 US presidential election. Nation-states and multilateral organizations continue to face challenges while attempting to counter false narratives, due to lack of familiarity and experience with online environments, limited knowledge and theory of human interaction with and within these spaces, and the limitations imposed by those who own and maintain social media platforms. In particular, these attributes present unique difficulties for the identification and attribution of campaigns, tracing information flows at scale, and

* This work was funded by the Department of the Navy, Office of Naval Research under ONR Grant No. N00014-18-1-2128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

identifying spheres of influence. Complications include the anonymity and competing motivations of online actors, poorly understood platform dynamics, and the sparsity of information regarding message transferal across communication platforms.

We propose that the use of social network analysis (SNA) can aid in addressing some of these challenges. We begin by providing a brief explanation of the field and its utility in understanding online communications. We discuss how theories drawn from SNA, which seek to make statistical inferences about relationships and information transfer, can be applied to the information operations domain. Specifically, we will focus on how current research in social influence, information diffusion, and cluster analysis can be immediately applied and identify opportunities for future research. We then demonstrate how these analytic techniques can work in practice, utilizing multiple online communication datasets. Finally, we conclude by discussing how the use of these methods can lead to the development of tactical approaches countering misinformation campaigns.

Keywords: *information operations, social network analysis, influence operations, information diffusion*

1. INTRODUCTION

A. Information Operations

Recent events require us to reconsider the role of information operations in modern conflict. The online infrastructure that facilitates civilian communication and organization also provides adversaries with new-found capabilities for exerting influence and disrupting democratic processes. Despite familiarity with information operations (IO) at a strategic level, adversaries' presence in the online environment and the intermingling between different actors complicates the development of countermeasures. How do we disrupt information campaigns without impacting civilian rights?

This paper does not promise universal solutions. Instead, we address the space between policy and practice, drawing upon current social network analysis and theory (SNA/T) research to propose alternative methodologies that can be used when detecting, analysing, and countering IO. The field of social network analysis (SNA) has developed theories and methods for understanding how humans relate, communicate, and spread information. Its relevance for understanding online social phenomena has cast the field into the spotlight. Although the application to IO is novel, SNA's study of the communication channels upon which IO relies makes it a natural fit.

Definitions of IO vary widely. Military descriptions, like those of NATO and the US Department of Defense, figure most prominently. In JP-313, the US military describes how using information-based systems can "… influence, disrupt, corrupt, or usurp the decision making of adversaries [1]". However, United Nations peacekeeping operations, like the 1999 operation in Kosovo [2], similarly invoke information campaigns to spread awareness and influence in "struggles for control over information identifiable in situations of conflict" [3]. These operations differ due to their alternative objectives and potential lack of adversary. Alternatively, the Canadian Forces' nation-state policy focuses less on assertive actions and more on peacetime strategies to: "deter conflict, protect… information and information systems, and [shape] the information environment" [4, 5].

In this report, we focus specifically on the deterrence of adversarial information campaigns. For clarity, we follow NATO's definition, which describes IO as "military information activities [that] create desired effects on the will, understanding, and capability of adversaries, potential adversaries, and other [North Atlantic Council] approved parties" [6].

We also avoid the term 'information warfare.' Offensive activities with national or international significance can be conducted by non-state groups, criminal organizations, or individuals for personal or economic benefit [7]. We thus choose the term 'operations,' which reflects the complexity of the online environment without implying a nation-state origin.

B. Challenges of the Online Environment

Online domains and social media have become platforms for advancing state-sponsored information campaigns. Most famously, Russian-backed accounts posed as US citizens to spread information prior to the 2016 presidential election [8]. The transferral of IO to the online domain introduces new complexities for developing countermeasures. The attributes below illustrate the unique challenges of the Internet.

1) Anonymity

Identifying information sources online remains difficult. People and organizations obscure their identity for purposes like fun, whistleblowing, trolling, criminal activity, and astroturfing [9, 7].

2) Ease of Coordination

The Internet enables people with similar interests, desires, or beliefs to coordinate more easily. This heightened capacity for ordinary civilians, communities, or organizations to mobilize at a national or international level creates a new social dynamic that is still not fully understood.

3) Virality

The online environment enables the rapid spread and evolution of information. The fast-paced, global spread of information online makes rumour containment challenging.

4) Multi-stakeholder Governance

The Internet's governance structure reduces state power online. Privacy laws [10], private domain limitations, and individuals' rights to counter government statements online illustrate some considerations that states must take when attempting to gain situational awareness or exert influence.

C. Why Use Social Network Analysis?

Social network analysis studies the underlying patterns of relationships and communications using models known as 'networks.' Network models enable us to address questions such as: 'What communities exist? How does information spread? What is a group's organizational structure?' To answer these questions and others, we combine an understanding of 'relational statistics' [11] with methodologies that ground research in social theories on the variables that influence behaviour.

The advent of the Internet and accompanying datasets inspired new computational techniques that apply SNA to large-scale social systems. As a result, there exists an expansive body of work that utilizes SNA approaches to map out communities, information flows and key actors in online environments. Relevant studies for countering information campaigns include network-based interventions for behaviour change [12], methods for identifying influential information sources [13], and approaches for identifying organizational structures of covert groups [14, 15]. In the next section, we will explain these aspects of SNA to show how they can enhance analytic processes for identifying and countering IO.

Note that SNA alone is not sufficient to develop counter-IO tactics. SNA characterizes content dissemination but not the content itself. Regardless, social networks can help illuminate the social influences and forces present that may spread or contain an IO.

2. COUNTERING ONLINE IO WITH SOCIAL NETWORK ANALYSIS

Scholars, policymakers, and members of the private and public sectors have debated varying measures to counter online disinformation (as defined in [12]). Using these resources, we propose the following linkages between identified needs and SNA/T contributions (Table 1). For the remainder of this section, we discuss each contribution alongside illustrations from relevant work.

TABLE 1: LINKAGES BETWEEN THE EUROPEAN COMMISSION'S HEG REPORT ON COUNTERING DISINFORMATION, NATO IO DOCTRINE, AND SNA/T RESEARCH.

Need [12]	Action [6]	SNA/T Contribution
Identify campaigns	Detect	Anomaly Detection
Monitor scale, techniques, tools, nature, impact of disinformation	Probe	Network Metrics
Identify and map sources and mechanisms	Expose, Deter	Attribution Strategies
Safeguard diversity and sustainability	Protect, Safeguard, Support	Influence Analysis
Counter IO efforts	Disrupt/Diminish/Negate/Prevent	Network Intervention

A. Anomaly Detection

Ruses, stratagems, deceits, camouflages and tricks are as old as war itself and their use… is written in the mists of time. – Paul Villatoux, translated [5]

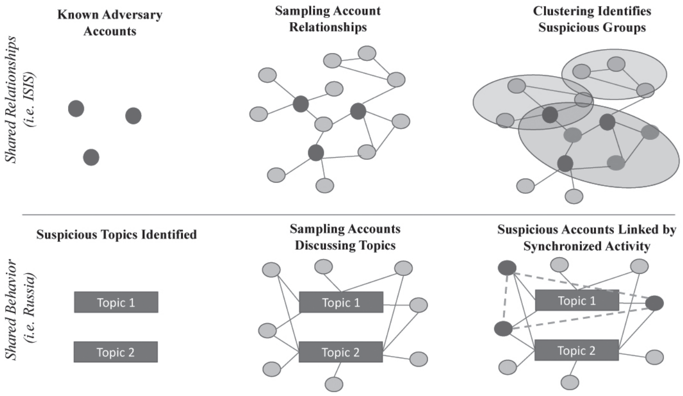
Identifying where information campaigns exist is a critical first step in the countering process. The frequency and velocity of online discussion makes this a non-trivial task considering the variety of ongoing 'influence' campaigns including product marketing, legitimate political efforts and various organic viral content. Malicious IO campaign detection must both identify the various campaigns and determine which are hostile.

The ability to characterize the interactions between online actors and their intended audiences makes SNA a common tool for information campaign detection [13, 14, 15]. Two popular examples are the Islamic State of Syria (ISIS) online recruitment campaigns and the Russian Federation interference in United States (US) elections. Note that Russian interference is not limited to only US elections, but US elections are a common topic of research and data.

The ISIS online recruitment campaign was a novel approach to manpower sourcing by a terrorist organization [16]. ISIS strongly relied on Twitter to spread propaganda and initialize recruitment across the world. Given ISIS's relatively unique messaging and tactics, SNA was heavily leveraged [13] for identifying ISIS users on Twitter. ISIS recruiters and propagandists were identified as seed actors, and users who interacted with their accounts were collected. While many users collected in such a manner had no relation to ISIS, the groups of ISIS supporters and non-ISIS Twitter users could be separated into communities through clustering, which is an SNA approach of grouping users into communities based on the attributes of their interactions such as frequency, similarity of connections and other metrics.

Russian influence campaigns opportunistically leverage world events to promote a diversity of objectives. Their use of both human and bot activity allows influence campaigns to scale with large 'astroturf' bot campaigns or targeted posts by humans [8]. Identifying this opportunistic targeting requires different approaches, such as the detection of synchronized actions [17] that appear to focus on a single topic, set of keywords or hashtags, or users. Prominent topic(s) or individual(s) in online discussion can be identified through various SNA metrics such as degree centrality measures, density, or clustering algorithms [14]. Figure 1 illustrates the differences between this approach and the one to identify ISIS accounts.

FIGURE 1. PROCEDURES TO IDENTIFY SUSPICIOUS ACCOUNTS IN ISIS AND RUSSIAN CAMPAIGNS.


Shared Relationships (i.e. ISIS)
Known Adversary Accounts
Sampling Account Relationships
Clustering Identifies Suspicious Groups
Shared Behavior (i.e. Russia)
Suspicious Topics Identified
Sampling Accounts Discussing Topics
Suspicious Accounts Linked by Synchronized Activity
Topic 1
Topic 2

SNA techniques can effectively detect change and time of change in networks based on stable relationships between accounts or group-level connections [18, 19]. Another common method for both ISIS and Russian-like campaigns is to pair SNA with machine learning (ML) methods to build systems for automated and possibly near-real-time campaign detection. SNA metrics are coupled with other features like post timing, content analysis and user-specific measures that are then fed into ML models to mine interactions and unique characteristics of the specific campaigns [13, 14, 15, 20].

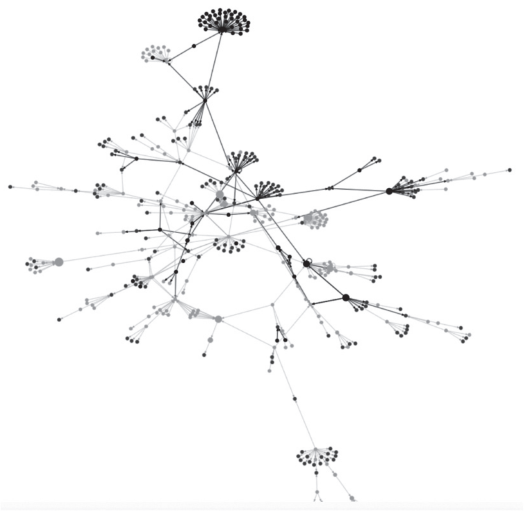
While joint SNA and ML methods have shown capability to rapidly detect malicious information campaigns, the adversaries continue to adjust tactics, techniques, and procedures (TTPs) to elude them. Along with changing adversary tactics, the limits of available data often curtail effective analysis. Many papers demonstrate detection capabilities on Twitter data, which is relatively easy to obtain. However, data from more secure and private platforms such as Facebook and Instagram are scarce. Furthermore, capabilities to map content and actors across online platforms are in early stages and, therefore, detecting cross-platform information campaigns is currently limited.

B. Network Metrics

Probe: to examine closely in order to evaluate a system or entity to gain an understanding of its general layout and/or perception. - Allied Joint Doctrine for Information Operations

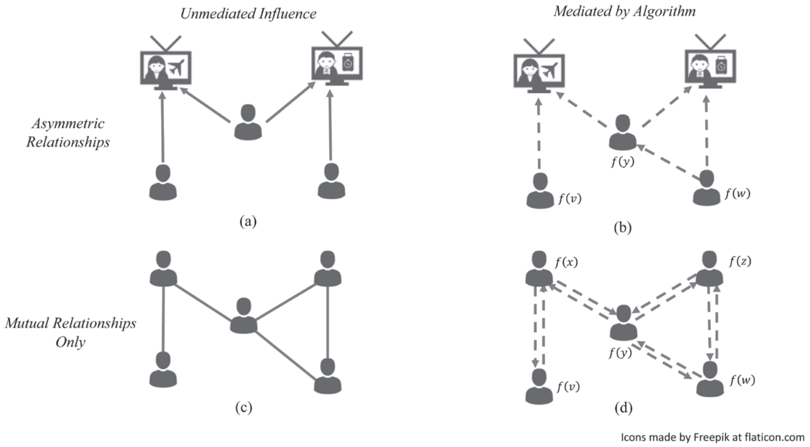
By finding ways to compare online campaigns, we can begin to build a strategic framework. SNA measures have been applied to study covert network organization [21] and may serve a similar purpose for comparing IO campaign structures. For example, centralization can tell us whether a campaign's communications rely on a few pivotal actors or if its propagation structure is dispersed. Cohesion describes how tightly interconnected people are, whereas modularity measures the extent to which they cluster into groups that infrequently mix. Finally, heterophily represents how often actors with different characteristics interact. In the context of electoral processes, this could measure how often people from differing political backgrounds communicate, thus indicating the likelihood that an IO narrative is shared across political party lines. Figure 2 illustrates how these measures can aid our understanding of a given campaign.

FIGURE 2. THE WHITE HELMETS, A SYRIAN VOLUNTEER RESCUE SQUAD, WERE EVACUATED TO SEVERAL COUNTRIES IN LATE JULY. AS PART OF A BROAD SAMPLING OF TWITTER MESSAGES REGARDING THE SYRIAN CONFLICT, THE AUTHORS COLLECTED ACCOUNTS WARNING THE RECEIVING COUNTRIES OF THE HELMETS' SUPPOSED TERRORIST TIES. THIS IS A TWITTER-MENTION NETWORK FOR THE 'CANADA/WHITE HELMETS/TERRORISM' NARRATIVE IN EARLY AUGUST. THE NETWORK HAS LOW CENTRALIZATION (0.044 USING DEGREE) AND LOW COHESION (0.003 USING EDGE DENSITY), REFLECTING THE LACK OF A DOMINANT ACTOR OR FREQUENT CROSS-NETWORK COMMUNICATION. BECAUSE GROUPS ARE HIGHLY SEPARATE, IT IS HIGHLY MODULAR (0.781 USING UNDIRECTED LOUVAIN). ASSORTATIVITY (0.219) IS ALSO LOW, WHICH REFLECTS THAT PEOPLE TEND TO MENTION OTHERS WITH SIMILAR VIEWS, THOUGH THE VIEWS OF MANY MENTIONED ACCOUNTS ARE UNAVAILABLE.



To create these measures, one must decide on a modelling approach. Network constructions differ by media platform. Figure 3 defines network models based on two dimensions. First, is it possible to have a relationship that is not reciprocated (i.e. to favourite or follow)? Asymmetric relationships are better represented by directed networks (top row) whereas mutual relationships (i.e. to friend) map to undirected networks (bottom row). The second dimension reflects whether algorithms impact the information an actor sees. Unaffected communications are dictated by personal choice and/or timing. When algorithmic influence is present, actors will see different message orderings based on their individual parameterizations.

FIGURE 3. A TAXONOMY OF NETWORK MODELS FOR MEDIA PLATFORMS ALONG TWO DIMENSIONS: ASYMMETRIC VS. MUTUAL RELATIONSHIPS, AND UNMEDIATED VS. ALGORITHMICALLY-MEDIATED COMMUNICATIONS. AS OF 2018, CATEGORICAL EXAMPLES ARE: (A) TRADITIONAL MEDIA OUTLETS (TV, NEWSPAPERS) TO USERS, BLOG LINKAGES, EARLY INSTAGRAM AND TWITTER IMPLEMENTATIONS; (B) CURRENT INSTAGRAM AND TWITTER, YOUTUBE; (C) CHATROOM-LIKE PLATFORMS INCLUDING WHATSAPP, FACEBOOK MESSENGER, SNAPCHAT, AND DISCORD; (D) THE FACEBOOK TIMELINE.



These categories exclude forums like Reddit and 4chan due to the difficulty of distinguishing users with a relationship from users with similar preferences. When this distinction is unnecessary, the mutual relationships/unmediated model can be applied.

Beyond measuring a campaign's organization, we may wish to evaluate its ability to engage and convert users. The innovation-decision process from diffusion of innovation theory maps five stages from awareness to adoption that can frame engagement levels and measures [22]. Characterizing users by stage in a network diagram may help gauge an information operation's impact on a target audience.

'Silent' intermediate objectives intend to shape the network environment and may precede message delivery. For example, researchers studying Russian influence on the US's white supremacy movement found themselves targeted by bot attacks: previously dormant bots followed the researchers *en masse* before flooding their Twitter notifications with messages [23]. This mass-following would be reflected as sudden changes in network measures, including cohesion and centralization.

C. Attribution Strategies

Doing attribution well is at the core of virtually all forms of coercion and deterrence.
– Ben Buchanan and Thomas Rid

Attribution is not a common aim for the SNA community [24], but SNA research may be practically applied to address the challenge of online anonymity. For example, methods that emphasize finding consistent patterns and inferring relationships could disambiguate groups across campaigns. Matching accounts across networks can also identify additional data sources for attributional clues.

1) Affiliation Networks

Affiliation networks construct possible relationships between people based on shared event attendance, group membership, or other commonalities [25]. We can use these networks to infer coordination among actors in otherwise potentially unrelated events. For example, Campana reconstructed a human trafficking network's structure using co-event data drawn from court files [26]. By comparing perpetrators' roles with their network positions, the author derived evidence that the trafficking ring was driven by specialized and independent actors rather than a unified organization.

Technical artefacts, including code similarities, media, or metadata, can also define affiliation networks. Saxe and Sanders built a network between malware samples based on shared icons, and found a cluster of linked Trojans. Through additional analysis, they proved that the clustered samples originated from the same source [27]. In IO, shared forums, slogans, or information sources could similarly be employed.

2) Structural Equivalence

If two actors are structurally equivalent, this means that their relationships are identical [28, 29]. This 'structural redundancy' can provide valuable clues. For example, in September 2014, Twitter began aggressively suspending ISIS accounts. That same month, there was a sudden surge of new ISIS-supporting accounts [30]. Preventing banned users from creating new accounts is difficult, but looking for structural equivalence over topics can help identify these 'rebound accounts.'

Analyzing actors across campaigns could be aided by ongoing research into how to compare roles between networks. Jeffrey Johnson's ethnographic approach of operationalizing social roles through detailed case studies is inspired by anthropology, but may be applicable for those actively participating in covert networks like dark web forums [31]. Some computational approaches include block modelling [32] and regular equivalence [33]. The practical need to identify TTPs, track operational consistency, or profile actor types may incentivize extending this theoretical work.

3) Network Deanonymization

Network deanonymization attempts to reconstruct actor identities in a network by matching them to the population of another network containing additional information. Ji et al. survey deanonymization techniques in [34] and provide a table (Table III) with information on their scalability, practicality, and computational efficiency. Most approaches require or are made significantly more effective with the presence of ‘seeds’, or successfully matched actors. Another challenge is identifying how well the known network’s population matches that of the hidden network. [35] explores methods to determine which auxiliary networks are most promising using the nodes’ network properties. Despite favourable results, follow-on efforts to explore and define its feasibility are still lacking. Due to ethical and regulatory concerns surrounding privacy [10], it is advised to fully understand one’s rights and limitations before attempting this approach. Regardless, it may prove useful for determining whether a known group has instigated a particular campaign.

D. Influence Analysis

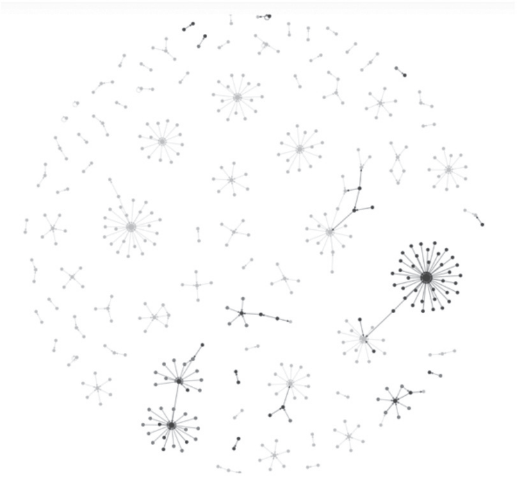
Power is unthinkable outside matrices of force relations; it emerges out of the very way in which figurations of relationships… are patterned and operate. – Mustafa Emirbayer

The European Commission Report states the need to ‘safeguard diversity and sustainability’ online [12]. The online environment is not a static system. Understanding how the rise and fall of influence is facilitated by social structures, dynamics, and platform design may guide the development of principles for future moderation efforts.

1) Identifying Key Actors

Many centrality measures exist for identifying important actors. Degree centrality helps identify particularly popular individuals. In an asymmetric network, the highest-degree actors are those most followed: examples include media outlets, influential bloggers, or maintainers of popular channels or podcasts. Figure 4 shows how one can track a message’s dominance by how frequently its proponents are quoted. Other roles within a network provide different types of influence. For example, one person may serve as a frequent mediator between two groups, such as a translator who interprets messages across linguistically bound communities. This may be captured using betweenness centrality, which measures how frequently an actor is present in communications across the network. Refer to [36] for further discussion of other centrality measures and their applicability and interpretability.

FIGURE 4. IN OUR ANALYSIS ON THE SYRIAN CONVERSATION, WE CREATED THIS NETWORK OF QUOTES AND RETWEETS FROM THE CONVERSATION ABOUT CANADA AND THE WHITE HELMETS. DARK NODES ARE USERS PUSHING THE WHITE HELMETS/TERRORIST NARRATIVE. THE MEDIUM NODES HAVE QUOTED THEM ON OTHER TOPICS, INDICATING THAT THEY WERE LIKELY EXPOSED TO THE NARRATIVE. ACTORS WITH THE HIGHEST DEGREE IN THIS NETWORK ARE THOSE WHO MOST SUCCESSFULLY HAD THEIR MESSAGE AMPLIFIED.



2) Creating Online Influence

Algorithms have an unseen effect on online communications. By altering communications between users, this mechanism changes what impact influencers can have on their connections. For example, YouTube's recommendation algorithm has been accused of promoting extreme content [37]. By recommending certain channels over others, this algorithm influences a user's choice of information sources. Algorithmic newsfeeds curate content based on a user's past preferences and actions [38, 39], thus shifting a user's likelihood of exposure to certain sources or posts. A content provider's ability to utilize these algorithms can determine their own influencing capabilities.

3) Influence Campaigns and Moderation

Online groups constantly seek to better promote their own personal or political beliefs, including state-based operations and extremist groups like ISIS. 4chan's famous trolling forum /pol/ attempted to influence the Google search algorithm to correlate racist terms with innocuous words [40]. Civilians have also used online platforms to increase their political influence, as seen in such high-profile cases as the Arab Spring, the 2017 Women's March, and the Gilets Jaunes.

Furthermore, maintaining the online influence space as a free and balanced marketplace of ideas is as much an economic challenge as it is a technical or political one. The monetization capability of influence has led to strategic product placement in influencers' posts, via allocated ad spaces, and even using false accounts [9] to promote word-of-mouth recommendations. Social interactions have financial value in the online world, and it is unclear to what extent this complexity has been considered in our current models of online communications.

Platforms and internet providers also have the ability to impact influencers' capabilities. Moderation efforts span from top-down driven administration, like Twitter's efforts to combat ISIS accounts [30], to Wikipedia's decentralized organization [41]. Censorship shocks on the Mandarin Wikipedia demonstrate the possibilities of Internet provider effects [42].

Some of the most recent developments in SNA are dedicated to better understanding these phenomena. Refer to [43] for a variety of techniques designed to tease out the source of a diffused message. Exponential random graph models (ERGMs) and stochastic actor-oriented models (SAOMs) are applied to test theories of how micro-behaviours lead to differences in network structure [44, 45, 46]. Relational event models (REMs) and Dynamic Network Actor Models (DyNAMs) consider the likelihood of an actor's actions based on their relationships and environmental factors [47, 48]. As the computational cost of dynamic modelling is reduced, ongoing work in this area holds promise for further illuminating the causes and influences of online dynamics.

E. Network Interventions

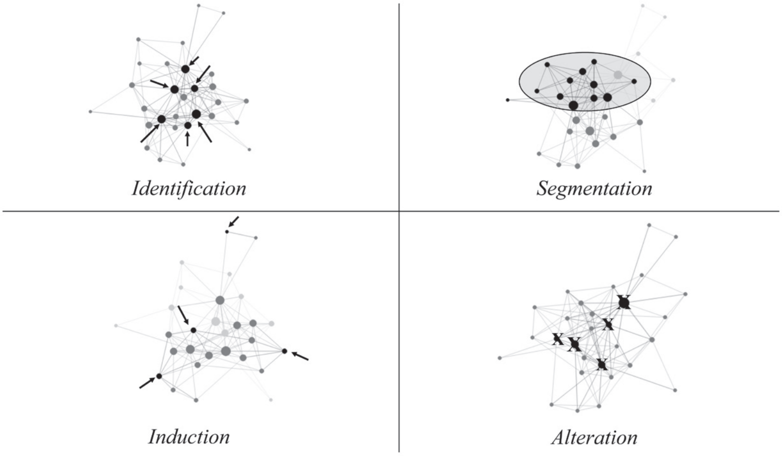
Example is not the main thing in influencing others. It is the only thing. - Albert Schweitzer

Network interventions use social influence forces to promote behaviour change [49, 22]. These interventions are based on the concept that exposure to a behaviour increases one's likelihood of adoption and that the influence of one's peers can be harnessed to spread desired behaviours. The authors have not identified conscious applications of network interventions to counter IO efforts; however, these interventions offer a comprehensive framework to classify suggested counter-IO tactics and inspire new approaches. Table 2 describes each intervention strategy and their unique capabilities, and Figure 5 demonstrates their potential for operationalization.

TABLE 2: NETWORK INTERVENTION STRATEGIES AND EXAMPLES

Strategy	Description	Example	Risks
Identification	Use network structure to identify actors to train to spread desired messaging or behaviour.	Opinion leaders on a social media site are identified and encouraged to spread counter-IO content.	Opinion leaders may not want to share desired content, or may share content that inadvertently increases belief in misinformation.
Segmentation	Simultaneously target actors that are in a well-connected group or in shared positions.	Clusters of highly connected accounts are located and targeted with counter-IO messaging.	Targeted accounts may view content as an attack on their community. Existing connections among members may reinforce current behaviour if members are not accepting of counter-IO information.
Induction	Promote communication across existing relationships in the network to disseminate desired messaging.	Civilians in affected area are encouraged to provide their accounts of on-ground activities.	Campaign hashtags or keywords may provide vehicles for continued misinformation spread. Civilian accounts may be framed to support an undesirable narrative.
Alteration	Change network structure to alter exposure and message spread.	Bridging actors on social media platforms are identified and trained in misinformation detection and handling to reduce the likelihood of their transmission of misinformation.	Actors may misunderstand intervention materials and increase the spread of misinformation through well-intended efforts to correct misinformation.

FIGURE 5. GIVEN THE CANADA/WHITE HELMETS/TERRORISM NARRATIVE, HOW COULD WE DESIGN AN INTERVENTION? IDENTIFICATION COULD TARGET HIGH-DEGREE NODES, WHILE INDUCTION WOULD BE MORE RANDOMLY DISPERSED. SEGMENTATION WOULD LOOK FOR CLUSTERS, AND ALTERATION COULD ADDRESS NODES THAT CONNECT DIVERSE POPULATIONS USING BETWEENNESS CENTRALITY.



Identification techniques engage actors in key positions in a network for training or messaging, with the expectation that their actions will impact the overall network. For example, rumour blocking simulations model the spread of misinformation and credible information simultaneously. Some researchers use identification techniques within these models to identify an optimized subset of users to spread credible information more effectively [50].

A tightly-knit group with few external relationships and frequent sharing of homogenous content can become an echo chamber. Segmentation methods can intervene with an echo chamber as a collective set so all actors receive content simultaneously.

Induction techniques reframe a narrative by actively encouraging people to communicate with one another. An example of this approach would be a word-of-mouth campaign that asks civilians to share their views on a topic with photos or other user-created media. Sharing user experiences from those close to an on-the-ground situation may aid in combating false information pertaining to that situation.

Finally, alteration methods modify network structure by adding or deleting links and/or nodes. Note that removing malicious bots or accounts from online platforms does not necessarily eliminate them: bot masters may make new accounts that are harder to detect or migrate to other platforms. However, not all forms of node removal require explicit removal. Analogous to how vaccinations prevent disease transmission, we can focus on techniques that reduce accounts' transmission of misinformation [49]. Training and messaging actors who play central roles in spreading information may effectively reduce an IO's diffusion through a network. Finally, link-based alteration strategies include encouraging people to connect or disconnect from particular accounts. A recent report suggested that actors that connect more with people that have differing opinions may reduce their belief in misinformation [51].

Node addition may be an overlooked tactic for network alteration. Self-identified bots could serve as assistive devices to provide just-in-time content to counter or distract from disinformation. For example, a monitoring account could analyse tweets and reply with an automated analysis of potentially coercive or emotionally evocative content, though the risk of false positives should be considered. Simulations have been conducted to inform optimal monitor placement within a network for misinformation detection for early containment [52].

3. CONCLUSION AND FUTURE DIRECTIONS

Through SNA, we gain a theoretical lens and applicable methodologies for examining and countering IO. Some of SNA's capabilities, like centrality measures and clustering, have been frequently applied to the online environment while others remain underutilized. Here we seek to broaden the audience's perspective of ongoing research in the field.

We note that social network analysis is not a cure-all for addressing IO. Because it is message-agnostic, theories related to the shaping and framing of a narrative are absent from this work. Furthermore, no tool replaces the need for collaboration among stakeholders.

Regardless, SNA has strong potential when combined with other technical and political techniques. As demonstrated above, SNA-combined approaches lead to more effective ways to identify information campaigns and extremist organizations than machine learning alone. Network-based measures and attributional information can help guide the decision-making process regarding whether to address potential campaigns. Finally, network intervention techniques provide potential strategies for implementing campaign countermeasures. We encourage policymakers and researchers alike to consider how SNA methodologies can further the development of countermeasures against online IO.

Acknowledgments
We thank the anonymous reviewers, Brad Ward, Dr David Silberberg, Dr Anthony Johnson, John Renda, Kristine Henry, and Daniel Kapellmann for their guidance on this work.

BIBLIOGRAPHY

[1] US Joint Staff, "Joint Publication 3-13 Information Operations," Government Printing Office, Washington DC, 2014.
[2] D. Lindley, "Untapped power? The status of UN information operations," *International Peacekeeping*, vol. 11, no. 4, pp. 608-624, 2004.
[3] K. Avruch, J. L. Narel and P. C. Siegel, Information Campaigns for Peace Operations, Washington DC: Office of the Assistant Secretary of Defense, Washington DC Command and Control Research Program (CCRP), 2000.
[4] Canadian Forces, "Canadian Forces Information Operations," Canadian Forces, Ottawa, 1998.
[5] R. Vandomme, "From Intelligence to Influence: The Role of Information Operations," Canadian Forces College, Toronto, 2010.
[6] NATO, "Allied Joint Doctrine for Information Operations," NATO, Talinn, Estonia, 2009.
[7] D. Denning, *Information Warfare and Security*, Reading, MA: Addison Wesley, 1999.
[8] S. Shane and M. Mazzetti, "The Plot to Subvert an Election," *The New York Times*, 20 9 2018.

[9] M. Kovic, A. Rauchfleish, M. Sele and C. Caspar, "Digital astroturfing in politics: Definition, typology, and countermeasures," *Studies in Communication Sciences*, vol. 18, no. 1, pp. 69-85, 2018.

[10] J. Soetbeer, "European Data Protection Regulation – Information Sheet," 1 3 2016. [Online]. Available: https://www.privacy-europe.com/blog/european-data-protection-regulation-information-sheet/. [Accessed 9 12 2018].

[11] U. Brandes, G. Robins, A. McCranie and S. Wasserman, "What is network science?" *Network Science*, vol. 1, no. 1, pp. 1-15, 2013.

[12] European Commission High Level Group, "A multi-dimensional approach to disinformation - Report of the independent High Level Group on fake news and online disinformation," Publications Office of the European Union, Belgium, 2018.

[13] M. C. Benigni, K. Joseph and K. M. Carley, "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter," *PloS one*, vol. 12, no. 12, p. e0181405, 2017.

[14] O. Varol, E. Ferrara, F. Menczer and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ Data Science*, vol. 6, no. 1, p. 13, 2017.

[15] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini and F. Menczer, "Detecting and tracking political abuse in social media," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona, 2011.

[16] J. M. Berger, "Tailored online interventions: The Islamic State's recruitment strategy," *CTC Sentinel*, vol. 8, no. 10, pp. 19-23, 2015.

[17] Q. Cao, X. Yang, J. Yu and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Scottsdale, AZ, 2014.

[18] I. McCulloh and K. M. Carley, "Detecting Change in Longitudinal Social Networks," *Journal of Social Structure*, vol. 12, no. 3, pp. 1-37, 2011.

[19] I. McCulloh, M. Webb and K. M. Carley, "Social Network Monitoring of Al-Qaeda," *Network Science*, vol. 1, no. 11, pp. 25-30, 2007.

[20] O. Varol, E. Ferrara, C. A. Davis, F. Menczer and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," *arXiv*, vol. 1703, no. 03107, 2017.

[21] S. F. Everton, *Disrupting dark networks* Vol. 34, New York: Cambridge University Press, 2012.

[22] E. Rogers, *Diffusion of Innovations*, 5th edition, New York: Free Press, 2003.

[23] J. Cox, "The Day an Army of Bots Turned on Bot Researchers," *The Daily Beast*, pp. https://www.thedailybeast.com/the-day-an-army-of-bots-turned-on-bot-researchers?ref=scroll, 29 8 2017.

[24] N. Hummon and K. M. Carley, "Social networks as normal science," *Social Networks*, vol. 15, no. 1, pp. 71-106, 1993.

[25] S. Borgatti, M. Everett and J. Johnson, "Analyzing Two-Mode Data," in *Analyzing Social Networks*, Los Angeles, CA, SAGE, 2013, pp. 267-286.

[26] P. Campana, "The Structure of Human Trafficking: Lifting the Bonnet on a Nigerian Transnational Network," *The British Journal of Criminology*, vol. 56, no. 1, pp. 68-86, 2016.

[27] J. Saxe and H. Sanders, "Identifying Attack Campaigns with Malware Analysis," in *Malware Data Science: Attack Detection and Attribution*, San Francisco, No Starch Press, Inc., 2018, pp. 54-58.

[28] H. C. White and F. Lorrain, "Structural equivalence of individuals in social networks," *The Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49-80, 1971.

[29] R. Burt, "Social contagion and innovation: Cohesion versus structural equivalence," *American Journal of Sociology*, vol. 92, no. 6, pp. 1287-1335, 1987.

[30] J. M. Berger and J. Morgan, "The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter," The Brookings Project on U.S. Relations with the Islamic World, Washington D.C., 2015.

[31] J. C. Johnson, C. Avenarius and J. Weatherford, "The Active Participant-Observer: Applying Social Role Analysis to Participant Observation," *Field Methods*, vol. 18, no. 2, pp. 111-134, 2006.

[32] H. C. White, S. A. Boorman and R. L. Breiger, "Social Structure from Multiple Networks," *American Journal of Sociology*, vol. 81, no. 4, pp. 730-780, 1976.

[33] D. R. White and K. P. Reitz, "Graph and semigroup homomorphisms on networks of relations," *Social Networks*, vol. 5, no. 2, pp. 193-234, 1983.

[34] S. Ji, P. Mittal and R. Beyah, "Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1305-1326, 2017.

[35] P. Govindan, S. Soundarajan and T. Eliassi-Rad, "Finding the most appropriate auxiliary data for social graph deanonymization," in *1st KDD Workshop on Data Ethics*, New York, New York, 2014.

[36] S. P. Borgatti, M. G. Everett and J. C. Johnson, “Centrality,” in *Analyzing Social Networks*, Los Angeles, SAGE Publishing, 2013, pp. 189-208.

[37] Z. Tufekci, “YouTube, the Great Radicalizer,” *The New York Times*, p. SR6, 10 3 2018.

[38] M. A. DeVito, “From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed,” *Digital Journalism*, vol. 5, no. 6, pp. 753-773, 2017.

[39] N. Koumchatzky and A. Andryeyev, “Using Deep Learning at Scale in Twitter’s Timelines,” 9 5 2017. [Online]. Available: https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html. [Accessed 9 12 2018].

[40] G. E. Hine, J. Onaolapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini and J. Blackburn, “Kek, cucks, and god emperor Trump: A measurement study of 4chan’s politically incorrect forum and its effects on the web,” in *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM*, Montreal, 2016.

[41] A. Forte, V. Larco and A. Bruckman, “Decentralization in Wikipedia Governance,” *Journal of Management Information Systems*, vol. 26, no. 1, pp. 49-72, 2009.

[42] A. F. Zhang, D. Livneh, C. Budak, L. Robert and D. Romero, “Shocking the Crowd: The Effect of Censorship Shocks on Chinese Wikipedia,” in *The 11th International Conference on Web and Social Media*, Montreal, Canada, 2017.

[43] Jiaojiao, S. Wen, S. Yu, Y. Xiang and W. Zhou, “Identiying Propagation Sources in Networks: State-of-the-Art and Comparative Studies,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 465-481, 2017.

[44] G. Robins, T. A. B. Snijders, P. Wang and M. Handcock, “Recent developments in exponential random graph (p*) models for social networks,” *Social Networks*, vol. 29, pp. 192-215, 2007.

[45] D. A. McFarland, J. Moody, D. Diehl, J. A. Smith and R. J. Thomas, “Network Ecology and Adolescent Social Structure,” *American Sociological Review*, vol. 79, no. 6, pp. 1088-1121, 2014.

[46] T. A. B. Snijders, “Stochastic actor-oriented models for network change,” *Journal of Mathematical Sociology*, vol. 21, no. 1-2, pp. 149-172, 1996.

[47] C. T. Butts, “A Relational Event Framework for Social Action,” *Sociological Methodology*, vol. 38, no. 1, pp. 155-200, 2008.

[48] C. Stadtfeld, J. Hollway and P. Block, “Dynamic Network Actor Models: Investigating Coordination Ties through Time,” *Sociological Methodology*, vol. 47, no. 1, pp. 1-40, 2017.

[49] T. W. Valente, “Network Interventions,” *Science*, vol. 337, no. 6090, pp. 49-53, 2012.

[50] I. Litou, V. Kalogeraki, I. Katakis and D. Gunopulos, “Real-Time and Cost-Effective Limitation of Misinformation Propagation,” in *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, Porto, 2016.

[51] D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, W. Hobbs and C. Mattison, “Combating fake news: An agenda for research and action,” in *Combating fake news: An agenda for research and action,* Cambridge, MA, 2017.

[52] H. Zhang, M. A. Alim, M. T. Thai and H. T. Nguyen, “Monitor placement to timely detect misinformation in Online Social Networks,” in *2015 IEEE International Conference on Communications (ICC)*, London, 2015.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Understanding the Strategic Implications of the Weaponization of Artificial Intelligence

Dr Joe Burton[1]
New Zealand Institute for
Security and Crime Science
University of Waikato
New Zealand

Dr Simona R. Soare
Institut d'Etudes Européennes,
Université Saint Louis - Bruxelles
Belgium

Abstract: Artificial Intelligence (AI) is expected to have a revolutionary impact across societies and to create economic displacement and disruption in security and defense. Yet the impact of AI on national security and military affairs has received relatively scant attention. The existing policy-focused literature has concentrated mainly on the technological, ethical or legal limitations of deploying AI and on the risks associated with it. This paper seeks to contribute to the debate by outlining the strategic implications of the weaponization of AI for international security. It explores how and in what ways AI is currently being utilized in the defense sector to enhance offensive and defensive military technologies and operations and assesses the ways in which the incorporation of AI into military platforms will affect war fighting and strategic decision-making. The paper is in four sections. Section one develops a typology of military AI that forms a foundation for the rest of the paper. The second section examines the uses of AI in cyberspace and the relationships between 'cyber weapons' and AI capabilities. The third section examines how the embeddedness of AI-based capabilities across the land, air, naval and space domains may affect combined arms operations. The final section distills the main strategic implications of weaponized AI, which include the speed of decision-making and action as well as enhanced domain situational awareness.

Keywords: *artificial intelligence, weaponization, cyber defense, strategy*

[1] This paper was drafted with equal contributions from both authors. The authors would like to thank the CyCon program committee and the two anonymous reviewers for their invaluable feedback on this paper.

1. INTRODUCTION

James Cameron's cult film *The Terminator* depicted a dystopian future in which Skynet, a malevolent Artificial Intelligence (AI), initiates a nuclear war against humans to ensure its own survival. The film was released in 1984, well before the advent of modern forms of AI, but was prescient in foreshadowing some of the concerns that have come to dominate debates about intelligent computer systems. The late renowned scientist Stephen Hawking described AI as the single greatest threat to human civilization,[2] which is not the first time scientific and technological innovation has been perceived as an existential threat,[3] and Henry Kissinger has warned that AI will change human thought and human values.[4] In recent years activists, scientists and governments[5] have sought to place UN-level bans on 'killer robots', including Lethal Autonomous Weapons Systems.[6] The technology that *The Terminator* films depicted is not yet with us, and a form of self-aware artificial intelligence described as 'general AI' is, according to most analysts, some decades away, yet the impact of AI in international security is beginning to receive sustained attention.

By some accounts, an AI arms race is emerging between the great powers, and the US, China and Russia in particular.[7] AI systems are already being incorporated into weapons platforms and military technologies, including missile defense systems, Unmanned Aerial Vehicles (UAVs), Unmanned Underwater Vehicles (UUVs), fighter aircraft and naval platforms.[8] In the realm of cyber security, AI could revolutionize how we protect computer systems from nefarious actors, but could also be used to develop much more sophisticated attack vectors, methods and technologies. The proliferation of AI to non-state actors, the rapid pace of technological change and the growing sophistication of the new technologies are also causing concerns, and there is a risk that policymakers are unprepared for sudden shifts in how AI technologies are used. This phenomenon is not new. Legislation gaps often occur with societal transitions to new technologies. It is, however, compounded by the fact that much of the technology is being developed by the private sector, including companies like

[2] Kharpal, A. (2017, November 06). Stephen Hawking says A.I. could be 'worst event in the history of our civilization'. Retrieved from https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html.

[3] AI is but one of a long list of threats to human civilization, including nuclear weapons, biological and radiological weapons, severe cataclysms and genetic experimentation.

[4] Kissinger, H. A. (2018, May 16). How the Enlightenment Ends. Retrieved from https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

[5] See for example: Open Letter on Autonomous Weapons. (n.d.). Retrieved from https://futureoflife.org/open-letter-autonomous-weapons.

[6] Busby, M. (2018, April 09). Killer robots: Pressure builds for ban as governments meet. Retrieved from https://www.theguardian.com/technology/2018/apr/09/killer-robots-pressure-builds-for-ban-as-governments-meet.

[7] Auslin, M. (2018, October 23). Can the Pentagon Win the AI Arms Race? Retrieved from https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race.

[8] Stewart, P. (2018, June 05). Deep in the Pentagon, a secret AI program to find hidden nuclear... Retrieved from https://www.reuters.com/article/us-usa-pentagon-missiles-ai-insight/deep-in-the-pentagon-a-secret-ai-program-to-find-hidden-nuclear-missiles-idUSKCN1J114J.

IBM, Google and Apple in the US, and Baidu, Alibaba and Tencent in China, leaving legislators struggling to regulate, control and mitigate some of AI's associated risks and to explore inherent opportunities. International organizations are beginning to respond to these challenges and governments are starting to develop their own national AI strategies and investment plans. In 2018, the EU, for example, released a civilian and economy-focused AI strategy,[9] and in the last several years a host of countries, including Canada, China, Denmark, Finland, France, India, Italy, Japan, Mexico, Singapore, South Korea, Sweden, Taiwan, the UAE, and the UK have released strategies to promote the use and development of AI.[10] In 2019, the US published its Department of Defense AI Strategy, which aims to accelerate the integration of AI across the US armed forces.[11]

Despite this growing attention, there are many areas of AI research in both the technical and political realms that are underdeveloped and have received surprisingly scant attention. This is especially true in the security and strategic studies disciplines in which the technical and practical aspects of AI development meet the political and doctrinal ones. How AI will affect military operations and how it can be harnessed to increase and enhance international security are questions that are only beginning to be addressed by security scholars.[12] Two schools of thought appear to be emerging in this nascent literature. The first argues that AI deployment in security and defense will have a revolutionary effect on operations (e.g. human-machine teaming), capabilities (e.g. swarms) and military structures (e.g. human-machine interfaces), and on how militaries interact with the civilian and political realms. Much of the literature in this school draws on the technical specifications of AI applications in the military field to derive conclusions about its likely revolutionary impact (which is arguable and speculative at this point in time). The second school of thought argues that AI will have a more evolutionary impact on international security, that its focus will be on increasing the efficiency of 'dull-dirty-and-dangerous' military tasks and on the speed of decision-making (through accurate situational awareness and actionable intelligence), and that it will not fundamentally change the nature of warfare.

[9] Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines. (n.d.). Retrieved from http://europa.eu/rapid/press-release_IP-18-3362_en.htm.

[10] Dutton, T. (2018, June 28). An Overview of National AI Strategies. Retrieved from https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

[11] US Department of Defense (2019: February 28). Summary of the Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity. Retrieved from https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

[12] Payne, K. (2018). Artificial Intelligence: A Revolution in Strategic Affairs? *Survival*, 60(5), 7-32. doi:10.1080/00396338.2018.1518374; Cummings, M. L., Roff, H. M., Cukier, K., & Parakilas, J. (2018, June 14). Artificial Intelligence and International Affairs: Disruption Anticipated. Retrieved from https://www.chathamhouse.org/publication/artificial-intelligence-and-international-affairs; Hoadley, D. S. and Lucas, N. J. (2019, January 30). Artificial Intelligence and National Security. *Congressional Research Service*. Retrieved from https://fas.org/sgp/crs/natsec/R45178.pdf; Sheppard, L. R. (2018, November 5). Artificial Intelligence and National Security: The Importance of the AI Ecosystem. Retrieved from https://www.csis.org/analysis/artificial-intelligence-and-national-security-importance-ai-ecosystem; Scharre, P., & Horowitz, M. C. (2018, June 22). Artificial Intelligence: What Every Policymaker Needs to Know. Retrieved from https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know.

In this context, we argue that empirical evidence and existing governmental AI strategies seem to suggest a middle path: that the role of AI will differ across military tasks. While AI may revolutionize tasks such as logistics and maintenance, it will be evolutionary for others, including decision-making (i.e. humans will continue to make political and military life-and-death decisions). In building this argument the aim of the paper is to shed further light on some of the crucial dynamics that will affect how AI is integrated into strategic planning and affect decision-making in relation to modern war and conflict. In particular, we focus on the process and implications of the weaponization of AI – meaning (a) how AI is and might be incorporated into weapons systems and platforms, and (b) how AI technologies themselves may be used with ill-intent to cause harm in the international arena. The paper seeks to understand the strategic implications of the process of weaponization and the results of that process, and in doing so to raise awareness and help contribute to emerging debates in the military and strategic studies communities about how AI affects military strategy.

The paper proceeds in four main sections. In the following section we outline the types of AI that are being developed that have usages in the military sector. This section works towards a typology of military AI that forms a foundation for the rest of the paper. The next section examines the uses of AI in cyberspace and the relationships between "cyber weapons" and weapons systems that are based on AI tools and capabilities. The following section examines how the embeddedness of AI across the land, air, naval and space domains may affect combined arms operations. The final section distils the main strategic implications of weaponized AI, which include changes in the speed of decision-making and action as well as implications for cross domain situational awareness.

2. TOWARDS A TYPOLOGY OF MILITARY ARTIFICIAL INTELLIGENCE

Much of the debate around the emergence of AI as a factor in military planning has suffered from a confusion about what exactly AI is and its various forms and utilities. This lack of clarity is not surprising given the complexity of the technology and the challenge of advancing scientific understanding in non-scientific communities. Across the international security and strategic studies disciplines, scholars are grappling with the implications of technologies that are opaque, highly technical, and developed by scientific disciplines with which they have had little interaction. The profusion of various forms of AI and their already widespread usage in the commercial sector has also complicated efforts to categorize and define the emerging AI marketplace. Voice recognition and commands are now built into everyday objects and platforms, and algorithms that predict and analyze information in real time are used extensively across a range of societal activity, including in the financial sector, market decision-making, and in software and computer hardware development. Yet often, the blanket term "AI" is used to describe a range of technologies, methods and processes which are different and distinguishable from one another.

At the most basic level, AI is a form of technology that exhibits human characteristics – most notably that of intelligence. Intelligence is the ability to reason and perform complex tasks, to understand and adapt to one's environment, and to learn from previous interactions and situations.[13] Intelligent machines will be able to perform complex tasks, be able to learn and improve operationally over time, and do so without human input. Moving beyond this basic definition, the first type of AI classification is a disciplinary one: *practical AI* refers to technological development and computing requirements associated with technical progress; and *fundamental AI* refers to the social, economic, psychological, philosophical and political implications of AI use.[14] Practical AI has seen its ups and downs since the 1950s. In the last decade there has been an exit from the "AI winter" of the previous several decades, a period where technological advancement stagnated, and there have been some rapid technological advancements. Fundamental AI, however, has struggled to keep up with the technological progress in practical AI. The growing gap between the two was well framed by Henry Kissinger, who has said we are in the presence of "a potentially dominating technology in search of a guiding philosophy".[15]

A further distinction in the contemporary literature on AI technology relates to the number of tasks it can perform at a time. The first category is *narrow AI*, which is the most common type of AI already in civilian and military use: this refers to technology that can perform a single task at a time – the task it has been specifically built to perform. It does not have the ability to migrate the knowledge or behaviors it is taught or has learned in one context to other situations. Scholars refer to this limitation as "catastrophic forgetting", meaning *narrow* AI cannot be repurposed for other tasks.[16] The systems involved are either *reactive*, in that they are not capable of forming memories or using past interaction to shape decisions, or have *limited memories*, in that they might process simple pieces of past information but are not capable of using that information systematically to influence or make decisions.

The second category is *general AI*, which is not yet deployed either in the civilian or the military realms. Through analogy with human intelligence, general AI is supposed to be able to perform several tasks at a time. It has the ability to understand context, to successfully apply information and behaviors learnt in one context to other situations

[13] Intelligence. (n.d.). Retrieved from https://www.merriam-webster.com/dictionary/intelligence.

[14] The authors would like to thank the anonymous reviewer for raising this point of difference.

[15] Kissinger, H. A. (2018, May 16). How the Enlightenment Ends. Retrieved from https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

[16] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521-3526. doi:10.1073/pnas.1611835114.

it encounters, and in circumstances other than the task it was designed to perform. In this category, intelligent machines will be able to adjust behavior depending on interaction with people and other technologies and understand the context, motivations and complex intentions of these actors. This type of AI has been referred to as "theory of mind" AI.[17] *General* AI in its most sophisticated form may become self-aware – this is a field of AI often referred to as artificial consciousness, machine consciousness, synthetic consciousness or singularity. The debates around the plausibility of the emergence of self-aware forms of AI are ongoing. In "Artificial Consciousness: Utopia or Real Possibility", Giorgio Buttazzo refutes the possibility that machines can exhibit consciousness,[18] but some scholars argue that AI may develop a level of sophistication *commensurate* with the human mind.

Thirdly, AI can be classified both as *software* and as *hardware*. Technically speaking, AI is an individual algorithm or system of algorithms (i.e. software). However, AI software is most generally deployed *together with and/or integrated on* physical platforms, be it robots, drones or systems of sensors. AI, either software or hardware, is dependent on being developed and deployed in a data ecosystem that it can monitor, exploit or adapt to achieve its tasks. In this sense, AI is fundamentally creating new capabilities and capacities for military institutions across the world, much like 'systems of systems' did in the late 1980s and early 1990s.

Another means of classification for AI refers to the *types of tasks or roles* it can perform. In the field of international relations and security, AI roles are generally considered to be analytical, predictive or operational.[19] Depending on the category, some roles are more important and likely to be more transformative than others: analytical roles provide decision-makers with actionable intelligence and improve situational awareness; predictive roles may have a significant transformative role at the tactical, operational and strategic level of military operations; whereas at the operational end of the spectrum, AI, robotics and automation are expected to take over a number of dull, dirty and dangerous tasks. Depending on the roles it is deployed to perform, AI software procurement is unlikely to result in easily quantifiable capabilities; AI in the form of lethal autonomous weapon systems, however, such as swarms, autonomous drones or autonomous underwater vehicles, will lead to the development of countable military capabilities. Swarm strategy and the intelligent collective behavior of these swarms is surely one of the most promising fields of AI R&D. Moreover, human-machine interaction, collectively and individually, and its

[17] Minsky, M. L. (2007). *The society of mind*. New York: Simon & Schuster Paperbacks; see also Azarian, B. (2018, November 8). Intelligent Social Robots Must Have a "Theory of Mind". Retrieved from https://www.psychologytoday.com/us/blog/mind-in-the-machine/201811/intelligent-social-robots-must-have-theory-mind.

[18] Buttazzo, G. (2001). Artificial consciousness: Utopia or real possibility? *Computer*, 34(7), 24-30. doi:10.1109/2.933500.

[19] Cummings, M. L., Roff, H. M., Cukier, K., & Parakilas, J. (2018, June 14). Artificial Intelligence and International Affairs: Disruption Anticipated. Retrieved from https://www.chathamhouse.org/publication/artificial-intelligence-and-international-affairs.

technical and legal interfaces, will also create new capabilities. Therefore, AI is likely to significantly impact the qualitative and quantitative international balance of power.

AI is a dual-use technology, and as with all dual-use technology its specifications determine the degree to which it is likely to spread in the military or civilian realms. At the present time, the forms of AI in usage in the military sector are predominantly narrow AI, including reactive and limited memory AI. These forms of technology have been incorporated into a wide range of military platforms, systems and processes. At the softer end of the security spectrum, AI is in use in logistics and training; augmented reality systems, for example, are already in use in the Royal New Zealand Navy for training engineers to work on naval platforms.[20] In its perhaps most widespread and currently consequential role, AI is being used for Intelligence, Surveillance and Reconnaissance (ISR). One controversial example is the National Security Agency's (NSA) 'Prism' program, which applied AI systems to big data for counter-terrorism purposes.[21] At the harder end of the military spectrum, AI is being incorporated into missile defense systems, drones and other unmanned vehicles capable of deploying military force, and in targeting for weapons systems. The Israeli Harpy drone – a loitering munition also known as a 'fire and forget' system – is, judging by its technical specifications alone, a fully-autonomous weapon system.[22] The Japanese military is also considering acquiring ballistic missile defense drones that are capable of autonomously tracking incoming missiles.[23]

Conceptualizing AI weaponization

While there is a wide range of usages of AI in the military sector, the more consequential series of concerns exist at the harder end of the security spectrum. Significant concerns have arisen over the weaponization of AI. In this article we use this term to refer to two connected processes. The first is the use and integration of AI technology in weapons systems and platforms across the four domains of warfare (land, air, sea, space) for strategic advantage. In this first category, AI is used to enhance and multiply the effects of military operations, to enable rapid dispersion and concentration of force, to increase the lethality, precision and destructiveness of the application of military power, to give offensive operations an advantage and to erode an adversary's ability to defend itself. The second way we conceive of weaponization is through the use of AI as a stand-alone capability to undermine, disrupt and destroy enemy systems through computer network-enabled operations. Weaponization thus refers to both its use to enhance the power of conventional military assets, and the weaponization of the software and data through and within cyberspace (the 5th domain). The latter is dealt with in a following section.

[20] Author visit to Devonport Naval Base, Auckland, NZ.

[21] Kalakota, R. (2013, June 17). NSA PRISM – The Mother of all Big Data Projects - DZone Big Data. Retrieved from https://dzone.com/articles/nsa-prism-—-mother-all-big.

[22] Harpy NG. (n.d.). Retrieved from http://www.iai.co.il/2013/36694-16153-en/Business_Areas_Land.aspx.

[23] Sakhuja, V. (2018, June 27). Asian Militaries and Artificial Intelligence. Retrieved from http://www.indiandefencereview.com/asian-militaries-and-artificial-intelligence/.

The process of weaponization has been studied in various security-related fields, the most prominent being the weaponization of nuclear materials and programs.[24] Similar concerns have been documented concerning the weaponization of toxins and biological and chemical agents, and the manipulation of weather and climate has even been examined in the concept of weaponization.[25] There is also a substantial literature on weaponizing outer space, most often referring to placing military assets and capabilities in earth's orbit. More recently, the notion that information is being weaponized has received significant attention, especially in the context of Russian information operations, active measures and the use of cognitive behavioral algorithms to achieve 'mass manipulation' effects.[26] Common to existing analyses of weaponization processes is the use of civilian or dual-use technologies for military purposes. This basic dynamic applies to nuclear, outer space, biological agents and much of the other weaponization literature. AI has widespread uses across societal functions and, unlike the internet, which was originally a military network, has not been developed with military purposes at the forefront of planning and funding. However, the military has clearly been interested in the functionality of AI technologies for some time, including for the purposes of achieving strategic surprise, achieving a military advantage over one's opponent or otherwise creating politically-driven military effects.

The process of weaponization – be it in the nuclear or information area – entails considerable risks. These are associated with the instability that the proliferation of technologies within the international arena creates, the prospect of arms races and security dilemmas, the risk that non-state actors will acquire weaponized agents, the risk that states will not be able to effectively control the weaponized technology, and that AI technologies will be uncontainable and result in unintended consequences when used. The risks associated with the weaponization of AI have not been outlined systematically[27] but include the development of bias within AI systems. This dynamic was demonstrated recently when a Microsoft chatbot called 'Tay' was given its own Twitter account and allowed to interact with the public and, as a result of being fed malicious data, began to exhibit racism, sexism, and extremist political viewpoints. If bias develops within AI that is integrated into military systems, either as a result of manipulation or by the nature of the algorithm or data it processes, it will not serve to enhance military effectiveness. Another significant risk with AI systems is that they can be manipulated, and their integrity altered by malicious actors and even

[24] Thakur, R. (2014). The inconsequential gains and lasting insecurities of India's nuclear weaponization. *International Affairs*, 90(5), 1101-1124. doi:10.4324/9781315749488-8.

[25] Pincus, R. (2017). 'To Prostitute the Elements': Weather Control and Weaponisation by US Department of Defense. *War & Society*, 36(1), 64-80. doi:10.1080/07292473.2017.1295539.

[26] Waltzman, R. (2017, April 27). The Weaponization of Information: The Need for Cognitive Security. Retrieved from https://www.rand.org/pubs/testimonies/CT473.html.

[27] Cummings, M. L., Roff, H. M., Cukier, K., & Parakilas, J. (2018, June 14). Artificial Intelligence and International Affairs: Disruption Anticipated. Retrieved from https://www.chathamhouse.org/publication/artificial-intelligence-and-international-affairs.

programmed to perform unintended functions.[28] AI has also created concerns over social manipulation. Sophisticated data algorithms were used to affect social media in the run-up to the 2016 US general election and to exacerbate societal tensions, thus exhibiting the utility of weaponization of information by authoritarian states to undermine democratic ones. There have also been several concerns highlighting the misalignment of goals between humans and machines, where an AI is programmed and intended to accomplish a specific task but may not proceed according to the expectations of the programmer.[29] The lack of transparency of most AI algorithms in performing designated tasks is a significant problem and creates obstacles to their deployment in active security and defence roles.

3. THE WEAPONIZATION OF AI IN CYBERSPACE

Enhancing cyber security is becoming increasingly challenging due to the growing number of internet-connected devices and the exponentially increasing volume of data produced that needs securing. These basic dynamics affect the deployment of AI in cyberspace directly. The volume of data produced is such that humans will never be able to monitor data networks without assistance from machines. Cyber networks are vast and carry vast amounts of data. Monitoring the security of these networks is an exponentially increasing challenge in the 21st century. The potential for AI to have a positive impact in this area is obvious, particularly in enhancing the ability of human operators to monitor and respond to adversarial and abnormal events. As Vinod Vasudevan argues:

> Today's systems generate so much security data that human experts are rapidly surpassed. People cannot find the attack elements fast enough or reliably enough. By comparison, computers excel at these operations. AI then helps them to make sense of what they find. It can even help by offering suggestions to security teams of processes to handle them.[30]

AI may thus help mitigate offensive actions. It may also help to more effectively attribute cyber-attacks to specific actors by enhancing information and digital evidence collection and by providing probabilistic models to assess contradictory and uncertain data.[31]

[28] Hoadley, D. S. and Lucas, N. J. (2019, January 30). Artificial Intelligence and National Security. Congressional Research Service. Retrieved from https://fas.org/sgp/crs/natsec/R45178.pdf.

[29] Worley, G. G., III. (2018, February 19). Formally Stating the AI Alignment Problem. Retrieved from https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991.

[30] Vasudevan, V. (2018, July 24). How AI Is Transforming Cyber Defense. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2018/07/24/how-ai-is-transforming-cyber-defense/#8b13293bb20a.

[31] Nunes, E., Shakarian, P., Simari, G. I., & Ruef, A. (2018). *Artificial intelligence tools for cyber attribution*. Cham, Switzerland: Springer.

But how does the deployment of AI in cyberspace relate to the weaponization debates introduced in this article? First, there has been increasing concern in scholarly and policy circles about the vulnerability of AI to malicious interference affecting the integrity and operability of those systems. As we have stated, AI is software that exists on hardware. It is present on computers and computer networks that are just as vulnerable to intrusion and exploitation as any other computer network. AI is also based on sophisticated algorithms which can be manipulated or corrupted in the same way that other data can. Hackers are already developing tools to manipulate AI and turn it against the controller/user. This is beginning to be interpreted as an emerging security crisis.[32] There are several crucial concerns here. The first is that AI may be fooled into seeing things that are not there, misclassifying objects and processes, and/ or failing to identify patterns or processes within data that has become corrupted or corruptible.[33] Researchers at University of California, Berkeley, for example, recently invented a stop sign that could fool driverless cars. The implications of this in the military realm are significant. If military vehicles are manipulated into taking or not taking actions that are based on adversarial mal-intent, then serious consequences could ensue. Military satellites could be fooled into misclassifying military assets, which could have negative implications for situational awareness and decision-making. Manipulation of AI-based image identifiers could also be used to deliberately misidentify terrorist suspects, for example.

Advances in AI may also make malware itself more damaging, more sophisticated and better able to precision-target its intended recipient. One recent example is the Deeplocker malware, developed by IBM Research, which is highly evasive and able to conceal its malicious intent before it reaches its target. The malware identifies targets through social media indicators, including facial recognition, geolocation and voice recognition, and avoids detection until delivering its ‘payload’. It has the potential to operate across millions of devices and was demonstrated recently as a mechanism to distribute the Wannacry virus covertly through video conferencing apps.[34] This is just one example in an expanding range of offensive capabilities enhanced or facilitated by AI. Others include spear-phishing campaigns that harness big data for more targeted social engineering attacks; ‘hivenets’ – artificial intelligence enabled botnets that harvest data to compromise additional devices; extensive-tailored attacks – which are large numbers of targeted attacks conducted simultaneously through the application of AI; and advanced obfuscation techniques – including efforts to misdirect defenders by learning from data from past campaigns.[35]

[32] Kobie, N. (2018, September 12). To cripple AI, hackers are turning data against itself. Retrieved from https://www.wired.co.uk/article/artificial-intelligence-hacking-machine-learning-adversarial.

[33] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J. Erhan, D., Goodfellow, I, and Fer, R. (2014, February 19). Intriguing properties of neural networks. Retrieved from https://arxiv.org/abs/1312.6199.

[34] Smith, Ms. (2018, August 08). Weaponized AI and facial recognition enter the hacking world. Retrieved from https://www.csoonline.com/article/3296098/security/weaponized-ai-and-facial-recognition-enter-the-hacking-world.html.

[35] Artificial intelligence technologies boost capabilities of cyber threat actors. (2018, February 28). Retrieved from http://thetimesofafrica.com/artificial-intelligence-technologies-boost-capabilities-cyber-threat-actors/.

A related concern is that AI could be used to enhance information operations and target populations with the intent of causing instability or division. In that way, AI might be a multiplier or amplifier of information warfare. More generally, the use of AI in cyber operations poses many risks similar to those that have been identified with 'cyber weapons' (loosely defined as malware designed and intended to cause damage). These have been amply documented elsewhere, but include the ability of states and non-state actors to reverse engineer malware, collateral damage (Wannacry and Stuxnet spread to hundreds of thousands of computer systems in over a hundred countries), the dangers that investment in cyber weapons can create security dilemmas and arms races within the international system,[36] that cyber weapons can be stolen and reused,[37] and the fear that proliferation of AI to less restrained and less deterrable non-state actors may create heightened levels of danger and instability.[38] In this sense, concerns over the weaponization of AI within cyberspace are closely related to (although not necessarily the same) as the weaponization of malware for strategic objectives.

4. BATTLEFIELD AI? USE OF AI IN COMBINED ARMS OPERATIONS

While AI can be weaponized within and through cyberspace and has the potential to cause considerable harm when used with malicious intent within computer networks, the ability to integrate AI into existing weapons systems or deploy it on next generation military platforms is equally apparent. In this section we explore how AI might be used on the battlefield in combined arms operations to achieve strategic objectives.

At this juncture, there are two possible paths through which AI could be utilized in joint operations to generate military advantage: either it will be integrated within existing doctrines and battle concepts (*evolutionary* perspective), including being deployed to enhance existing capabilities, or to improve the speed of action and effectiveness of the human environment. Alternatively, the application of AI in the military field, either independently or in conjunction with other emerging technologies such as quantum computing, big data analytics, advanced robotics, human enhancement technologies, and automation, will lead to the development of new doctrines that defy the existing physical and legal boundaries of today's battlefield (*revolutionary* perspective). The application of AI into combined armed operations will likely depend more on the national models of inclusion of AI into the military field and the usefulness of this

[36] Buchanan, B. (2017). *The Cybersecurity Dilemma: Hacking, trust and fear between nations*. Oxford: Oxford University Press.

[37] Baram, G. (2018, June). The Theft and Reuse of Advanced Offensive Cyber Weapons Pose A Growing Threat. Retrieved from https://www.cfr.org/blog/theft-and-reuse-advanced-offensive-cyber-weapons-pose-growing-threat.

[38] Maurer, T. (2018). *Cyber Mercenaries: The state, hackers, and power*. Cambridge: Cambridge University Press.

emerging technology, rather than a general set of technical specifications. Application of AI in combined operations, however, will likely, at a minimum:

(a) Facilitate real-time analysis and improve situational awareness of the battlefield;
(b) Provide troops on the ground with actionable intelligence and enhanced decision-making;
(c) Facilitate dispersion or rapid concentration and application of lethal power, thereby enhancing mission precision and improving military effects;
(d) Act as a logistical aide by providing predictive maintenance and supply for military equipment, increasing the safety of operating equipment, reducing operational costs and thereby improving the readiness and deployability of troops;
(e) Enable robotics systems to serve a variety of military functions, including the use of lethal force;
(f) Fulfill jobs in the military that are dull, dangerous or dirty, including enhancing force protection and reducing casualties.

At a broader level, the effect of the application of AI in the military field will affect the balance of power at least through doctrinal changes and adaptations or through the creation of new capabilities; a new computer powerful enough to perform real-time big data analysis in ISR and discern actionable intelligence, for example. It will also affect the interplay between different levels of action, creating opportunities for tactical maneuvers (especially because of superior speed of decision and action) to have operational or even strategic effects, particularly through *fait accompli*, increasing strategic surprise and creating perceptions of first mover advantage (i.e. intensifying the security dilemma).

AI will likely create the conditions for the return of warfare operations ‘in mass’ again. Mass will become increasingly important, whether in data and intelligence or in actual capabilities deployed on the ground. In this context, as well as in the context of AI-cyber jointly, it is interesting to consider the idea of attrition: are these new capabilities likely to be used for attrition purposes or for disruption purposes, or both? This leads us to the question: is AI, together with cyber and a number of emerging technologies, likely to lead to the emergence of a new era of weapons of mass attrition or weapons of mass disruption? For example, active measures doctrine in Russia is a type of attrition in that it seeks to deplete the opponents’ sources of power (be it the integrity of their democratic institutions, the integrity of their information systems, and public support) but it may also act as a type of disruption, including disrupting the functioning of a national power apparatus and incapacitating the opponent from acting at the speed of relevance. Russia has not released a formal strategy for AI and is encumbered in some areas of technology by a lack of industrial and technological innovation, but its operational doctrine appears to suggest that the main current function of its AI capability is attrition – i.e. it is aimed at undermining the political cohesiveness and solidarity of the 'West' over time. That is not to say, however, that the Russian government will not use the technology for mass disruption, especially at a time of armed conflict and or international crisis.[39]

Interoperability will be increasingly affected by AI. Developing and deploying AI that is compatible across different branches of the armed forces will be challenging. The ability of two or more different AI-enabled systems to cooperate seamlessly in pursuit of combined mission objectives will be critical to achieve military advantages and mission effects. There are a number of states developing AI-enabled capabilities that have expressed an interest in maintaining interoperability with allies and partners,[40] but there are equally powerful protectionist forces in the defence industry which may present obstacles to seamless multinational interoperability.

Critical decision-making at the political level and on the battlefield will remain human in the age of AI. However, human-machine teaming and other blending solutions will enhance the application of power. Ultimately, it is unlikely that humans will be able to exert *full* control and authority over AI systems *at all times*. The notion that has been often stated on the military side of the LAWS debate, that there will always be an element of human control, appears to be fanciful in the current context. Trust will be an integral factor – military decision-makers will have to either trust from ignorance or from verification. In this context, testing and exercises involving AI and the generation of data pertaining to reliability and integrity will be paramount. This also raises questions about process, and how military decisions are made, including the centralization of command functions relating to AI. This is an old issue in many ways – centralized command structures have always had to adapt to the deployment of new battlefield technologies. In the field of AI, however, we believe it will be important to assess and resolve the balance between AI-based decision-making being distributed to commanders in the field, based on actionable AI generated intelligence, and the slower (but perhaps safer) centralization of AI command and decision-making.

Politics in this respect will be integral to outcomes in the deployment of military AI across domains. Strategy has always been the use of force to achieve political objectives, but we assert that politics will shape how AI is used as much as being the goal of the deployment of AI. What AI will not be able to do for combined arms operations, or any other type of operation for that matter, any time in the near future,

[39] Polyakova, A. (2018, November 16). Weapons of the weak: Russia and AI-driven asymmetric warfare. Retrieved from https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/.

[40] For example, the 2019 US Department of Defense AI Strategy, the EU's 2019-2020 Work Programme for the European Defence Industrial Development Programme and the 2019 Work Programme for the Preparatory Action on Defence Research reference interoperability in AI-enabled capabilities.

is lift the fog of war: the veil of uncertainty around the interests driving opponents' actions. It will not alleviate the security dilemma and may complicate arms control and disarmament efforts as barriers to entry are lowered due to the acceleration of technological progress in the civilian sector.

5. STRATEGIC IMPLICATIONS

The purpose of this article is not to provide definitive conclusions as to how AI will affect strategy. As Clausewitz often stressed, the unseen complexities involved in military affairs do not allow for clear answers.[41] The purpose of this paper is rather to enhance understanding of different aspects of what policymakers and military officials will face as AI technologies are integrated into war and conflict. In that spirit, we see several considerations as paramount to current and future strategy and policy.

The first is the requirement for and the simultaneous challenge of greater military-civilian fusion. We recognize this as a tautology that has always been true. However, it seems clear that militaries will need to develop much closer cooperation with the private sector in the development and use of AI technology through 'spin in' effects. China has already recognized this, as detailed by Elsa Kania in a recent report, and is working to fuse military and state-owned enterprise efforts to enhance China's AI capabilities and technologies.[42] In this respect, the extent to which China has an inherent advantage over the US because of state control of private enterprise is likely to influence the emerging power struggle over AI. China certainly has some advantages, including a productive and innovative economic and industrial base, and the clear articulation of national strategies around AI, but the notion that direct control over industry confers an advantage should be questioned. Much of historical innovation in technology has been derived from research conducted in private enterprises and research labs, sometimes with government funding. China's technological progress has also been driven, at least in part, by illegal appropriation of technologies and copyright theft, largely through cyber espionage. This has been amply documented.[43] The latest research suggests that China faces significant challenges in developing technologies due to the exponential increase in the complexity of military technology and the difficulties involved in replication and imitation.[44] In the US and Europe, conversely, the challenge will be to develop effective cooperation between the

[41] Otte, T. (2002). Educating Bellona: Carl von Clausewitz and Military Education. In G. Kennedy & K. Neilson (eds.), *Military education: Past, present, and future*. Westport, CT: Praeger.

[42] Kania, E. B. (2017, November). Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power. Retrieved from https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power.

[43] Laskai, L. L., & Segal, A. (2018, December 6). A New Old Threat: Countering the Return of Chinese Industrial Cyber Espionage. Retrieved from https://www.cfr.org/report/threat-chinese-espionage.

[44] Gilli, A., & Gilli, M. (2019). Why China Has Not Caught Up Yet: Military-Technological Superiority and the Limits of Imitation, Reverse Engineering, and Cyber Espionage. *International Security*, 43(3), 141-189. doi:10.1162/isec_a_00337.

military and private sector in the development of AI, while managing concerns around ethics and privacy. Recent reports suggest that the US military is now more trusted to develop AI systems than some of the big tech companies such as Google and Facebook, reflecting recent controversies around social media being used as a platform for AI-enabled information warfare and data privacy breaches.[45] However, it is our contention that technology must be jointly and collaboratively developed, and that military control of AI innovation will ultimately be counterproductive, largely because of the need to apply the technology across a wide range of societal activity.

Second, we expect that there will be an ongoing evolution (not revolution) from information warfare to intelligent warfare and that this process will define technology's use in conflict.[46] The outcomes of military conflict will not just be decided by who controls the information environment, but the application of AI to that information, to monitor it, to manipulate it, to degrade it and to harness it with the aim of achieving political ends. We recognize that there is no AI without information processing and that AI is already a social and collective technology that relies on information being fed into it. But the acceleration of this process as a result of big data trends is clearly significant. Battlefield commanders will need to gain an accurate view of the operational environment and achieve an understanding of how information flows through it, the extent to which AI systems can better inform military decisions, enhance insight, better predict what enemy forces might be planning, and minimize error. Access to information and large volumes of data will be paramount, and there will be increased competition, particularly in the early stages of military conflicts, over gaining access to and denying adversaries information.

Third, there will be a scale of human involvement depending on the military function. To express this simply, there will always be human control over AI pertaining to the deployment of nuclear weapons; authority is unlikely to be delegated to computers and algorithms at the high end and in the most destructive areas of military power. However, military decision-making and autonomous decision-making are likely to occur in other military functions such as logistics and situational awareness, for example. In this respect there is a spectrum of decision-making in AI and not a binary with humans involved or not. The novelty of AI should be noted here. We already have AI platforms – such as in the area of missile defense, the Israeli Harpy drone, and automated Russian tanks – that are fully capable of being autonomous, but they have not yet been fully deployed or relied upon. This is because of: (a) the fallibility of human control or decision-making; (b) the competition between states restricting the extent of deployment; (c) the lack of determination of the acceptable uses of AI; and

[45] Kahn, J. (2019, January 10). U.S. Military Trusted More Than Google, Facebook to Develop AI. Retrieved from https://www.bloomberg.com/news/articles/2019-01-10/u-s-military-trusted-more-than-google-facebook-to-develop-ai.

[46] Kania, E. B. (2017, November). Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power. Retrieved from https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power.

(d) the shadow of the future – i.e. fear of the normative and political consequences of AI's use in the battlefield.

Fourth, we expect that situational awareness both within computer networks and on the battlefield in tactical and operational environments will be considerably enhanced. There are already trials of battlefield AI that can significantly enhance the awareness that soldiers have of the environment, allowing them to be notified of enemy troop presence and movements, and these will lead to a more proactive approach to threat identification and mitigation. Mission control has always been based on sensing, perception, comprehension and prediction (battlefield situational awareness) and has always been meant to provide effective real-time decision support.[47] AI will accentuate the importance of these functions. Trials of these types of battlefield AI have already taken place, such as those developed by the Defence Science and Technology Laboratory (Dstl) and UK industry partners (SAPIENT).[48] Because of this, we expect that the role of humans in the battlefield will be reduced: drones, for example, have enabled us to place distance between ourselves and violence, and this trend will likely accelerate with advances in AI. Automated systems will be increasingly capable of doing the dirty work that soldiers used to do, and AI will enable commanders to keep forces out of harm's way more effectively.

Relatedly, while AI has been presented in certain debates (and certainly in *The Terminator* films) as posing a great threat to humankind, the prospect that 'killer robots' might take the place of human combatants is not without its benefits. Military commanders will likely be focused on harnessing AI to minimize danger, for force protection, and for deterrence as much as for offensive actions. In this respect, while the weaponization of AI is likely to be an ongoing driver of AI adoption in the military, the technology can clearly be harnessed to enhance security as well as destroy.

6. CONCLUSIONS

This article has sought to highlight some of the key strategic implications resulting from the weaponization of AI, but it is but one of a handful of early scholarly ventures into the strategic use of AI technologies. We are sure it will not be the last. The state of AI research in the strategic studies and security studies areas is still in its infancy. In the next decade, the literature is likely to expand, just as the cyber security literature has done in the previous decade. This will bring much-needed answers to questions over how AI will affect war, conflict, and strategy.

[47] Endsley, M. R. (2002). *Designing for situation awareness: An approach to human-centered design*. London: Taylor & Francis.

[48] Evans, V. W. (2018, September 24). Artificial intelligence weaponry successfully trialled on mock urban battlefield. Retrieved from https://www.telegraph.co.uk/news/2018/09/24/artificial-intelligence-weaponry-successfully-trialled-mock/.

Overall, we believe AI will continue to shape the battlefield and provide a driving force for the evolution of strategy itself as we move further into the 21st century. It will do so because AI systems will continue to be integrated into weapons systems and used to enhance the precision, lethality and destructiveness of the use of military force. Furthermore, AI will have varied and influential impacts on cyber defense and offense and is likely to continue to be weaponized – to be used with the intent to cause harm and damage – within and between computer networks. We see several other key impacts related to the emergence of AI. These include the magnification of the cognitive ability of military commanders, and, provided AI can be secured from intrusion and manipulation, that decision-making will become more intelligent and less prone to error. Again, this will be a revolutionary or evolutionary process depending on the task AI is set to perform and the domain it is activated in. Clearly the structure of militaries will also need to adapt to AI – especially as swarm technologies and multi-agent systems are developed – and new decision-making processes will need to be adopted. We are at the early stage of that process. Relatedly, constant attention will need to be given to the legal, ethical and strategic debates around human enhancement – including the physical and cognitive development and evolution of military forces, and how psychical and cognitive processes might change and evolve as weaponized AI is increasingly integrated into war fighting.

This leaves us with some big questions. Is weaponization desirable? Should the international community be seeking to control and stop these processes, and what effect might that have on non-military uses of AI? In this respect we believe that the sometimes hyperbolic debate about 'killer robots' somewhat misses the point. AI is already being weaponized and the debate about banning fully autonomous weapons systems ignores much of the other weaponization processes pertaining to AI that are already in full swing. A final point for further theoretical and scholarly reflection is what role AI will play in multilateral fora such as NATO, and how the use of AI within multilateral security missions will be shared and harnessed among contributing nations. Developing common operational standards, requirements and ethical guidelines for AI-enabled capabilities through NATO's Defence Planning Process (NDPP) and Science and Technology Organization (STO), or through the EU's European Defence Fund (EDF), Coordinated Annual Review on Defense (CARD) and Permanent Structured Cooperation (PESCO), will be both necessary and challenging.[49] NATO has taken a big step forward in announcing the use of offensive cyber operations by its members to support its missions,[50] but this leads to the question of how AI will be integrated into operations such as those in Afghanistan involving dozens of allies and partners deployed to highly complex, fractured intra-state conflicts.

[49] European Commission – the Independent High-Level Expert Group on Artificial Intelligence. (2019, April 8). Ethical Guidelines for Trustworthy AI. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[50] Ricks, T. E. (2017, December 07). NATO's Little Noticed but Important New Aggressive Stance on Cyber Weapons. Retrieved from https://foreignpolicy.com/2017/12/07/natos-little-noticed-but-important-new-aggressive-stance-on-cyber-weapons/.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Hidden in the Shadow: The Dark Web – A Growing Risk for Military Operations?

Robert Koch
Fraunhofer FKIE
Bonn, Germany
robert.koch@fkie.fraunhofer.de

Abstract: A multitude of leaked data can be purchased through the Dark Web nowadays. Recent reports highlight that the largest footprints of leaked data, which range from employee passwords to intellectual property, are linked to governmental institutions. According to OWL Cybersecurity, the US Navy is most affected. Thinking of leaked data like personal files, this can have a severe impact. For example, it can be the cornerstone for the start of sophisticated social engineering attacks, for getting credentials for illegal system access or installing malicious code in the target network. If personally identifiable information or sensitive data, access plans, strategies or intellectual property are traded on the Dark Web, this could pose a threat to the armed forces.

The actual impact, role, and dimension of information treated in the Dark Web are rarely analysed. Is the available data authentic and useful? Can it endanger the capabilities of armed forces? These questions are even more challenging, as several well-known cases of deanonymization have been published over recent years, raising the question whether somebody really would use the Dark Web to sell highly sensitive information. In contrast, fake offers from scammers can be found regularly, only set up to cheat possible buyers. A victim of illegal offers on the Dark Web will typically not go to the police.

The paper analyses the technical base of the Dark Web and examines possibilities of deanonymization. After an analysis of Dark Web marketplaces and the articles traded there, a discussion of the potential risks to military operations will be used to identify recommendations on how to minimize the risk. The analysis concludes that surveillance of the Dark Web is necessary to increase the chance of identifying sensitive information early; but actually the ‘open’ internet, the surface web and the Deep Web, poses the more important risk factor, as it is – in practice – more difficult to surveil than the Dark Web, and only a small share of breached information is traded on the latter.

Keywords: *Dark Web, military operations, data breaches, data leaks, data sale, marketplaces, anonymity, Tor, deanonymization, operational security, OPSEC, PII*

1. INTRODUCTION

The so-called Dark Web has been in the focus of the media in recent years, regularly in a negative context. With the takedown of the ‘Silk Road’ website in October 2013 by the FBI, the Dark Web entered the awareness of large parts of the population. In February 2015, the FBI took the infamous Dark Web site ‘Playpen’ offline, which hosted more than 23,000 child pornographic images and videos and had more than 215,000 users. As part of the preparation for the terrorist attacks in Paris in November 2015, the communication was anonymized by using the software Tor; while the weapon used in the shooting rampage in Munich in July 2016 was also acquired over the Dark Web. Beside drugs, weapons, and child pornography, every kind of information is sold via marketplaces on the Dark Web: from credit cards to sensitive information captured during data leaks or hacking attacks. The latter can pose new challenges for the armed forces.

Since sensitive data is repeatedly looted (see the overview of the world’s biggest data breaches (McCandless 2018)), the possibilities of the Dark Web can increase the motivation of attackers even further: based on the anonymity of the users, as well as the easy to use but (in the sense of the user, not fully traceable transactions) hard to track digital currencies like Bitcoin, illegal activities can be executed with apparently low risk for criminals.

To analyse the possible influence of the Dark Web on military operations, an overview is provided in Section 2, including an analysis of the technical background. Based on that, possibilities of deanonymization attacks are discussed; the security and reliability of the Dark Web may have an influence on the offered content. Next, an analysis of Dark Web marketplaces and the goods traded there is provided in Section 3, followed by a discussion of the resulting potential risks for military operations in Section 4. Finally, the main arguments of the paper are summarized in Section 5.

2. THE ONION ROUTER AND ANONYMITY

To understand the opportunities and weaknesses when using the Dark Web, some knowledge of how anonymization networks work is required. Therefore, terms with respect to the Dark Web are explained. These are often mixed, but must be clearly separated. This is followed by an investigation into the security levels of the Dark Web, since this is fundamental for an evaluation of the transactions to be expected there.

A. Terminology

Quite often, the terms *Darknet, Deep Web* and *Dark Web* are improperly mixed or used interchangeably. Due to insufficient separation and misuse of terms, data and evaluations can be incorrectly assigned and falsify the actual situation.

Deep Web. The Deep Web "refers to any Internet information or data that is inaccessible by a search engine and includes all websites, intranets, networks and online communities that are intentionally and/or unintentionally hidden, invisible or unreachable to search engine crawlers" (Janssen 2018). The term, Deep Web, "relates to deep sea/ocean environments that are virtually invisible and inaccessible" (Janssen 2018). Therefore, the Deep Web "contains data that is dynamically produced by an application, unlinked or standalone Web pages/websites, non-HTML content and data that is privately held and classified as confidential. Some estimate the size of the Deep Web as many times greater than the visible or Surface Web" (Janssen 2018).

Darknet. From a technical and historical point of view, the term 'Darknet' is used to describe the part of the IP address space which is *routable, but not in use*. This must be differentiated from addresses, which should not be routed by definition. In the still predominantly used internet addressing architecture, Internet Protocol version 4 (IPv4), specific addresses are defined as private.[1] By using them, a router can provide connectivity to numerous attached devices by using its own public address, translating the traffic between the private network and the internet. The respective private addresses are not visible on the internet; therefore, they should *not* be routed, and only routable addresses can be seen. By monitoring these unused but routable addresses, a lot of observations with respect to security can be made: normally, nobody should interact with them. So if some interaction can be seen, the underlying behaviour is typically malicious, e.g., an automated worm run looking for target addresses to infect. This security-relevant part of the address space is called the *Darknet*.

One of the early uses of the term with regard to digital content can be found in an article about content protection. It described Darknets as a 'collection of networks and technologies used to share digital content' (Biddle 2002). Nowadays, the term

[1] Subnetworks 10.0.0.0/8, 172.16.0.0/16 and 192.168.16.0/24, RFC1918.

is mainly used for o*verlay networks* providing anonymous network connectivity and services. An overlay network is a layer of virtual network topology on top of the physical layer, which directly interfaces with users (Zhang 2003). Tor is an example of an overlay network, and the biggest and most widely used anonymisation network; but there are numerous others, such as I2P, Freenet or ZeroNet.

It is important to recognize that the term *Darknet* originally refers to the network itself, and therefore the *technical base* like the protocol and devices; but not the content which may be transported through the network, or can be found on its respective servers.

Dark Web. The Dark Web refers to the websites which are hosted within overlay networks, and are *normally*[2] not accessible without special software like the Tor Browser. Nowadays, usage of the Tor network is easy and straightforward: the Tor Browser is a complete bundle ready to use without installation by providing a fully configured Firefox Browser. As in the case of the Deep Web, search engine crawlers are not able to index the websites of the Dark Web. But in contrast to it, its most important feature is that the users of a service stay anonymous - neither a provider of a website can identify the visitors, nor can a visitor identify the service provider. Given this, the respective services are also called ‘hidden services’; more recently, ‘onion services’.

B. Anonymity on the Internet

The history of privacy-enhancing technologies dates back to 1981, with a technique to hide the communicating participants of an electronic mail system and their messages (Chaum 1981). Since then, much work has been done in the area of anonymization techniques, with the Tor Project one of the most well-known. The acronym Tor stands for ‘The Onion Router’, based on the underlying principle of onion routing (Reed 1998). It was developed as a research project of the Naval Research Laboratory in the 1990s, with the purpose of protecting the online communication of US intelligence agents. The first pre-alpha of Tor was published in 2002 (Dingledine 2002). In 2004, the second generation of the system was published (Dingledine 2004), and the code released under a free licence.

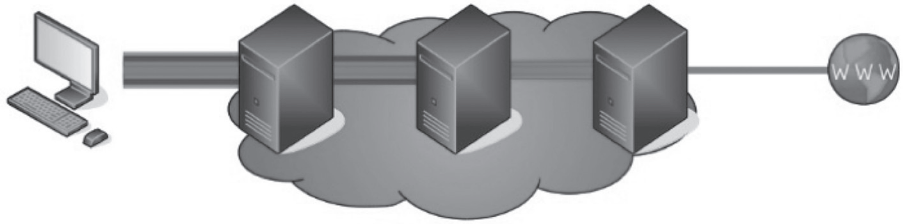
Becoming Anonymous. Two basic modes of application are offered by Tor: anonymous access to the internet, and onion services. In the first case, the traffic is routed through the Tor network and returns to the internet via so-called Exit relays. When accessing a website on the internet, it does not see the real IP address of the user, but that of the Exit relay; the IP of the user is not traceable. In the second case, the traffic stays *within* the Tor network: users can offer services like websites or instant messaging servers,

[2] Nowadays, there are also ways to access the content of the Dark Web without the use of special software. For example, the website tor2web.org enables browsing and accessing content on the Dark Web without the use of Tor software; though one must be aware when using this service that only the provider of the content stays anonymous, not the requesting user.

while others can access them via so-called ‘rendezvous points’. Both sides, the visitor as well as the service provider, stay anonymous.

To get a better idea of how Tor works, anonymous access to the internet is briefly described. Tor generates an overlay network in which each relay maintains a Transport Layer Security (TLS) connection to every other relay. Based on that, Tor establishes a circuit - a random pathway through the network - by selecting an Entry, Middle, and Exit relay.[3] The Exit relay is chosen based on a weighted random selection and changes regularly.[4] When sending data through Tor, the client encrypts it multiple times with the relays’ keys, including the predecessor’s and successor’s addresses for their respective relays. Each relay has the key for only one layer, uses the key to remove that layer, then forwards the data. In this way, it sees only the IP address of where the packet came from and where it must go. The Exit relay sends the packet to its final destination, which sees only the exit relay’s IP address. When the answer returns, each relay adds its encryption layer only the sender can finally remove them all and thus read the answer. Figure 1 visualizes the routing and anonymizing process of Tor.

FIGURE 1. FUNCTIONAL PRINCIPLE OF ONION ROUTING. EVERY RELAY ADDS RESPECTIVELY REMOVES ONE LAYER OF ENCRYPTION, AND ONLY KNOWS ITS IMMEDIATE PREDECESSOR AND SUCCESSOR.



Becoming Deanonymized. Due to the broad application possibilities of the Tor network, positive as well as negative/illegal ones, there is a strong interest in deanonymizing providers as well as users of onion services. For example, repressive regimes can try to locate those who use Tor for freedom of expression; while government agencies can try to fight illegal drug trafficking or child pornography. Therefore, many efforts to deanonymize users have been made and three basic categories can be identified, which will be explained briefly:

CAT 1 The first category includes attacks at the technical level. This is the most dangerous, but in practice also the rarest type of deanonymization attack. These can be

[3] Tor can extend the circuit by adding relays; but a circuit typically has only one Middle relay, so that communication latency remains at an acceptable level.

[4] By default, the circuit for a new TCP stream is rotated all 10 minutes to avoid profiling attacks; long-lasting single TCP streams (e.g., an IRC connection) are not rotated and will stay on the same circuit (Tor 2015).

directed against implementation flaws of the Tor software, but also attack weaknesses in the design of the network protocol of Tor. Attacks based on actual technical shortcomings of Tor are rare, but can have severe impact. An important example is the 'relay early' traffic confirmation attack, which was identified and executed between January 30, 2014 and July 4, 2014 by the Software Engineering Institute of Carnegie Mellon University (Dingledine 2014). The identified IP addresses were subpoenaed by the FBI and used in the trial against Brian Farrell:

> The record demonstrates that the defendant's IP address was identified by the Software Engineering Institute ("SEI") of Carnegie Mellon University (CMU") [sic.] when SEI was conducting research on the Tor network which was funded by the Department of Defense ("DOD") […] Farrell is charged with conspiracy to distribute cocaine, heroin, and methamphetamine due to his alleged role as a staff member of the Silk Road 2.0 dark web marketplace (Cox 2016).

CAT 1b Another attack on a technical basis is much more common – but not directed against the Tor software or the protocol itself, but against *the used browser*. While Tor can be used with any browser, this must be configured accordingly. The Tor Browser, which is based on a Mozilla Firefox browser, makes this much easier, as it just needs to be downloaded and started; it is preconfigured and no installation is required, which should make it particularly attractive to many users. Therefore, *vulnerabilities of the browser* can present an interesting target and be exploited to deanonymize the users. A famous example is the shutdown of the 'Playpen' Dark Web child pornography website by the FBI in February 2015. The FBI used a so-called 'Network Investigative Technique' (NIT), which was exploiting a non-publicly-known vulnerability of the Mozilla browser to break into suspected visitors' computers and identify their real IP addresses (Cox 2016). Instead of shutting down the website, the FBI continued to run it from a government server for 13 days to collect the IP addresses of potential visitors. In further action, the FBI broke into more than 8700 computers in 120 countries due to a court decision of a single judge. The procedure was heavily criticized. Of the 100,000 people worldwide who visited the site, 8700 were hacked but only 214 were arrested.

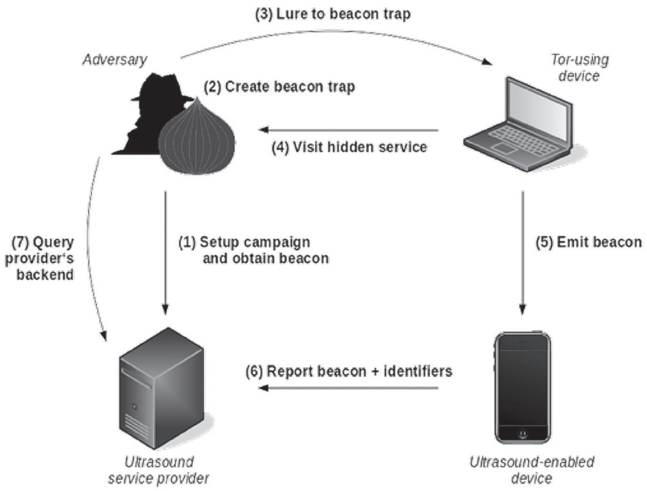
Because of deanonymization attacks like that one, the Tor Project provided a hardened version of the Tor Browser, beginning from November 2015 (Tor Browser 5.5a4-hardened), providing additional hardening against the exploitation of memory corruption bugs and adding debugging features. Anyway, in part because of, *inter alia*, the confusion among users caused by the two series, regular and hardened, the second one was discontinued in April 2017.

CAT 2 The second attack category is not based on technical characteristics of the Tor Browser or the respective protocol, but exploiting *indirect shortcomings, which are not based on technical vulnerabilities*. A prominent example is the use of default configurations: on most distributions, the Apache server ships with a feature called `mod_status` enabled, which provides a website at `/server-status`, containing statistics like resource usage and virtual hosts. For security reasons, this page is by default only reachable from localhost. Yet the Tor demon for onion services *is* running on localhost, which allows connections to the status page from external clients if the configuration is unchanged. Due to this, sensitive information can be leaked; even a .onion search engine was identified as having the module enabled, exposing *all* search queries sent to the page.

Another example highlights the endangerment of the indirect attack vectors included in this category even better: back in 2014, a new advertising technique called ‘ultrasound cross-device tracking’ (uXDT) was deployed. The idea behind uXDT is embedding unique sound codes, inaudible to humans, into advertisements. The inaudible sounds are replayed when the ad is presented to a user. Unknown and unrecognizable to the user, the sound pattern may be noticed by another device nearby. Software supporting uXDT is listening for such patterns; if it recognizes one, it sends it back to a central server - together with information about the device. The central server knows the pattern as it was created in a unique way, and therefore knows the targets to which it was sent. In this way, it is possible to identify and merge multiple devices owned by a user, optimizing ad campaigns to all their devices, even if they were never involved in an action like searching for a specific product, resulting in a purposive ad.

Even worse, this technique can be used for deanonymization attacks on Tor users as well (Mavroudis 2017). If someone enters the Dark Web, they will quickly recognize there are a lot of ads, for example embedded in well-known search pages and even in popular marketplaces. Using the default configuration of the Tor Browser, these ads are presented to the user. Therefore, if someone opens a web page which presents an ad with an embedded uXDT sound, there is the risk that a device nearby, maybe a smartphone, another computer or even one of the numerous IoT gadgets which are now so popular, is listening. By applying the same technique, sending back such a unique beacon trap to a central server, the attacker can directly merge the anonymized access to the regular, public connection, and easily deanonymize the user. Figure 2 illustrates the attack scheme.

FIGURE 2. ULTRASOUND TRACKING BASED ATTACK SCHEME TO DEANONYMIZE TOR USERS. VISUALIZATION BASED ON (MAVROUDIS 2017).


(3) Lure to beacon trap
Adversary
Tor-using device
(2) Create beacon trap
(4) Visit hidden service
(7) Query provider's backend
(1) Setup campaign and obtain beacon
(5) Emit beacon
(6) Report beacon + identifiers
Ultrasound service provider
Ultrasound-enabled device

These two examples highlight the wide range of opportunities through which Tor users can be deanonymized if they are not extremely careful when using the network. CAT 3 In fact, user mistakes and human behaviour are the most common reason for deanonymization. A prominent example is the shutdown of the Dark Web marketplace, "Silk Road", which specialized in drug trafficking and was one of the first of its kind on the Dark Web. The creator, Ross Ulbricht, who used the pseudonym "Dread Pirate Roberts", revealed himself by several momentous mistakes. First, he used the pseudonym "altoid" to announce and promote his marketplace in early January 2011. In October of the same year, the same pseudonym was used for a post on a Bitcoin talk, and his email address was included as a contact opportunity for interested users: rossulbricht@gmail.com. This was discovered by the authorities, enabling them to trace Ulbrich back, eventually resulting in his imprisonment.[5] Blake Benthall failed to heed this; he was arrested in November 2014 for establishing and running the Silk Road 2 marketplace, after the first one was closed. Benthall could be identified because he registered the server where the anonymous website was running with his email address, blake@benthall.net; the same category of mistake as that of Ulbricht.

Another example was an online drug dealer, caught in 2017 because he was conspicuous at the post office. To avoid fingerprints, he always delivered the postal packages wearing latex gloves at the counter. However, this eventually caught the attention of the postal employees, so they informed the police. When the dealer was

[5] Also, there was a report that Ulbricht ordered several fake IDs to rent the required servers for the Silk Road website. The fake IDs were sent from Canada to the US, and found at the border as part of a routine mail search. The packet contained nine fake IDs - each with a different name, but all of them with the same photo: a *real* photo of Ulbricht. As the packet was even addressed *directly* to Ulbricht, that was another low-hanging fruit for the officers. However, the careless handling of the pseudonym 'altoid' seems to have been the root cause of the identification.

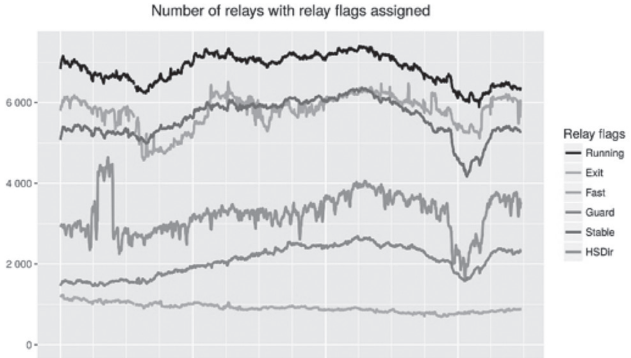
arrested, further traces on his mobile phone linked him to an entry on a Reddit website about drug dealers on the Dark Web. He had not deleted the history.

As an interim conclusion, it can be stated that the protection afforded by the Dark Web for criminal activities can be quickly lost through numerous possibilities of deanonymization. This can involve particularly careless behaviour by users, but can also be originated by attacks on the software or the protocols.

Since the need for secure anonymization can be anticipated when dealing with information relevant to military operations on the Dark Web, a closer look should be taken at the functional principles and their weaknesses. In particular, the question arises whether the Dark Web offers sufficient protection when used cautiously.

Traffic Analysis and its Relevance. To answer that question, a closer look at the working scheme of the overlay network, and the resulting possibilities of deanonymization without an exploitation of protocol and programming vulnerabilities, should be taken. As such an analysis would go beyond the scope and technical depth of this article, only a few key findings are outlined as follows. Tor is the largest, most widely used anonymization network; yet it has the problem that the number of relays in the network is relatively limited and barely growing. In some cases, it can even be observed that the number of relays involved is decreasing – which may also be due to legal reasons.[6] However, there are also relatively spontaneous, very large changes in the number of specific relays – often a sign that an attack on the Tor network is being attempted again, or that research institutions or other bodies are trying new analyses. Figure 3 shows the development of the number of relays since January 2015, as provided by The Tor Project (Tor 2018). In particular, the Exit relays stagnated for years and only increased again recently.

FIGURE 3. DEVELOPMENT OF THE NUMBER OF RELAYS SINCE JANUARY 2015 (TOR 2018).

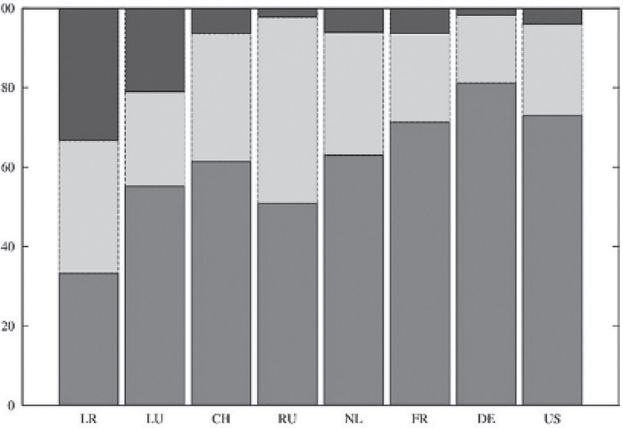


[6] For example, because of violations of copyright infringement when the Exit Nodes are misused (Ferner 2017).

Some areas of the curves are striking: a sudden, rapid increase in the number of Hidden Service Directory (HSDir) relays can be observed from mid-April 2015 until the end of May 2015. On the other hand, a sudden drop of HSDir and Stable relays can be identified in December 2017: this was affected by a DDoS attack on the Tor network. Multiple servers went down because of the attack; the HSDir relays were badly affected, because if such a system goes down, it does not get back the HSDir server flag immediately after rebooting, but takes 96 hours. The loss of HSDir relays also affected the reachability of onion services (Goulet 2017). Other strong jumps in the number of relays may also be related to, e.g., C&C infrastructure ran over the Tor network or bots.

As we can see, only a small number of relays are providing the core functionality of the Tor network, and the chances are high that they include quite a number of malicious ones. Moreover, not only is the number of Exit relays already quite low, but the way they are selected by the underlying algorithms reduces the actually used relays significantly. Figure 4 provides an example of the actual Exit relay use per country relative to available Exit relays based on a three-week observation (Koch 2016). Each bar shows the ratio of available Tor relays (red) to relays configured as Exit relays (green) to selected Exit relays (blue). Nearly a quarter of all nodes were located in the US, but Tor selected only 5.53 per cent of these (blue section of US bar). Likewise, 8.53 per cent of all exit nodes were located in Germany (green section of DE bar), but Tor selected only 2.22 per cent of these (blue section of DE bar).

FIGURE 4. RATIO OF AVAILABLE TOR RELAYS TO EXIT RELAYS TO SELECTED/USED EXIT RELAYS. THE SMALL SHARE OF ACTUALLY USED EXIT RELAYS SIMPLIFIES TRAFFIC ANALYSIS ATTACKS (KOCH 2016).



It can be seen that only a small fraction of the available Exit relays is selected and used. This simplifies attacks that analyse traffic flows through the Tor network, as the number of relays to be monitored drops sharply. But not only Exit relays are endangered. With respect to onion services, malicious HSDir relays can be used to identify new onion services on the Dark Web. For example, more than 100 snooping HSDir relays were identified on the Tor network (Noubir 2016) – a technique typically used by companies providing Dark Web intelligence, or by federal agencies.

These intense activities of various actors, which aim at the analysis of actions up to the deanonymization of Dark Web users, should be kept in mind ahead of the further discussion.

3. DARK WEB MARKETS AND DATA

Based on the knowledge of the function, opportunities and weaknesses of anonymising networks, an analysis of Dark Web marketplaces and their trading is performed, before specifically looking into the trading of sensitive information.

A. Data Economy and Marketplaces

Of course, a central aspect of the question whether the Dark Web is a growing risk for military operations involves the nature, extent and quality of information which can be found there. While crawling the Dark Web can be challenging, e.g., finding new websites or entering closed marketplaces, DARPA's Memex program sought to develop software to advance search capabilities, especially with regard to the *Deep* Web, and a series of tools was made public (DARPA 2014). Some studies tried to shed some light by analysing onion services in the Dark Web provided by Tor. e.g., 39,824 hidden service descriptors were analysed on 4 February 2013 (Biryukov 2014). After scanning the hosts, 3,050 HTTP services were identified, and the content classified. Only hidden services offered in the English language had been analysed: 2,618 services in total. From these pages, 805 showed a default page and no actual content; 44 per cent of the identified topics were devoted to drugs, adult content, counterfeit, and weapons, while 56 per cent were devoted to topics like politics.

Another study identified a share of 57 per cent in services with illicit content (Moore 2016). The used categories are shown in Table 1.

TABLE 1. CATEGORIES AND ACCESS NUMBERS OF CONTENT IN THE DARK WEB (MOORE ET AL. 2016).

Category	Category
None	2,482
Other	1,021
Drugs	423
Finance	327
Other illicit	198
Unknown	155
Extremism	140
Illegitimate pornography	122
Nexus	118
Hacking	96
Social	64
Arms	42
Violence	17
Total	5,205
Total active	2,723
Total illicit	1,547

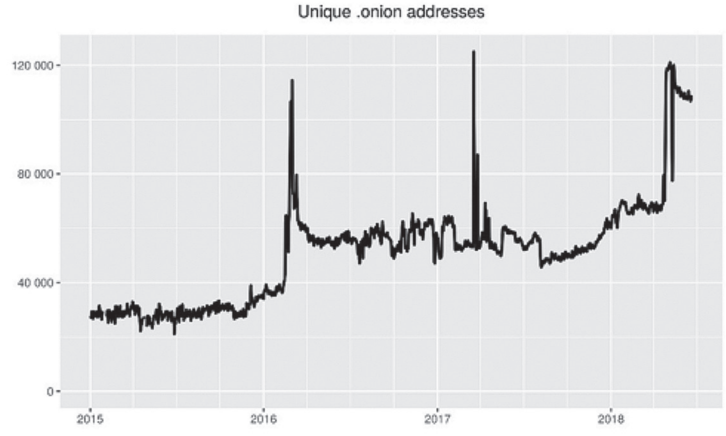
Repeatedly, it is argued that most parts of Tor traffic are illicit; the rough numbers seem to confirm this. A study presented by the University of Portsmouth even highlighted that 80 per cent of traffic to Tor hidden services is related to child pornography. While these are shocking results at first glance, a closer look at the underlying data reveals that the corresponding values are highly uncertain and only marginally justify such statements: based on the nature of the Dark Web, respective measurements can typically only be made indirectly. Regularly, requests to (malicious, therefore especially set up for the measurement task) hidden service directories will be counted. The respective numbers are often used to derive relative numbers of users; but they say more about the behavioural differences of different types of users (Mathewson 2014). Another important and often unnoticed aspect is that child protection agencies also regularly crawl the Dark Web for websites containing illicit pornography. Law enforcement agencies do so too. Therefore, it is interesting to look at the evaluations of these agencies to get a better idea of the actual situation. The results presented in the recent reports of the Internet Watch Foundation (IWF) are highlighted in Table 2.

TABLE 2. URLS CONFIRMED CONTAINING CHILD SEXUAL ABUSE IMAGERY AS SEEN BY THE IWF (IWF 2015, IWF 2016, IWF 2017).

Year	URLs to Child Porn	Hidden Service Proportion	Proportion of the Dark Web
2015	68092	79	0.116
2016	57335	41	0.071
2017	78589	44	0.056

We can see that the number of identified hidden services related with child pornography is small in contrast to the actual identified links to websites with child pornography in the surface or Deep Web. The IWF highlights that ‘hidden services commonly contain hundreds or even thousands of links to child sexual abuse imagery that is hosted on image hosts and cyberlockers on the open web’ (IWF 2017). This must be combined with the fact that the Dark Web is very small and growing only very slowly. Figure 5 shows the number of unique .onion addresses between January 2015 and June 2018. There is only a slow increase in the number of onion services; and the numbers are often quite constant over longer periods of time, sometimes even declining. Very fast, large increases are typically indicative of an experiment or attack and do not represent a sudden increase in the number of available pages. It should also be noted that nowhere near all pages have content; many only present the default page of the web server, such as that already shown in the above-referenced analyses.

FIGURE 5. NUMBER OF UNIQUE .ONION ADDRESSES FOR SERVICE VERSION 2 FROM 1 JANUARY 2015 TO 23 JUNE 2018.



The Tor Project – https://metrics.torproject.org/

Independent from the number of onion services marketplaces, but very important, is the trading volume. Some calculations have been made of the sales volume of the ecosystem, including several famous and heavily used marketplaces like Silk Road, Black Market Reloaded and Silk Road 2.0 (Soska 2015). The trading volume was higher than previously thought, and is also subject to strong fluctuations. However, the total volume does not experience exponential growth. The study identified that "in the short four years since the development of the original Silk Road, total volumes have reached up to $650,000 daily (averaged over 30-day windows) and are generally stable around $300,000-$500,000 a day, far exceeding what had been previously reported" (Soska 2015).

It is important to keep these dependencies in mind, as it is the base from which to focus on the most significant aspects. Looking at leaked data, most occurrences are on the clear internet – and while there may be trades of the data on the Dark Web, the result normally provides a link to a page in the surface or Deep Web, where it can be found and downloaded; but normally, it is not hosted on the Dark Web. Paste services like pastebin are popular for that.

We can conclude that the growth and therefore, the evolution of the importance of and danger posed by, the Dark Web is often over-estimated. In particular, the sometimes assumed exponential growth of the Dark Web cannot be demonstrated by any measurable numbers: neither the number of onion services and Dark Web marketplaces, nor the traffic itself, nor the trading volume.

B. Trading Sensitive Data

Looking at the most important trading categories of the Dark Web marketplaces: drugs, counterfeit and adult, most of them are not really able to affect military operations.

Some companies are offering Dark Web intelligence, highlighting the footprints of companies on the Dark Web, based on data they find. For example, OWL Cybersecurity published a so-called 'Darknet [sic.] Index' which aims to measure how the availability of breached data affects the overall cybersecurity of a company (OWL 2017). For this purpose, OWL Cybersecurity has set up a database, which is "automatically and continuously updated with between 10 to 15 million pages per day, from more than 24,000 domains on the Tor network alone, as well as other darknet networks" (OWL 2017). It highlighted that every company in the 2017 Fortune 500 is exposed on the darknet [sic.]; the companies with the largest footprint are shown in Table 3.

TABLE 3. TOP 10 ENTRIES OF THE 'DARKNET [SIC.] INDEX' FOR THE FORTUNE 500 COMPANIES PRESENTED BY OWL CYBERSECURITY (OWL 2017).

DARKINT Rank	Company Name	Darknet [sic.] Index Score
1	Amazon.com	19.16
2	Alphabet (Google)	17.21
3	Apple	15.98
4	Facebook	14.99
5	eBay	14.55
6	American Express	13.33
7	Frontier Communications	13.29
8	Netflix	13.19
9	Texas Instruments	12.99
10	FedEx	12.58

OWL Cybersecurity presented additional evaluations focusing on specific sectors, e.g., for IT companies. Moreover, based on the Fortune 500 evaluation, it analysed the US government to compare the results with the commercial sector. Key points of their conclusions are that the "U.S. Government scored worse than expected as compared to the largest U.S. companies. The U.S. Government averaged 1.6 points higher than the average Fortune 500 company, meaning that the government has a comparably larger amount of darknet exposure" (OWL 2017). The analysis identified that the US Navy has the most extensive footprint of all government agencies examined, and that

> military and defense groups overall are the largest target, closely followed by Cabinet agencies. A target's attractiveness stems from the desirability of its protected information. Whether personal or proprietary, it would appear that the groups more closely linked to defense have data that cyber criminals find attractive (OWL 2017).

To what extent these footprints represent a real threat to the company in question is not easy to estimate. That the footprints of state organizations are very large is fundamental here. Table 4 presents the force numbers by service branch for 2016, as published by the DoD in December 2017.

TABLE 4. FORCE NUMBERS BY SERVICE BRANCH AND RESERVE COMPONENT FOR 2016. SOURCE: DOD, DECEMBER 2017.

Branch	Employees
Army Active Duty	471,271
Army National Guard	344,862
Navy Active Duty	320,101
Air Force Active Duty	313,723
Army Reserve	306,272
Marine Corps Active Duty	183,501
Navy Reserve	108,864
Marine Corps Reserve	106,581
Air National Guard	105,887
Air Force Reserve	104,520
Coast Guard Active Duty	39,597
Coast Guard Reserve	8,123
Sum	2,413,302
Active	1,778,942

In addition to these numbers, associated authorities, civilian employees, etc. must be added to the reflected attack surface. In comparison, only Walmart has 2.2 million employees, far more than any other Fortune 500 company. Next are McDonald's (420,000), IBM (412,000) and Kroger (400,000), while the average number of employees at the Fortune 500 companies is about 50,000. Given this, leaks with elements affecting one or another employee of the governmental sector are likely and possibly adding to the footprint. Therefore, there may not be a *direct* risk for a company; but of course, there is always the risk of social engineering attacks.

With respect to the data available on the Dark Web, it can be assumed that an evaluation of the importance or possible impact is usually very difficult. While extensive reputation systems have been established in the area of illicit drug trafficking or trading in stolen credit card numbers, this is not so easy for the trade in leaked data. Typically, the data will often come from different sources and sellers will be unknown. Here, we can look at other areas of the Dark Web struggling with similar 'problems': the arms trade and hitman services. There are multiple Dark Web websites offering these services. Yet such is the nature of the Dark Web, many scams can be found: since a buyer of illegal weapons or the client to a murder can hardly go to the police after they have paid, but have not received what they were promised, scammers can earn easy money here.

A prominent example is the 'Besa Mafia' website. While the page was very well set up and many discussions focused on the question of whether it was real or not, eventually it was shown to be a scam. The scammers had been able to collect money from different potential customers, but never executed an assassination (Jeffries 2017). Also, according to federal investigators, Ross Ulbricht ordered six murders over the Dark Web; but five never happened, and the sixth turned into an indictment because the supposed hitman was actually a federal agent working undercover (Jeffries 2017).

The same applies for the illegal arms trade. While it is possible to buy a weapon on the Dark Web, it is actually quite difficult, as the case of the Munich shooting rampage has shown. A study analysed the role of the Dark Web in facilitating trade in firearms, ammunition and explosives (RAND 2017). After collecting one week of data during September 2016,[7] it was systematically analysed and discussed in workshops and interviews. RAND concluded that the Dark Web is an enabler of the circulation of illegal weapons but also highlighted the limitations of the study, especially "the impossibility to determine with certainty the nature of a vendor (scammer, law enforcement or real vendor)" (RAND 2017). Some verified examples like the Munich case are mentioned, but the number is very small. Moreover, in terms of the weapons trade, the activities of scammers and undercover cops supersede real offers by far. For example, Agora stopped selling guns altogether when it was the largest market on the Dark Web, because of "scamming by dishonest vendors" (Cox 2015). Of course, the trade in 3d-printing plans is much easier to do and can lead to increasing proliferation.

Taking a look again at data that may have an impact on military operations, direct and indirect effects have to be differentiated. For example, trading in mission plans or classified reports and evaluations, as well as access credentials to systems or services, can generate a direct impact; while personally identifiable information (PII) can generate an indirect impact.

However, based on the available reports and experiences, it can be assumed that trading data like mission plans and classified reports is not easy and not very likely on the Dark Web. Sales on the Dark Web are mainly financial data, login access, access to online services and identities including fake IDs like passports (Ablon 2014, McFarland 2015, Ray 2017). The same applies for the governmental sector: PII is the most compromised record type, counting 57.4 per cent of available data from breaches in the governmental sector (Huq 2015).

Although evidence can be provided about the authenticity of the data – for example, the provision of individual screenshots or excerpts from documents – due to its peculiarity (as opposed to the dumping of credit card numbers, etc.), the sale will be much more difficult, and will attract undercover agents. Rather, it can be assumed

[7] 19-25 September 2016.

that such data is traded outside of the Dark Web, in the traditional way. For instance, while access to some SCADA systems was offered for sale on the Dark Web in 2015 (Aharoni 2015), three years later, this is still a rare case and not yet a new trend. More likely is trade in credentials or PII as part of leaks, which may not even be directly affecting the military, but indirectly affects its personnel. Another indirect impact may also be generated by more inconspicuous services available on the Dark Web: namely, the proliferation of attack tools regarding knowledge, which then can be used to implement and execute attacks on military communication systems:

- Weaknesses, 0-days, 1-days
- Exploit code
- Malware frameworks
- Ransomware as a Service (RaaS), Crime as a Service (CaaS)
- Botnet access/rent for the execution of DDoS attacks
- Jamming devices

These categories may pose a special, indirect danger for military operations. While this is no direct trade in mission-critical information, specially crafted malware used in social engineering campaigns, or the offer to hack social media accounts can be a starting point to access a mission-critical environment. There are regular data leaks available; and hence, a lot of PII with which to identify potential targets: with numerous servicemen and women possibly affected, too. For example, the xDedic marketplace is offering easy access to legitimate organizational servers; different advertisements for hacking email or social media accounts can be found (Paganini 2017).

Based on this broad background – the technical functioning of Tor and the possibilities of user deanonymization, the activities which can be observed in Dark Web marketplaces and a realistic estimate of their importance compared to the surface and Deep Web – the actual risk to military operations from the Dark Web can now be discussed.

4. DISCUSSION

Several studies have been published highlighting the apparently predominantly illegal use and content of the Dark Web; but this only holds true at first glance. The actual numbers show that criminal activities committed on the Dark Web are only a very tiny portion, while a vast amount happens on the surface web and the Deep Web. In fact, the Dark Web page provided by Facebook at facebookcorewwwi.onion to allow users in countries with surveillance and repression to access the service is the most widely used site on the (Tor) Dark Web.

Dark Web marketplaces can have several hundred thousand dollars in sales per day, but the focus of trade is drugs and financial fraud, while a lot of PII is traded, too, which can be the enabler for social engineering and targeted attacks. Even more, CaaS with offers like hacking social media accounts are services which we must consider. Accordingly, a threat to military operations may result if social media or system accounts of soldiers are hacked in order to gain access to a target system. The trade in PII from data leaks can additionally support this. Nevertheless, the process is time-consuming and long, opening various options for detection and early warning.

The greatest threat seems to arise if PII is not made available for sale but publicly available. Automatically monitoring the relevant forums and pages is relatively easy for a tech-savvy user to do, so data deployed there can be used very quickly for (especially) social engineering attacks, often before those affected have heard of the original leaks. For example, the recent so-called Germany-Leaks, including details of German lawmakers up to Angela Merkel, were distributed by a hacker with the pseudonym 'Orbit' in December 2018, with subsequent comprehensive media coverage in the beginning of January 2019 (Times 2019). The original links are no longer available, but the material and alternative links still can be found quickly on corresponding websites on the Dark Web. In this context, it should also be mentioned that on the same forum where this data and other leaks were provided, no military-related record or post could be identified.

In addition to the requirement to first find respective leaked data, the question is also whether a targeted attack against a *particular* mission will be feasible – or whether 'only' an endangerment of a 'random' mission may arise. Moreover, the past few years have repeatedly shown police operations in which Dark Web marketplaces were shut down and those responsible were held to account. Studies on the Dark Web also continue to regularly show that a high proportion of the nodes involved are run by governmental agencies, research laboratories and universities; and numerous monitoring measures are implemented. For example, there are also fingerprints for the website and distribution 'TAILS' in the xkeyscore monitoring program of the NSA: if an attacker succeeds in manipulating the Tor Browser or a relevant distribution during the download – for example, inserting a backdoor – anonymity can be broken from the beginning. The numerous incidents and attacks which are known about, and extensive research on the topic of deanonymization, all make it questionable if someone is willing to sell sensitive data which is important for the success (or failure) of a military operation on the Dark Web – and equally, whether another party is willing to buy it there.

On the other hand, it should also be noted that the security of onion services will increase significantly in the near future – and thus the effort to deanonymize the services or their users will become more challenging. This is due to the recent introduction of onion services version 3. Currently, companies providing tracking and intelligence services for the Dark Web benefit highly from design weaknesses of long-used hidden services of version 2. For example, placing malicious HSDirs is a very popular and heavily used technique to identify new services in the (hidden service v2) Dark Web. Recently, researchers found more than 100 of these malicious HSDirs, reflecting the intense activities of companies providing Dark Web intelligence, as well as researchers and public authorities. With the availability of the new onion service v3, the exploited design shortcomings of the predecessor are fixed. Several design decisions and measures guarantee much better protection of users than before, and thus a much higher degree of anonymity. This is realized, among other things, by the following properties (Tor 2013, Tor 2017):

- Use of stronger cryptographic building blocks: SHA3/Ed25519/Curve25519 instead of SHA1/DH/RSA1024 in version 2
- Improved directory protocol with less metadata leaked to directory servers
- New pseudo random variables to prevent predictable Tor uses
- Better onion address security against impersonation: new addresses with 56 characters instead of 16 characters in version 2
- A cleaner, more modular code base

Therefore, tracking opportunities for the companies mentioned above decrease significantly, while attacks on services are more challenging. With the new name space of the services and the protocol adaptions, finding new, as yet unknown pages on the Dark Web will become much more difficult. This could again lead to much greater use of the Dark Web for criminal activities, but the question is: what kind of activities?

When talking about data which can pose a risk to military operations, there are two scenarios: a ‘random’ hack or a ‘targeted’ hack. If a hacker obtains the data more or less by chance, they will also offer it more visibly in order to make money; available contacts to interested parties are not to be expected here. This increases the likelihood of detecting traces of sensitive data, even on the Dark Web, in a timely manner. However, in the case of a targeted attack, possibly even controlled by a state, the interested party is clear; and a particularly visible offer is unnecessary and unlikely.

In the case that mission-critical information is available on the Dark Web, another thought must also be taken into consideration: finding and recognizing it may not be enough, or may be too late with respect to a current mission. While early detection of a new set of credit card numbers available for sale on the Dark Web can be used to disable and exchange the affected cards, protecting customers from financial damage even before the data can be exploited, this can be much more difficult with respect to an ongoing operation. Therefore, another approach can be beneficial too: the deliberate introduction and monitoring of honeydata: consciously placed, realistic looking records.

Based on these considerations, a comprehensive data management strategy must include the following elements:

1. Continuously tracking the surface and Deep Web as well as the Dark Web for the appearance of new leaked and stolen data. This requires the creation of fingerprints (hashes) for sensitive files, which then can be used to search for leaked data on the surface web as well as the Deep and Dark Web. Here, services like PwnedList can be integrated too.
2. The implementation of honeydata to increase detection probabilities.
3. The preparation (and testing!) of action plans and guidelines for fast, accurate handling of detected data leaks, including procedures to initiate the deletion of data from typically used platforms like pastebin.

Another aspect involves using the Tor network in essentially the way it was invented for – to hide the communication and identity of agents. Offensive actions may be executed by using anonymization networks like Tor; but as the analysis has shown, it is quite easy to monitor the Exit Nodes and very easy to blacklist them. Therefore, monitoring the IPs of the Exit Nodes can be used for an early warning if someone is willing to execute an attack over the Tor network.

Summing up these arguments, we can conclude that the new, more secure anonymous onion services will certainly lead to an increase in the popularity of illegal exchanges, but sensitive data important for military operations will still not be the focal point. More dangerous is the overall trade of data from breaches and leaks, which *may* contain details connected to the military; and in the broader sense, to military operations. For example, data records from dating agencies or sports applications may be assigned to soldiers, which can make them targets for social engineering, blackmailing or just make them (and therefore, their unit) trackable. While such information can be an element in a much broader mission to eventually influence a military operation, the risk factor is significantly lower than in the case of directly trading data on such operations. For the military, this means that a threat intelligence capability, monitoring potential risks associated with data breaches, is increasingly important. The main focus remains on the surface and especially the Deep Web; but monitoring the Dark Web is also beneficial.

5. CONCLUSION

Anonymization networks like Tor can be used to hide someone's identity or trade illegal goods on the Dark Web. Numerous data-related incidents and the trade of the corresponding records represent an increasing challenge. The availability of specially crafted malicious software or CaaS over the Dark Web can also generate new risk potential.

On the one hand, a closer look at the Dark Web, its technical base and the available data identifies no direct endangerment of armed forces capabilities. Scammers, law enforcement and surveillance opportunities do not make the Dark Web a reliable vector for sophisticated attackers. Therefore, monitoring the Dark Web does not play a superior role; the main activities, which can pose a risk for military operations, take place on the surface and the Deep Web. On the other hand, due to the multitude of available PII, which can also affect servicemen and women when being used for, say, social engineering campaigns, timely detection of sensitive information is of particular importance. While such data cannot be routinely targeted against an operation or military capability, it can open access to somewhere in the system and thus be the beginning of a longer attack path. Accordingly, it is important to monitor all parts of the web continuously through a holistic strategy, and develop and regularly practise emergency plans for rapid response to recognized data loss.

REFERENCES

Ablon, L., Libicki, M., and Golay, A. 2014. *Markets for Cybercrime Tools and Stolen Information: Hackers' Bazaar*. Tech Rep. RAND Corporation.

Aharoni, I. 2015. SCADA Systems Offered for Sale in the Underground Economy. Available online: http://www.infosecisland.com/blogview/24608-SCADA-Systems-Offered-for-Sale-in-the-Underground-Economy.html accessed on 30. June 2018.

Biddle, P., England, P., Peinado, M. and Willman, B. 2002. The Darknet and the Future of Content Protection. In *ACM Workshop on Digital Rights Management*, Springer-Verlag Berlin Heidelberg; pp. 155-176, ISBN 3-540-40410-4.

Biryukov, A., Pustogarov, I., Thill, F., and Weinmann, R. 2014. Content and Popularity Analysis of Tor Hidden Services. In *Distributed Computing Systems Workshops*, IEEE 34th International Conference; pp. 188-193.

Chaum, D. 1981. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. In *Communications of the ACM*, vol. 24 no. 2.

Cox, J. 2015. Scams and Undercover Cops Are Denting the Dark Web Gun Trade. Available online: https://motherboard.vice.com/en_us/article/wnx88q/scams-and-undercover-cops-are-denting-the-dark-web-gun-trade accessed on 30. June 2018.

Cox, J. 2016. Confirmed: Carnegie Mellon University Attacked Tor, Was Subpoenaed by Feds. *Motherboard*.

Cox, J. 2016. The FBI Used a ‘Non-Public’ Vulnerability to Hack Suspects on Tor. *Motherboard*.

DARPA. 2014. Memex Tools and Components. Available online: https://github.com/darpa-i2o/memex-program-index accessed on 9 March 2019.

Dingledine, R. 2002. *Pre-Alpha: Run an Onion Proxy Now!* SEUL Project Archives.

Dingledine, R. and Mathewson, N. and Syverson, P. 2004. *Tor: The Second-Generation Onion Router*. Naval Research Lab Washington DC.

Dingledine, R. 2014. Tor Security Advisory: “Relay Early” Traffic Confirmation Attack. Tor Blog.

Eddy, M. 2019. Hackers Leak Details of German Lawmakers, Except Those on Far Right. *The New York Times*.

Ferner, J. 2017. Betreiber eines TOR-Exit-Nodes kann für Urheberrechtsverletzungen haften. Available online: https://www.ferner-alsdorf.de/urheberrecht__betreiber-eines-tor-exit-nodes-kann-fuer-urheberrechtsverletzungen-haften__rechtsanwalt-alsdorf__56543 accessed on 09. March 2019.

Goulet, D. 2017. *Ongoing DDoS on The Network - Status*. The Tor Project.

Hug, N. 2015. *Follow the Data: Analysing Breaches by Industry. Trend Micro Analysis of Privacy Rights*. Clearinghouse.

Hargreaves, S. 2015. IWF Annual Report 2015. Internet Watch Foundation.

Hargreaves, S. 2016. IWF Annual Report 2016. Internet Watch Foundation.

Hargreaves, S. 2017. IWF Annual Report 2017. Internet Watch Foundation.

Janssen, D. and Janssen, C. 2018. *Deep Web*. Techopedia 2018.

Jeffries, A. 2017. People Keep Falling for this Murder-for-Hire Dark Web scam. Available online: https://theoutline.com/post/932/people-keep-falling-for-this-murder-for-hire-dark-web-scam accessed on 30. June 2018.

Kadianakis, G. and Loesing, K. 2015. *Extrapolating Network Totals from Hidden-Service Statistics*. The Tor Project.

Koch, R., Golling, M. and Dreo, G. 2016. How Anonymous is the Tor Network? A Long-Term Black-Box Investigation. *IEEE Computer* no. 3; pp. 42-49.

Lacey, D. and Salmon, Paul M. 2015. It’s Dark in There: Using Systems Analysis to Investigate Trust and Engagement in Dark Web Forums. In *Engineering Psychology and Cognitive Ergonomics*, Springer International Publishing, Cham, Switzerland; pp. 117-128, ISBN 978-3-319-20372-0.

Mathewson, N. 2014. Some Thoughts on Hidden Services. Tor Blog.

Mavroudis, V., Hao, S., Fratantonio, Y., Maggi, F., Kruegel, C. and Vigna, G. 2017. On the Privacy and Security of the Ultrasound Ecosystem. In *Proceedings on Privacy Enhancing Technologies*; De Gruyter Open; pp. 95-112.

McCandless, D. 2018. World’s Biggest Data Breaches. Available online: http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/ accessed on 30. June 2018.

McFarland, C., Paget, F. and Samani, R. 2015. *The Hidden Data Economy - The Marketplace for Stolen Digital Information*. McAfee LLC.

Moore, D. and Rid, T. 2016. Cryptopolitik and the Darknet. In *Survival* Vol. 58 Nr. 1 2016, Taylor & Francis; pp. 7-38.

Noubir, G. and Sanatinia, A. 2016. *Honey Onions: Exposing Snooping Tor Hsdir Relays*. DEFCON 24.

OWL Cybersecurity. 2017. The OWL Cybersecurity Darknet Index: Reranking the Fortune 500 using Darknet Intelligence (DARKINT). OWL Cybersecurity, Denver, Colorado.

OWL Cybersecurity. 2017. DARKOWL Press Room. Available online: https://www.darkowl.com/news/ accessed on 30. June 2018.

Paganini, P. 2017. Digging into the Darkweb. CTI - EU Cyber Threat Intelligence ENISA.

Persi Paoli, G., Aldridge, J., Ryan, N., and Warnes, R. 2017. *Behind the Curtain*. RAND Corporation.

Ray, V. 2017. Exploring the Cybercrime Underground: Part 4 - Darknet Markets. Available online: https://researchcenter.paloaltonetworks.com/tag/cybercrime-underground/ accessed on 30. June 2018.

Reed, M., Syverson, P. and Goldschlag, D. 1998. Anonymous Connections and Onion Routing. *IEEE Journal on Selected Areas in Communications* vol. 16 no. 4 1998; pp. 482-494.

Soska, K. and Christin, N. 2015. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. *USENIX Security Symposium*; pp. 33-48.

The Tor Project. 2013. 224-rend-spec-ng.

The Tor Project. 2015. Tor Project: FAQ. Available online: https://www.torproject.org/docs/faq.html.en\#ChangePaths accessed on 9 March 2019.

The Tor Project. 2017. Tor 0.3.2.2-alpha is Released.

The Tor Project. 2018. *Welcome to Tor Metrics!*

Zhang, X. 2003. *System/Application Designs, Optimization and Implementations on Overlay Networks*. High Performance Computing and Software Lab; Ohio State University.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

SamSam and the Silent Battle of Atlanta

Kenneth Kraszewski
LLD Candidate
Faculty of Law
University of Helsinki
Helsinki, Finland
kenneth.kraszewski@helsinki.fi

Abstract: The SamSam ransomware attack on Atlanta in early 2018 crippled municipal services in a major American city without the firing of a single shot, epitomizing the notion of a “Silent Battle”. Atlanta was not the only battlefield. Municipal governments in Colorado and New Mexico, as well as medical associations in Indiana, Virginia, New York and Buffalo, were all targets. While other ransomware or ransomware-like attacks have been larger-scale events, the SamSam ransomware attacks deserve an international law analysis.

This article examines the SamSam attacks on health care providers and municipal government through the lens of the second Tallinn Manual. First, it explains the SamSam ransomware itself and Gold Lowell, the group presumed to be behind it. Second, this article explores how the SamSam incidents might be classified under international law. This article asks whether ransomware attacks are internationally wrongful acts – breaches of international obligations attributable to a State. This entails considering whether a ransomware attack may be legally classified as a use of force, an intervention, a violation of sovereignty, or a breach of an international law obligation. Finally, this article discusses the possible legal responses to the SamSam ransomware attacks available to the United States: countermeasures, the plea of necessity, acts of self-defense under Article 51 of the U.N. Charter, and acts of retorsion.

Keywords: *attribution, cyber attack, due diligence, non-intervention, ransomware, sovereignty*

1. INTRODUCTION

In March 2018, the municipal government of Atlanta was "brought to its knees" by a ransomware attack deemed "one of the most sustained and consequential cyberattacks ever mounted against a major American city".[1] The city's court – "the busiest court" in the South-eastern United States[2] — was unable to validate warrants, policer officers were forced to issue citations by hand, and the city's employment application portal was shut down.[3] Years of digital files were rendered inaccessible.[4] The attack was costly. Its perpetrators demanded $51,000 to restore Atlanta's systems to full functionality, but the city followed the advice of federal authorities and refused payment. One month later, Atlanta had spent over $2.6 million to restore its systems;[5] an additional $9.5 million was later requested.[6] Atlanta is not alone in its misery. The same hacking group and malware have been implicated in attacks on hospital and health services providers and municipal governments across the United States.

In 2016, hospital systems in Baltimore were infected.[7] The following year, Buffalo's primary trauma center was hit. With computers offline, staff resorted to paper charts, transmitted messages in person, and viewed X-rays on traditional light boxes.[8] Clinics and doctors' offices in Virginia lost access to patient files when the systems of an electronic health records company were infected in early 2018.[9] A hospital in Greenfield, Indiana was infected simultaneously, leaving 1,400 files, including patient medical records, inaccessible.[10]

While Atlanta received more attention, other municipal governments were also victims. Two thousand computers at the Colorado Department of Transportation were encrypted in late February 2018. Colorado spent up to $1.5 million to remediate the

[1] Alan Binder & Nicole Perlroth, *A Cyberattack Hobbles Atlanta, and Security Experts Shudder*, N.Y. TIMES, Mar. 27, 2018, https://nyti.ms/2Gf7oRX.

[2] Rhonda Cook, *Court Hit by Hack Struggles to Recover*, ATLANTA J.-CONST., June 10, 2018, at B1, 2018 WLNR 17814216.

[3] Binder & Perlroth, *supra* note 1.

[4] Charles Bethea, *The Seemingly Random and Definitely Worrisome Cyberattack on Atlanta*, THE NEW YORKER, Mar. 29, 2018, https://perma.cc/E982-5NL3.

[5] Lily Hay Newman, *Atlanta Spent $2.6M to Recover From $52,000 Ransomware Scam*, WIRED, Apr. 23, 2018, https://perma.cc/3CBJ-PF2M.

[6] *Atlanta Officials Reveal Worsening Effects of Cyber Attack*, 6/6/18 Reuters News 22:50:01, June 6, 2018.

[7] Ian Duncan et al., *MedStar Hackers Demand Ransom*, BALT. SUN, Mar. 31, 2016, at 1, 2016 WLNR 9768566.

[8] Henry L. Davis, *How ECMC Got Hacked by Cyber Extortionists*, BUFF. NEWS, May 20, 2017, 2017 WLNR 15750503.

[9] Cathy Dyson, *Fredericksburg Clinic, Doctors' Offices Crippled by Virus—the Computerized Kind*, FREE LANCE-STAR (Fredericksburg, Va.), Jan. 22, 2018, 2018 WLNR 2228939.

[10] Vic Ryckaert, *Hospital Pays $50K Ransom for Patient Data*, INDIANAPOLIS STAR (Indianapolis, Ind.), Jan. 18, 2018, A01, 2018 WLNR 1767864.

effects.[11] SamSam ransomware shut down systems in Farmington, New Mexico, disrupting bill paying and record processing services.[12]

The WannaCry, Petya and NotPetya ransomware incidents of 2017 have garnered greater media coverage than SamSam. WannaCry infected hundreds of thousands of systems across the world, wreaking havoc on the United Kingdom's National Health Service, the Russia Interior Ministry, and India's Andhra Pradesh police department.[13] Petya, like WannaCry, made use of code stolen from the U.S. National Security Agency and leaked online.[14] It began as an attack on Ukrainian government and business computer systems on the day before a holiday marking the adoption of Ukraine's first post-Soviet constitution.[15] Petya spread to affect systems across the globe. Soon thereafter, a variant of Petya struck in Ukraine: deemed "NotPetya", this follow-on event was determined to not be a traditional ransomware attack. Instead, researchers have concluded that the attack, which targeted the computer systems of banks, energy firms and an airport, primarily in Ukraine, was carried out by Russian government hackers. The ransomware component was a ruse designed to trick its victims into believing the attacks were being conducted by a "mysterious hacker group".[16]

While WannaCry, Petya and NotPetya were larger scale events, the SamSam ransomware also deserves an international law analysis; because its effects manifested in a single State, the analysis is perhaps more straightforward. This article considers the attacks on health care providers and municipal government through the lens of the *Tallinn Manual 2.0 on International Law Applicable to Cyber Operations ("Tallinn Manual 2.0")*.[17] The SamSam ransomware and the group behind it are explained in Part 2. Part 3. explores how the SamSam incidents might be classified under international law, and Part 4. discusses the possible responses available to the United States.

This article purposely avoids considering the ransomware campaign under the auspices of the Convention on Cybercrime of the Council of Europe ("Budapest Convention")[18] in order to consider how such attacks may be analyzed through the

[11] Tamara Chuang, *After Online Derailment, CDOT Mostly on Track*, DENV. POST, Apr. 6, 2018, 14A, 2018 WLNR 10601275.

[12] Hannah Grover, *City of Farmington Recovering After SamSam Ransomware Attack*, DAILY TIMES (Farmington, N.M.), Jan. 18, 2018, 2018 WLNR 1861786.

[13] Michael Schmitt and Sean Fahey, *WannaCry and the International Law of Cyberspace*, JUST SECURITY, Dec. 22, 2017, https://perma.cc/QJ7W-GY7K.

[14] Nicole Perlroth et al., Cyberattack Hits Ukraine Then Spreads Internationally, N.Y. Times, June 27, 2017, https://www.nytimes.com/2017/06/27/technology/ransomware-hackers.html.

[15] *Id*.

[16] Ellen Nakashima, *Ukraine Attack Used a Ransomware Ruse*, WASH. POST, June 30, 2017, at A12, 2017 WLNR 20082512

[17] INT'L GRP. OF EXPERTS, NATO COOP. CYBER DEF. CTR. OF EXCELLENCE, TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (Michael N. Schmitt ed., 2017) [hereinafter TALLINN MANUAL 2.0].

[18] Council of Europe, Convention on Cybercrime, European Treaty Series, No. 185 (Budapest, opened for signature 23 Nov. 2001, entered into force 1 July 2004).

Tallinn Manual 2.0. While the Budapest Convention may, in certain circumstances, be a better vehicle for bringing the perpetrators of malicious cyber incidents to justice, it has significant drawbacks. It does not apply to State actors or the nationals of non-member States, and its scope differs significantly from that for the *Tallinn Manual 2.0*. The former focuses on harmonizing national laws to counter cybercrime, whereas the latter is principally concerned with whether and how international law applies to malicious activities in cyberspace. This article, in keeping with the approach of the *Tallinn Manual 2.0*, will consider whether the SamSam attacks may be characterized as internationally unlawful acts and the possible responses available to the United States, rather than considering whether they should be treated as cybercrimes under the Budapest Convention and the remedies available under that instrument.

2. GOLD LOWELL AND SAMSAM

The group behind the SamSam ransomware attacks has been named "Gold Lowell" by cybersecurity researchers.[19] Gold Lowell's members were first believed to reside in Eastern Europe,[20] but later alleged to be Iranians.[21] Security researchers presume that the group's members are not native English speakers based on "linguistic errors" in the ransom notes and transaction communications.[22] Gold Lowell is believed to have privately developed the SamSam ransomware.[23]

Unlike other forms of ransomware, SamSam is directly targeted. Attacks are focused on healthcare providers and municipal governments. SamSam is not commodity ransomware sold to other actors on online forums. The software is closely held and updated frequently to thwart antivirus detection.[24] Gold Lowell has utilized different means to gain access to servers. In 2015 and 2016, they scanned for Java vulnerabilities. Later, the group moved on to target Microsoft's IIS, file transfer protocol, and remote desktop protocol ("RDP"). As of May 2018, the group was primarily focused on accessing networks through "single-factor" external access protocols, such as RDP or virtual private networks.[25] Several tools are used once the group has gained access to the network, and Gold Lowell "is known to move from file to file, manually encrypting hundreds of systems".[26] Once encryption is complete, an apologetic message is displayed demanding payment of a certain sum in exchange for decryption.[27] The SamSam group purposely sets the price at a level

[19] Secureworks, *SamSam Ransomware Campaigns*, Feb. 15, 2018, https://perma.cc/L4EP-J2W6.

[20] Steve Ragan, *SamSam Explained*, CSO, Apr. 18, 2018, https://perma.cc/DP4W-YJUH

[21] Nicole Perlroth & Katie Benner, *Iranians Accused in Cyberattacks*, N.Y. TIMES, Nov. 28, 2018, https://www.nytimes.com/2018/11/28/us/politics/atlanta-cyberattack-iran.html.

[22] Secureworks, *supra* note 19.

[23] Ragan, *supr*a note 20.

[24] *Id*.

[25] *Id*.

[26] Nicole Perlroth, *Digital Thieves Rely on Ransom*, HOUS. CHRON., May 14, 2017, at A001, 2017 WLNR 15134229.

[27] Christopher Boyd, Malwarebytes, *SamSam Ransomware*, May 1, 2018, https://perma.cc/3LAT-VGGV.

deemed affordable. The rate charged to decrypt one system is set at around $10,000, while all systems on the network can be decrypted for $50,000. The group has even offered to decrypt one non-essential system for free to demonstrate their ability and willingness to release the data if their demands are met. The following sections of this article consider whether the SamSam ransomware attacks were internationally wrongful acts and how the United States might legally respond.

3. INTERNATIONALLY WRONGFUL ACTS

For a cyber operation to constitute an internationally wrongful act, it must be attributed to a State and must breach an international obligation owed by that State to another State.[28] Setting aside the question of attribution for the moment, this article first explores whether the SamSam attacks were breaches of an international law obligation. In the context of cyber operations, the most relevant obligations are the prohibition on the use of force, the prohibition on intervention, respect for the sovereignty of other States, and due diligence. Each obligation is examined in detail.

A. Breach of International Obligation

1) Use of Force

The SamSam ransomware attacks were not breaches of the prohibition on the use of force because the scale and effects of the attacks were neither sufficiently severe, immediate, direct, invasive, nor measurable to be considered uses of force. Nor were the SamSam ransomware attacks prohibited threats to use force because although the demands for ransom payments were communicative in nature, the action threatened in the messages was not itself an unlawful use of force.

The United Nations Charter ("U.N. Charter") prohibits the threat or use of force by one State against the territorial integrity or political independence of another.[29] The threshold for what constitutes the use of force in cyberspace is unsettled. However, the prohibition of the use of force is not limited to simply uses of kinetic force. There was general agreement amongst the International Group of Experts (the "Experts") involved in drafting the *Tallinn Manual 2.0* that cyber operations causing death, destruction, injury, or damage are uses of force. Nevertheless, the level of damage inflicted must not be more than de *minimis*.[30]

Whether a cyber activity crosses the use of force threshold can be determined by applying a scale and effects test. The test considers how widespread and of what nature the effects of the cyber activities are. Crucial to the determination is whether

[28] Int'l Law Comm'n, Responsibility of States for Internationally Wrongful Acts, G.A. Res. 56/83 annex, U.N. Doc. A/RES/56/83, art. 2 (December 12, 2001) [hereinafter Articles on State Responsibility].

[29] U.N. Charter art. 2(4).

[30] TALLINN MANUAL 2.0, *supra* note 17, at 334.

the effects of the cyber activities are comparable to those of a kinetic action or a non-kinetic action that qualifies as a use of force. If the activity's effects are comparable, then the cyber activity can also be considered a use of force. If not, the activity is unlikely to qualify as a use of force.

The *Tallinn Manual 2.0* proposes that States are likely to consider eight factors: severity, immediacy, directness, invasiveness, measurability of effects, military character, State involvement, and presumptive legality.[31] Severity is the most important factor. If the scope, duration and intensity of the effects of a cyber activity are severe, it will be likely be considered by States to be a use of force.[32] All the other seven factors are contextual. The more immediate, direct, invasive, measurable, presumptively legal, military in nature, and involving a State the effects are, the more likely it is that the activity will be judged a use of force. Immediacy concerns the time between the cyber activity and its effect.[33] Directiveness involves the nexus between the activity and its effect.[34] Invasiveness describes the activity's degree of penetration into the cyber system of the victim, with the caveat that highly invasive activities that merely exfiltrate data without causing damage will be considered internationally lawful acts of cyber espionage, not uses of force.[35] Measurability of effects gauges the quantifiability of the effects and is linked to the severity factor.[36] Military character is considered relevant because the U.N. Charter is especially concerned with military actions.[37] Presumptive legality is premised on the Lotus principle that international acts not expressly forbidden are permitted.[38] Thus, absent express treaty or accepted custom to the contrary, several prominent cyber activities are presumptively judged not to be uses of force: psychological operations, dissemination of propaganda, espionage, and economic coercion.[39] State involvement, finally, concerns the nexus between the State and the activity.[40] States are also likely to take into account a prevailing political environment, including the relationship between the victim State and the State to which the cyber activity is attributed, when judging whether a cyber activity is a use of force.

[31] *Id*. at 334–36.

[32] *Id*. at 334.

[33] *Id*.

[34] *Id*.

[35] *Id*. at 334–35. Most scholars agree that peacetime espionage is not the breach of an international obligation, but several has disagreed. *See*, *e.g*., Ingrid Delupis, *Foreign Warships and Immunity for Espionage*, 78 AM. J. INT'L L. 53, 67 (1984) (reasoning that peacetime espionage is illegal under international law if it involves an intrusion of foreign territory); Manuel R. Garcia-Mora, *Treason, Sedition and Espionage as Political Offenses Under the Law of Extradition*, 26 U. PITT. L. REV. 65, 79–80 (1964) (labeling peacetime espionage "an international delinquency and violation of international law"); Quincy Wright, *Legal Aspects of the U-2 Incident*, 54 AM. J. INT'L L. 836, 849 (1960) (stating that peacetime espionage is an "illegitimate enterprise[] because [it] manifest[s] a lack of respect for foreign territory").

[36] TALLINN MANUAL 2.0, *supra* note 17, at 335–36.

[37] *Id*. at 336.

[38] S.S. Lotus (Fr. v. Turk.), Judgment, 1927 P.C.I.J. (ser. A) No. 10, at 3, 18 (Sept. 7).

[39] TALLINN MANUAL 2.0, *supra* note 17, at 336.

[40] *Id*.

Even assuming that the SamSam incidents can be attributed to a State actor, it is unlikely that their scale and effects are such that they should be considered at least uses of force. Crucially, their overall severity was low. While the potential for serious harm to result from the disruption of normal hospital and municipal functions is high, in none of the incidents did such harm actually occur. The consequences of the SamSam attacks did not follow immediately from the cyber activities. In most cases, the penetration of the affected systems occurred weeks before the ransom notice was directed to the victim, and monetary costs incurred by the victims to recover data and restore their systems followed weeks or months thereafter. Nor were the effects of the SamSam attacks directly connected to the underlying cyber activity. While the attacks did have indirect consequences, in the form of the costs incurred to restore backed-up data and to implement improved security, the directness of the attacks' causes and effects is in no way comparable to the direct harm caused to people or objects by an explosion. Gold Lowell did indeed invasively probe the networks of municipal governments and healthcare providers; however, these were not top-secret networks that were necessarily intended to have the highest level of security. And the networks that the hackers did access were not amongst the most secure maintained by the victims: for instance, Atlanta's emergency response networks were untouched. The effects of the SamSam attacks cannot be calculated with certainty, even if a numerical sum can be affixed to the remediation costs. There is no suggestion that the attacks had a military character: no link has been publicly asserted between the hackers and the military of any State, nor were American military forces the target of the ransomware campaign. Likewise, no State is publicly alleged to have been involved, either directly or indirectly, in the campaign. Finally, the reconnaissance and network probing activities of the Gold Lowell group are qualitatively similar to espionage activities, which are not per se regulated under international law and are not presumptively judged to be uses of use. On consideration of each one of the foregoing factors, the SamSam attacks fail to meet the criteria of a use of force.

Finally, the U.N. Charter prohibits not only unlawful uses of force but also threats of the use of unlawful force.[41] The elements of a prohibited threat of the use of force include that the threat be communicated to the victim and that the threatened action be an unlawful use of force. The *Tallinn Manual 2.0* considers a cyber activity to be a prohibited threat of the use of force when "the threatened action, if carried out, would be an unlawful use of force".[42] The SamSam attacks do involve the communication of a threat that if a ransom is not paid, the victim's data will be lost. But, following the analysis of the previous paragraph, the threatened action is not a use of force. Moreover, by the time Gold Lowell communicated the ransom notice to its victims, it had already undertaken the action of encrypting their files, causing an effect. The group was simply offering the chance to mitigate the effects of its action for a price.

[41] U.N. Charter art. 2(4).

[42] TALLINN MANUAL 2.0, *supra* note 17, at 338.

The SamSam incidents were neither unlawful uses of force nor unlawful threats of the use of force.

2) Intervention

A cyber activity that falls below the threshold of a use of force may still be a breach of the customary international law principle of non-intervention. In the cyber context, the principle of non-intervention prohibits "coercive intervention, by cyber means, by one State into the internal or external affairs of another".[43] Thus, an intervention consists of two elements: a cyber activity relating to the internal affairs or external affairs of the target State, and the activity must be coercive.

A State's internal affairs or *domaine réservé* comprises those matters "in which [it] is permitted by the principle of sovereignty, to decide freely".[44] In particular, a State's *domaine réservé* includes the "choice of a political, economic, social, and cultural system, and the formulation of foreign policy".[45] According to the *Tallinn Manual 2.0*, the State's choice of political system and its organization lie most clearly within a State's *domaine réservé*.[46] Excluded from a State's *domaine réservé* are all matters that the State has committed to international law. For example, a State bound by human rights obligations that severely restricted the freedom of speech of its citizens could not argue that a cyber operation by another State enabling the first State's citizens to communicate more freely was an unlawful intervention in its *domaine réservé*. By entering into a human rights treaty, the first State had committed such matters to international law and removed them from its *domaine réservé*. In addition to *domaine réservé*, the principle of non-intervention also protects the external affairs of the target State. Thus, matters such as the State's choice of diplomatic and consular relations, recognition of foreign States and governments, membership of international organizations and participation in the drafting of or entry into treaties are all protected. A cyber operation coercively interfering in the *domaine réservé* or the external affairs of the target State is a breach of the principle of non-intervention.[47]

The second component in an unlawful intervention is that it be coercive.[48] While its coercive effect may be indirect, the act must be designed to deprive the target State of the freedom of choice in either its *domaine réservé* or external affairs. The intervening State's action must intentionally cause the target to either act in a way it would otherwise not act or refrain from acting in the manner that it otherwise would have. The mere threat of action can meet the threshold of intervention if it coerces the target State into acting or refraining from action.

[43] *Id*. at 312.

[44] *Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.)* [hereinafter Nicaragua], Judgment, 1986 I.C.J. 14 (June 27), para 205.

[45] *Id*., para. 205.

[46] TALLINN MANUAL 2.0, *supra* note 17, at 315.

[47] *Id*. at 317.

[48] *Nicaragua*, 1986 I.C.J. 14, para. 205 ("The element of coercion . . . defines, and indeed forms the very essence of prohibited intervention.").

The SamSam attacks were not coercive interventions in the *domaine réservé* or external affairs of the United States. There is no suggestion that the SamSam incidents in any way involved the external affairs of the United States, but certain SamSam attacks did implicate its *domaine réservé*. For example, the conduct of the Atlanta traffic police or the operation of the Colorado Department of Transportation are certainly fields of activity not committed to international law. It is less likely that the attacks were coercive efforts designed to influence outcomes in those fields of activity.[49] While Gold Lowell may have manipulated hospitals and municipal governments into making a choice between paying a ransom or spending considerably more to remedy the effects, that choice was not coercive in the sense that it was designed to compel the United States to adopt a particular policy with regard to traffic police, hospitals, or municipal policy. Instead, the coercion was intended to compel the payment of ransom.

3) Violation of Sovereignty

While neither violations of the use of force nor prohibited interventions, the SamSam ransomware incidents, if attributable to a State, were violations of U.S. sovereignty because they caused severe losses of functionality and interfered with the performance of inherently governmental functions. "Sovereignty in the relation between States signifies independence. Independence in regard to a portion of the globe is the right to exercise therein, to the exclusion of any other State, the functions of a State".[50]

A violation of sovereignty may take one of two forms: a violation of the territorial State's borders or an interference or usurpation of an inherently governmental function of the territorial State. The violating action must be undertaken by or attributable to another State.[51] In cyberspace, a violation of territorial integrity is difficult to identify, especially if the cyber activity is conducted remotely. The *Tallinn Manual 2.0* approach judges whether a violation of territorial integrity is a violation of sovereignty on the basis of "the degree of infringement upon the target State's territorial integrity".[52] Causing physical damage within the territorial State is a violation of sovereignty; causing a loss of functionality to the cyber infrastructure of the territorial State may sometimes be.[53] For instance, the 2012 Shamoon virus, which caused thousands of computers maintained by Saudi Arabia's state oil company to malfunction to the point of necessitating their repair or replacement, was a violation of Saudi Arabia's sovereignty, assuming it could be attributed to a State.[54] A cyber activity that necessitates reinstallation of the operating system would likewise be a

[49] TALLINN MANUAL 2.0, supra note 17, at 318 ("[M]ere coercion does not suffice to establish a breach of the prohibition of intervention [. . . . Instead,] the coercive effort must be designed to influence outcomes in, or conduct with respect to, a matter reserved to a target State.").

[50] Island of Palmas (Neth. v. U.S.), 2 R.I.A.A. 829 (Perm. Ct. Arb. 1928).

[51] TALLINN MANUAL 2.0, *supra* note 17, at 17.

[52] *Id*. at 20.

[53] *Id*.

[54] *Id*. at 21.

violation.[55] However, whether a cyber activity that causes neither physical damage nor a loss of functionality constitutes a breach of the territorial State's sovereignty is unclear.[56]

An interference with or usurpation of an inherently governmental function of the territorial State, regardless of whether damage is caused, also qualifies as a violation of sovereignty.[57] The territorial State enjoys the exclusive right to perform inherently government functions—e.g., delivering social services, conducting elections, collecting taxes, and conducting diplomacy. Inherently governmental function is a narrower concept than *domaine réservé*: whereas the latter concerns an area over which the State has exclusive control, the former deals with specific State functions. Stealing money from a State tax collector is not an interference with or usurpation of the State's inherently governmental tax collection function, whereas preventing the State from collecting taxes or usurping its authority to collect taxes is.

The SamSam ransomware attacks, if attributable to a State, are violations of the sovereignty of the United States. While the attacks did not cause physical damage, they resulted in severe losses of functionality. Medical services were disrupted. Municipal offices were forced offline for weeks. The loss of functionality required spending considerable sums of money to remedy. Moreover, the SamSam incidents also interfered with the performance of inherently governmental functions. Atlanta's court and police operations are inherently governmental functions, which although not usurped were certainly interfered with. Thus, the attacks were violations of the United States' sovereignty and, if attributable to a State, constitute internationally wrongful acts.

4) Due Diligence

The SamSam attacks may also have been breaches of the international obligation of due diligence if the State controlling the territory from which they were launched had a requisite level of knowledge about their occurrence and failed to take feasible actions to prevent them. A territorial State is in breach of its international due diligence obligation to a target State when it has actual or constructive knowledge of and fails to take feasible measures to stop an action affecting the rights of and causing serious adverse consequences to the target State emanating from within the territorial State's territory.[58] In the cyber context, a State must exercise due diligence in not allowing territory under its control to be used for cyber operations that affect the rights of and cause severe adverse consequences to another State.[59]

Breaches of the duty of due diligence do not require that the act in question be

[55] *Id.*
[56] *Id.*
[57] *Id.*
[58] *See Corfu Channel (UK v. Alb.)*, 1949 I.C.J. 4, 22 (Apr. 9).
[59] TALLINN MANUAL 2.0, *supra* note 17, at 30.

attributable to a State. Instead, the duty of due diligence assumes the role of three parties: the target State toward which the cyber operation is directed; the territorial State; and a third-party author of the cyber operation.[60] The third party may be another State, a non-State group, or a private person. Thus, if the State that controls the territory from which the Gold Lowell group is operating has knowledge of those operations, the operations affect the rights of and cause serious adverse consequences to the United States, and the United States intimates that the State take action to stop the breach of an international norm, that State has a duty to take feasible action to stop the SamSam actions. While the harm caused by a cyber activity must be serious, the due diligence principle does not require that there be physical damages to objects or injuries to persons.[61]

The SamSam ransomware incidents affected the U.S. sovereign right to perform inherently governmental functions – operating courts and police departments. It is questionable, however, whether there were serious adverse consequences. While the incidents certainly had the potential to cause serious adverse consequences – if, for example, the encryption of medical files had led to improper medical care resulting in injury to or death of patients – no such serious adverse consequences were reported.[62]

Knowledge, actual and constructive, is a constitutive element of the duty of due diligence. A State is in breach if even if it is unaware of cyber activity conducted from its territory but "objectively should have known that its territory was being used".[63] There is too little publicly available information to determine whether the State from whose territory the Gold Lowell group is operating actually knows or objectively should know about its operations or whether any actions have been taken to stop the SamSam ransomware attacks. Thus, the analysis need not go further.

B. Attribution

To constitute an internationally wrongful act, the SamSam ransomware attacks must not only be breaches of an international obligation owed by one State to another but must also be attributable to the former. Attribution is especially difficult in cyberspace.[64] A cyber operation is attributable to a State when it is carried out by organs of that State or by organs of another State placed at its disposal. A cyber operation can also be attributed to a State when it is carried out by non-State actors pursuant to the State's

[60] *Id*. at 32.

[61] *Id*. at 37–38.

[62] *See, e.g.*, Duncan, *supra* note 7 (quoting a Baltimore doctor as saying "while things have moved more slowly, patients were getting treated"); Ryckaert, *supra* note 10 ("Life support and other critical hospital services were not affected, and patient safety was never at risk.").

[63] TALLINN MANUAL 2.0, *supra* note 17, at 41.

[64] *See, e.g.*, William Banks, *State Responsibility and Attribution of Cyber Intrusions After Tallinn 2.0*, 95 TEX. L. REV. 1487, 1505–08 (2017); Christian Payne & Lorraine Finlay, *Addressing Obstacles to Cyber-Attribution*, 49 GEO. WASH. INT'L L. REV. 535, 559–566 (2017). *See also* Thomas Rid & Ben Buchanan, *Attributing Cyber Attacks*, 38 J. STRAT. STUD. 4, 7 (2015) (proposing a "Q model" for attribution, combining tactical, operational, and strategic aspects).

instructions or under its direction or control, or when the State acknowledges and adopts the operation as its own. From the publicly available evidence, it appears that the SamSam attacks cannot be attributed to a State actor because they were not the acts of a State organ, acknowledged and adopted by a State, or carried out by Gold Lowell pursuant to a State's instructions or under a State's direction or control.

1) Attribution of Acts by State Organs and State Organs Placed at the Disposal of Another State

The law of State responsibility defines "organs of a State" broadly to include any State organ, whether it exercises legislative, executive, judicial or any other functions, whatever its position in the organization of the State, and whatever its character as an organ of the central or regional government of the State.[65] An organ of a State also includes "any person or entity which has that status in accordance with the internal law of the State".[66] Thus, if the SamSam attacks were carried out by any governmental unit of a State or if the attackers were a State organ under the State's internal laws and the attacks are found to be breaches of an international obligation owed to the United States, each attack is an internationally unlawful act. However, there is no suggestion in any of the public reporting concerning the SamSam incidents that Gold Lowell is a State organ. No formal announcement has been made to that effect, which contrasts with charges made by the United States against North Korea in the aftermath of the WannaCry malware in 2017.[67] Without further information, it is speculative to presume that Gold Lowell is an organ of any State.

2) Attribution of Acts by Non-State Actors

Even if Gold Lowell is not a State organ, its actions may be attributable to a State if conducted pursuant to that State's instructions or under its direction or control or retroactively acknowledged and adopted.[68] No State has acknowledged and adopted the SamSam attacks. Thus, to attribute the campaign to a State, it must be shown that Gold Lowell was "acting on the instructions of, or under the direction or control of, [a] State".[69]

When a non-State actor is acting upon the instructions of a State, the analysis is simple. If the non-State actor functions as the State's "auxiliary", its actions are attributable to the State.[70] For instance, if a State hires a group of hackers to identify vulnerabilities in an adversary's cyber infrastructure, the group's actions are attributable to the State. Whether a non-State actor is under the "direction or control" of a State is less straightforward. Direction indicates a longer-term relationship between the State

[65] Articles on State Responsibility, *supra* note 28, art. 4(1).

[66] *Id*., art. 4(2).

[67] *See* Michael Schmitt & Sean Fahey, *WannaCry and the International Law of Cyberspace*, JUST SECURITY, Dec. 22, 2017, https://perma.cc/QJ7W-GY7K.

[68] *See* TALLINN MANUAL 2.0, *supra* note 17, at 94.

[69] Articles on State Responsibility, *supra* note 28, art. 8.

[70] TALLINN MANUAL 2.0, *supra* note 17, at 95.

and the non-State actor, and control indicates that the State exercises a high degree of control over the non-State actor's actions. Together, direction and control can be likened to the notion of "effective control" devised by the International Court of Justice in *Nicaragua* and reiterated in *Genocide*.[71] In the cyber context, a State having "effective control" over a non-State actor would determine the execution and course of the cyber operation carried out by the non-State actor and would have authority to order its commencement and cessation.[72] Simply participating in the planning and supervision of non-State actor's cyber operation is not exercising "effective control". Nor is the mere provision of financial or other support.[73]

The SamSam attacks are not attributable to another State because Gold Lowell, according to public sources of information, was not acting under the instruction or "effective control" of another State. Without State attribution, it is impossible to establish that the SamSam incidents constitute an internationally wrong act on the basis of a breach of the prohibition on the use of force, an unlawful intervention, or a violation of U.S. sovereignty. Although it was judged that the SamSam incidents neither constituted a use of force nor a prohibited intervention, they were violations of sovereignty. However, because the actions of the Gold Lowell group cannot be attributed to a State, those violations alone do not constitute internationally unlawful acts. The principle of due diligence does not require that the underlying wrongful action be attributable to a State. Thus, if the State controlling the territory from which the attacks were launched had a requisite level of knowledge and failed to take feasible actions to prevent them, it breached of its duty of due diligence.

4. POSSIBLE RESPONSES

Having established that the SamSam attacks, according to public information, do not meet the criteria of an internationally unlawful act, this section examines the options available for the United States to take in response. Cyber operations may, in general, be met with four responses under international law: countermeasures, the plea of necessity, self-defense, and retorsion. For the reasons explained below, only retorsion is suitable.

A. Countermeasures

Countermeasures are actions are would be unlawful but for the fact that they are taken in response to another State's internationally wrongful act and are designed to terminate that unlawful act or compel the State to which it is attributable to make reparations.[74]

[71] *Nicaragua*, 1986 I.C.J. 14, para. 115; *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosn. and Herz. v. Serb. and Montenegro)*, Judgment, 2007 I.C.J. Rep. 108 (Feb. 26), para. 400.

[72] TALLINN MANUAL 2.0, *supra* note 17, at 96.

[73] *Nicaragua*, 1986 I.C.J. 14, para. 115.

[74] Articles on State Responsibility, *supra* note 28, art. 49.

However, the object of countermeasures must be a State,[75] and it is not possible to attribute the SamSam attacks to a State. Moreover, there must be an internationally wrongful act to justify countermeasures.[76] Even if there was, countermeasures should be limited to ensuring that the unlawful act stops, potentially obtaining assurance and guarantees of non-repetition from the responsible State,[77] and compelling the responsible State to make reparations.[78] Because the SamSam incidents have stopped, countermeasures would have to be limited to compelling the responsible State to guarantee that the incidents not resume and providing compensation for damages. Countermeasures may not be punitive or have a retaliatory effect.[79]

Additionally, the United States would be advised not to engage in countermeasures in response to the SamSam attacks even were they attributable to a State because if the countermeasures were to violate a legal obligation owed to a third State, the United States would itself be in breach of international law. The wrongfulness of such a breach is not precluded by the validity of the countermeasure against the responsible State.[80] Thus, the United States could find itself in breach of its international law obligations by too aggressively seeking to curtail Gold Lowell's campaign.

B. Plea of Necessity

The plea of necessity allows a State to act in exceptional cases when there is grave and imminent peril to an essential interest of the State and action is the sole means of safeguarding that interest.[81] Even then, the plea of necessity requires that the injured State's action be balanced with the interests of any States that would be affected and with those of the international community.[82] The injured State's action may not seriously impair the essential interests of affected States.[83] The plea of necessity is not available to injured State that have substantially contributed to their own injury.[84] However, the plea of necessity can be asserted to take action against non-State actors and can justify actions that violate the rights of non-responsible States, if the threat to an essential interest of the injured State is sufficiently grave and imminent and no other means of safeguarding the interest are present. State attribution is not a precondition for action based on the plea of necessity.

A State's "essential interest" is not clearly defined. It would certainly include healthcare, justice, and policing. Thus, the SamSam attacks on healthcare service providers and

[75] TALLINN MANUAL 2.0, *supra* note 17, at 112.

[76] *Id*. at 114.

[77] *Id*. at 142–44 (discussing the responsible State's duty to cease an internationally wrongful act and, if appropriate, provide assurances and guarantees of non-repetition).

[78] *Id*. at 144–52 (discussing the responsible State's obligation to make full reparation for injuries suffered by the injured State).

[79] Michael N. Schmitt, *"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law*, 54 VA. J. INT'L L. 697, 714 (2014).

[80] TALLINN MANUAL 2.0, *supra* note 17, at 133.

[81] Articles on State Responsibility, *supra* note 28, art. 25(1)(a).

[82] *Id*., art. 25(1)(b).

[83] *Id*.

[84] *Id*., art. 25(2)(b).

Atlanta's police and court systems certainly impaired essential interests of the U.S. It is unlikely that the temporary interruption in functionality the ransomware caused was sufficient to put those essential interests in grave and imminent peril and that no other means existed to safeguard those interests. In any case, the ransomware attacks have abated, if temporarily, and the plea of necessity could only be invoked to end the harmful activity.

C. Self-defense

A State may respond with force to a cyber operation that qualifies an "armed attack" pursuant to the customary international law right of self-defense, codified in Article 51 of the U.N. Charter. Most commentators consider only grave uses of force – typically, those that kill or injure persons or damage or destroy property—to be armed attacks.[85] The U.S., however, takes an outlier position, consistently arguing that any use of force is an armed attack.[86] In *Nicaragua*, the I.C.J. identified "scale and effects" as criteria upon which to judge whether a use of force constitutes an armed attack. In the Court's view, only "the most grave" uses of force do so.[87] Thus, only cyber operations that kill persons or cause significant damage to, or destruction of, property would constitute armed attacks.[88] Because the SamSam ransomware campaign fails to meet the criteria of use of force, even accepting the United States' outlier opinion, it was not an armed attack triggering the right to self-defense.

D. Retorsion

Retorsion, "lawful retaliation in kind for another country's unfriendly or unfair action",[89] is the best legal response available to the United States in dealing with the SamSam attacks. Acts of retorsion are lawful, albeit unfriendly.[90] For example, a State may respond to another State's unfriendly or unfair action by suspending diplomatic relations with the responsible State, restricting travel rights or expelling foreign nationals of the responsible State, or preventing the use of its cyber infrastructure for communications from the responsible State.[91] Retorsion is only way for the United States to respond to the SamSam ransomware campaign without a determination that another State has breached an international obligation owed to it.

5. CONCLUSION

The SamSam ransomware campaign disrupted healthcare organizations and municipal services in numerous locations across the United States. Undoubtedly, the attacks

[85] *Nicaragua*, 1986 I.C.J. 14, para. 95.

[86] US Department of Defense, Office of the General Counsel, Law of War Manual (June 2015), paras. 1.11.5.2, 16.3.3.1.

[87] *Nicaragua*, 1986 I.C.J. 14, para. 191.

[88] TALLINN MANUAL 2.0, *supra* note 17, at 341.

[89] Black's Law Dictionary (10th ed. 2014).

[90] TALLINN MANUAL 2.0, *supra* note 17, at 112.

[91] *Id*.

were malicious cyber operations carried out by foreign actors, implicating the rights of the United States under international law. To be considered internationally unlawful acts, the ransomware attacks would have to constitute the breach of an international law obligation owed to the United States and be attributed to a State. The attacks were neither uses of force nor coercive interventions in the *domaine réservé* of the United States. While violations of U.S. sovereignty, the attacks are not attributable to a State according to publicly available reporting. Likewise, it is unknown whether the United States has asked any State to fulfil its due diligence obligation to use all feasible measurable to end the attacks. Thus, the SamSam attacks do not qualify as internationally unlawful acts, limiting the possible recourse for the United States. Even if the ransomware attacks could be attributed to a State, countermeasures would be ill-advised because they would be limited to forcing a State to comply with its legal obligation. Because the attacks are not presently ongoing, the United States would risk engaging in punitive or retaliatory action, for which countermeasure are not allowed. The plea of necessity likewise cannot be invoked to respond to action that has stopped. Because the ransomware was not a use of force, the United States cannot invoke its customary law and Article 51 right of self-defense. Thus, retorsion is the best response available to the United States.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

The Contours of 'Defend Forward' Under International Law

Jeff Kosseff
Assistant Professor
Cyber Science Department
United States Naval Academy[1]
Annapolis, MD
kosseff@usna.edu

Abstract: In 2018, United States Cyber Command announced a new operational concept to "defend forward" against other states whose cyber operations against the United States have been hostile, but short of an armed attack. Defend Forward supports the U.S. strategy of persistent engagement, which recognizes the need to continuously engage to inhibit incessant adversarial cyber operations against the United States. Although the public Defend Forward description was short on details, it consists of three general components: (1) *positioning* to degrade cyber operations; (2) *warning* to gather information about threats and inform defenses; and (3) *influencing* adversaries to discourage them from deploying cyber operations against the United States. In the year since the announcement of the Defend Forward concept, there has been vital debate about whether the United States *should* defend forward. This paper examines a related but distinct question: *Could* the United States defend forward under international law, and if so, what limits does the law impose? This paper concludes that international law provides the United States with significant leeway to position itself to degrade adversaries' cyber operations, gather information about cyber threats, and discourage other states from acting against the United States in cyberspace. Although international law imposes vital limits on operational concepts such as Defend Forward, there is a significant gap between those boundaries and how the United States has defended against cyber aggression short of armed conflict.

Keywords: *cybersecurity, countermeasures, defense, espionage, retorsion*

[1] The views expressed in this paper are only those of the author, and do not represent the United States Naval Academy, United States Department of Navy, United States Department of Defense, or any other party. Thanks to those who provided incredibly valuable feedback on earlier drafts, including Tracy Emmersen, Evan Field, Michael Fischerkeller, Emily Goldman, Joe Hatfield, Chris Inglis, Eric Jensen, Martin Libicki, Kurt Sanger, and Andrew Slack.

1. INTRODUCTION

The headline in the September 20, 2018 edition of *The Washington Post* was unambiguous: "White House Authorizes 'Offensive Cyber Operations' to Deter Foreign Adversaries."[2] Reporting on U.S. National Security Adviser John Bolton's discussion of a new U.S. cyber posture authorized by the classified National Security Presidential Memorandum 13, the *Post* declared it "a new policy that eases the rules on the use of digital weapons to protect the nation."[3] Yet in the same article, the Post reported that Bolton, speaking at a news conference announcing the federal government's new cyber strategy, "did not elaborate on the nature of the offensive operations, how significant they are, or what specific malign behavior they are intended to counter."[4]

Such is the challenge of describing a nation's cyber strategy. As the United States and its allies face constantly evolving cyber threats from Russia, China, North Korea, Iran, and non-state actors, the recently elevated U.S. Cyber Command has taken an increasingly active stance in cyberspace, with a "defend forward" operating concept that supports its strategy of "persistent engagement." This stance reflects the reality that continuous engagement with cyber adversaries, rather than case-by-case responses, are necessary in light of the constant threats that the United States faces.[5] While the public statements of Cyber Command indicate that the United States military will increasingly move beyond operating within its cyber perimeter, the inherently classified nature of cyber operations makes it difficult to know, with certainty, what precisely the government means when it promises to "defend forward."

This paper fills some of these gaps by defining the outer limits that international law imposes on the U.S. ability to defend forward. Although the United States has exercised considerable restraint in cyber operations to date, this has largely stemmed from operational concerns, such as the impact on international relations.[6] To be sure, international law imposes significant constraints on even some mild forms of cyber offense; however, the United States has been operating far below those legal limits. The paper first outlines the limited public statements that the United States has issued regarding Defend Forward. Based on those high-level statements, the paper then assesses the scope of permissible actions under international law. In short, the paper argues that international law provides the United States with significant leeway to use countermeasures, espionage, and retorsion to "defend forward" and conduct cyber operations in the systems and networks of others.

[2] Ellen Nakashima, *White House Authorizes 'Offensive Cyber Operations' to Deter Foreign Adversaries*, WASH. POST (Sept. 20, 2018)

[3] *Id.*

[4] *Id.*

[5] *See* Dave Weinstein, *The Pentagon's New Cyber Strategy: Defend Forward*, LAWFARE (Sept. 21, 2018).

[6] *See* Ben Buchanan, *The Implications of Defending Forward in the New Pentagon Cyber Strategy*, COUNCIL ON FOREIGN RELATIONS (Sept. 25, 2018) ("the Obama administration in particular exhibited a tremendous caution in the world of offensive cyber operations").

2. DEFINING ‘DEFEND FORWARD’

To understand the significance of the U.S. adoption of the operational concept of “defend forward” and its accompanying strategy of “persistent engagement,” it is useful to examine the development of U.S. cyber policy over nearly a decade. In July 2011, the Defense Department issued its Strategy for Operating in Cyberspace. Among the most noteworthy parts of the strategy was “active cyber defense,” which the Department stated was intended “to prevent intrusions and defeat adversary activities on DoD networks and systems.”[7] The 2011 Strategy suggested that this defense would take place within the Defense Department’s network.[8] In April 2015, the Defense Department issued a new Cyber Strategy, which focused on protecting not only Defense Department networks but also civilian government and private sector networks.[9] The strategy stated that the U.S. Defense Department could be directed to “use cyber operations to disrupt an adversary’s command and control networks, military-related critical infrastructure, and weapons capabilities” during “heightened tensions or outright hostilities”[10] but did not explicitly brand such operations as “offensive.”[11]

The formal articulation of a “defend forward” operational concept occurred in 2018. In March, Cyber Command released a 10-page Command Vision: “Defending forward as close as possible to the origin of adversary activity extends our reach to expose adversaries’ weaknesses, learn their intentions and capabilities, and counter attacks close to their origins.”[12] The Command Vision stresses the need for “continuous engagement”, which “imposes tactical friction and strategic costs on our adversaries, compelling them to shift resources to defense and reduce attacks.”[13] Although the Command Vision provides little detail as to what sorts of “friction” and “costs” the United States might impose, the focus on stopping cyber threats *before* they hit the United States was soon hailed as a marked shift in U.S. cyber strategy.[14]

The National Security Presidential Memorandum 13, signed in August 2018,

[7] DEPARTMENT OF DEFENSE STRATEGY FOR OPERATING IN CYBERSPACE (July 2011) at 7.

[8] *Id*. (“As intrusions may not always be stopped at the network boundary, DoD will continue to operate and improve upon its advanced sensors to detect, discover, map, and mitigate malicious activity on DoD networks.”).

[9] *See* DEPARTMENT OF DEFENSE CYBER STRATEGY (April 2015) at 10 (“In addition to DoD’s own networks, a cyberattack on the critical infrastructure and key resources on which DoD relies for its operations could impact the U.S. military’s ability to operate in a contingency.”).

[10] *Id*. at 14.

[11] *See* Herb Lin, *Two Observations About the New DOD Cyber Strategy*, LAWFARE (April 24, 2015) (“[O]ne must *infer* the offensive character of the operations being discussed at various points in the document.”).

[12] U.S. CYBER COMMAND, ACHIEVE AND MAINTAIN SUPERIORITY IN CYBERSPACE: COMMAND VISION FOR U.S. CYBER COMMAND (March 2018) at 6.

[13] *Id*.

[14] *See* Richard Harknett, *United States Cyber Command’s New Vision: What It Entails and Why It Matters*, LAWFARE (March 23, 2018) (“These operational orientations recognize that previous U.S. approaches ultimately left the U.S. playing ‘clean-up on aisle nine,’ too often dealing with adversaries inside our networks (or in the aftermath of their exploitations), rather than stopping them before entering.”)

reportedly supported a more flexible approach. The memorandum is classified, and the Defense Department released an unclassified summary of its cyber strategy the next month. The summary states that Defend Forward was intended to "disrupt or halt malicious cyber activity at its source, including activity that falls below the level of armed conflict."[15] The unclassified summary discusses the plan to "defend forward to halt or degrade cyberspace operations targeting the Department[.]"[16] Observers quickly recognized the significance of the new operational concept.[17] Defend Forward is the clearest indication of the U.S. recognition that cyber threats do not merely take the form of discrete events but are also continuous operations that must be defended against in real time. Gen. Paul M. Nakasone, commander of U.S. Cyber Command, elaborated on the purpose of "defend forward" and "persistent engagement" in a 2019 article, further confirming the intent to operate beyond U.S. military networks: "Persistent engagement of our adversaries in cyberspace cannot be successful if our actions are limited to DOD networks," he wrote. "To defend critical military and national interests, our forces must operate against our enemies on their virtual territory as well."[18]

A more detailed description of Defend Forward appeared in an unclassified 2018 Cyber Command newsletter that received little public attention. Cyber Command wrote that Defend Forward is part of its Persistent Engagement strategy, which "focuses on an aggressor's confidence and capabilities by defending against, countering, and contesting on-going strategic campaigns short of armed attack."[19] Cyber Command identified three "broad lines of effort" that comprise defending forward:

- **Positioning:** Perhaps the biggest shift in U.S. cyber operations under Defend Forward is Cyber Command's recognition of the need for "a forward cyber posture that can be leveraged to persistently degrade the effectiveness of adversary capabilities and blunt their actions and operations before they reach U.S. networks."[20] The positioning focuses on America's "most capable and dangerous adversaries in cyberspace, thereby allowing diplomatic, law enforcement, security, and private actors to address lesser threats against which they have the authorities and capacity to defend" and "may also support a strategy of deterrence and warfighting."[21]

[15] Summary, Department of Defense Cyber Strategy 2018.

[16] *Id*. at 2.

[17] *See* Nina Kollars & Jacquelyn Schneider, *Defending Forward: The 2018 Cyber Strategy is Here*, WAR ON THE ROCKS (Sept. 20, 2018) ("Reactive strategy might focus on hack-backs, while a preemptive strategy might focus on operations that prevent an adversary's cyber unit from accessing the Internet."); Lyu Jinghua, *A Chinese Perspective on the Pentagon's Cyber Strategy: From 'Active Cyber Defense' to 'Defending Forward,'* LAWFARE (Oct. 19, 2018) ("The evolution in Defense Department cyber documents suggests that the U.S. cyber force is expanding its scope of operations in terms of geography, timing and potential adversaries.").

[18] Paul M. Nakasone, *A Cyber Force for Persistent Operations*, 92:1 JOINT FORCE QUARTERLY (2019) at 10.

[19] U.S. Cyber Command, CYB3R CYPH3RS, Vol. 4., No. 1, at 5.

[20] *Id*.

[21] *Id*.

- **Warning:** The Defend Forward concept gives the United States "enhanced warning of adversary actions, intentions, and capabilities," and allows the United States "to better defend government and civilian networks, data, and platforms."[22] Obtaining information about the adversaries' cyber operations before they are deployed "allows cyber mission forces to assess the threat, develop mitigations, and disseminate threat information across allies, partners, and industry."[23]
- **Influence:** The Defend Forward concept also "encourages stability by disabusing adversaries of the idea that they can operate with impunity in cyberspace" and "signals U.S. commitment to confront hostile activities and impose cumulative costs for ongoing malicious actions."[24] Cyber Command discusses an approach of "shadowing" dangerous cyber actors to "keep them constantly on-guard and off-balance" and "signal their national leaders that attribution and response to cyber aggression will be swift."[25]

3. LEGAL CONTOURS OF 'DEFEND FORWARD'

This section examines the limits and obligations that international law imposes on the three components of Defend Forward: positioning, warning, and influence. Positioning is likely to raise the most concerns under international law, and therefore will be discussed most extensively. Even under a conservative application of international law, however, the United States will have significant leeway to implement the newer defend forward concept.

A. Positioning

A noteworthy aspect of "Defend Forward" is the focus on "positioning" activities. Cyber Command's public definition of positioning is not terribly specific, likely stemming from an understandable aversion to describing particular techniques. The public description suggests that these operations might require the United States to access non-DOD networks or systems in order to adequately position itself.

Positioning might be akin to the kinetic concept of "preparing the battlefield." As Robert Chesney wrote, the cyber equivalent of battlefield preparation might include "[i]ntrusions into the systems of potential adversaries in order to secure access of a kind that can be exploited for disruptive or destructive effect if and when the need later arises."[26] Positioning supports the strategy of persistent engagement by inhibiting the

[22] *Id*.

[23] *Id*.

[24] *Id*.

[25] *Id*.

[26] Robert Chesney, *The 2018 DOD Cyber Strategy: Understanding 'Defense Forward' in Light of the NDAA and PPD-20 Changes*, LAWFARE (Sept. 25, 2018). To the extent that the access is conducted for the purpose of deterrence, Chesney distinguishes it as a "hold at risk" operation rather than battlefield preparation. *Id*.

adversary's planning and execution of cyber campaigns targeting U.S. interests. Such active measures are the category of the Defend Forward approach that is most likely to raise international law concerns. However, when they are aimed at nations that are continuously acting against the United States in cyberspace, there is significant leeway for the United States to respond. Under Defend Forward, such response might take place on non-U.S. military networks.[27]

Cyber Command's limited public description states that Defend Forward addresses activities that fall below armed conflict.[28] This reflects the realities of the steadfast aggression that the United States confronts in cyberspace.[29] Accordingly, this paper examines how the United States should address continuous campaigns of hostile actions that are not sufficiently grave to constitute armed attacks; therefore, U.S. positioning in this situation cannot rise to the level of the use of force. It is difficult to predict with absolute certainty whether a cyber operation to establish the capability to degrade an adversary's capabilities would be seen as a use of force.[30] However, there is a strong argument that narrowly focused Defend Forward operations would not constitute a use of force.[31] An operation may be less likely to constitute a use of force if its effects have a limited "scope, duration, and intensity."[32] For instance, the analytical framework set forth in the *Tallinn Manual 2.0* suggests that if the United States determines that a particular IP address is the repeated source of malware that is harming U.S. computers, an action would be less likely to qualify as a use of force if it was focused on positioning the ability to degrade operations from that individual IP address for a limited period of time rather than positioning across a much broader region.[33] Similarly, ensuring that the operation does not cause physical damage, bodily harm, and, most importantly, casualties, will substantially reduce the likelihood of it being viewed as a use of force.[34] It is unlikely that mere positioning activities, separate from leveraging that position, would rise to that level.

[27] *Id*. (stating that defend forward "plainly concerns activity outside of U.S. networks" and that it "entails operations that are intended to have a disruptive or even destructive effect on an external network: either the adversary's own system or, more likely, a midpoint system in a third country that the adversary has employed or is planning to employ for a hostile action.").

[28] *See* Department of Defense *supra* note 15 at 2; *see also* Weinstein, *supra* note 5 ("This is an important principle: the United States simply cannot allow the current levels of sub-armed conflict in cyberspace to persist unmitigated.").

[29] *See* Gary Corn & Eric Talbot Jensen, *The Use of Force and Cyber Countermeasures*, 32 TEMPLE INT'L & COMP. L. J. 127 (2018) ("Happily, this situation of threatened armed attack is not the norm in today's world, whether through cyber or non-cyber operations. However, the continuous and pervasive use of cyber capabilities to conduct unfriendly and even internationally wrongful acts presents a potentially destabilizing influence on the international community.").

[30] *See* Michael N. Schmitt, *"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law*. 54 VA. J. INT'L L. 697, 719 (2014) ("[U]ncertainty will sometimes exist as to whether a cyber operation taken in response to an internationally wrongful act reached the use of force threshold and thereby failed to qualify as a countermeasure.").

[31] *See* Michael N. Schmitt (ed.), TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (2017) (hereinafter, "Tallinn Manual") at 333 (setting forth a multifactor test to determine whether a cyber operation constitutes a "use of force").

[32] Tallinn Manual at 334.

[33] *Id*. ("Severity is the most significant factor in the analysis.").

[34] *See* Andrew C. Foltz, *Stuxnet, Schmitt Analysis, and the Cyber 'Use of Force Debate*, JFQ (2012) ("cyber operations resulting in physical damage or injury will almost always be regarded as use of force.").

Assuming that the operation does not constitute a use of force, U.S. positioning operations still might infringe on the sovereignty of the target nation or violate another legal obligation.[35] The literature is not settled as to whether merely establishing a position to degrade ongoing adversarial cyber actions – rather than the degradation itself – constitutes a violation of sovereignty.[36] The United States would have a strong argument that mere positioning against persistent adversarial campaigns does not raise sovereignty issues, though this will likely depend on which network or system is the focus of a positioning operation, how the operation is deployed, and the impacts of the positioning.

Based on Cyber Command's public description of positioning, it appears that positioning helps to establish a posture that the U.S. could leverage to degrade adversaries' capabilities. Accordingly, any legal analysis of Defend Forward must examine *both* the positioning *and* the use of that position to degrade an adversary, even though degradation is not explicitly among the three stated prongs of Defend Forward. Once the United States *leverages* its position to degrade the adversary's operations, that act might be more likely to raise sovereignty issues.

To the extent that the operations do raise concerns about sovereignty,[37] these activities could be legally justified as countermeasures[38] if conducted to inhibit a persistent campaign of illegal acts against the United States, provided that they are not uses of force.[39] (There is no indication in Cyber Command's publicly disclosed strategy that positioning activities or use of the position would rise to the levels of use of force or armed attack.) The non-binding draft Articles on Responsibility of States for Internationally Wrongful Acts allow an injured state to exercise countermeasures to cause a state to cease the commission of internationally wrongful acts or to provide reparation.[40] Therefore, even if U.S. positioning activities violated sovereignty or other legal obligations to another nation, the United States could justify them as countermeasures aimed at ceasing further illegal actions against the United States.

[35] *See* Tallinn Manual at 17 ("A State must not conduct cyber operations that violate the sovereignty of another State.").

[36] *See Id*. at 21 ("no consensus could be achieved as to whether, and if so, when, a cyber operation that results in neither physical damage nor the loss of functionality amounts to a violation of sovereignty.").

[37] *See* Schmitt, *supra* note 30 at 705 ("While monitoring activities in another State may merely constitute espionage, which is not prohibited, emplacement of malware into a system, destruction of data, and hacking into a network to identify vulnerabilities would seem to pierce the veil of sovereignty.").

[38] *See* Tallinn Manual at 111 (defining "countermeasures" as "actions or omissions by an injured State directed against a responsible State that would violate an obligation owed by the former to the latter but for qualification as a countermeasure.").

[39] *See* Oona A. Hathaway, *The Drawbacks and Dangers of Active Defense*, PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT (2014) ("There is little legal support for the proposition that countermeasures doctrine provides a legal end-run around the prohibition on the use of force in Article 2(4) of the UN Charter.").

[40] DRAFT ARTICLES ON RESPONSIBILITY OF STATES FOR INTERNATIONALLY WRONGFUL ACTS (2001) (hereinafter "Articles on Responsibility") at 75 ("In certain circumstances, the commission by one State of an internationally wrongful act may justify another State injured by that act in taking non-forcible countermeasures in order to procure its cessation and to achieve reparation for the injury.").

The unrelenting nature of cyber threats increases the likelihood of success of a countermeasures justification.

If positioning or the use of that position to degrade are justified as countermeasures, they are constrained by the legal rule that countermeasures are limited to the purpose of terminating the other party's illegal activities.[41] For instance, the analytical framework in the *Tallinn Manual 2.0* suggests that if an adversary conducts cyber operations against the United States that damage U.S. data, systems, or connectivity, but fall short of an armed attack, such activities may nonetheless violate U.S. sovereignty and justify countermeasures.[42] Similarly, the draft Articles on Responsibility suggest that the United States may only degrade an adversary's capabilities temporarily until the adversary has resumed compliance with legal obligations.[43] Of course, in light of the continuous nature of cyber threats that prompted the persistent engagement strategy, the United States would have a reasonable argument that positioning and degradation are necessary over the long term as the adversaries' persistent aggression is unlikely to cease.

Who is a legitimate target of positioning actions? The United States may only direct countermeasures at a state that has violated international legal obligations to the United States.[44] Relatedly, the United States may only respond to the operations of a *state* that has violated an international legal obligation. If, for instance a private company in another nation has violated U.S. sovereignty, the United States is entitled to deploy countermeasures only if the company's actions are attributed to the state,[45] such as when the state "instructs or directs or controls cyber operations launched by a non-state group or by individuals."[46] To be sure, attribution is not an easy task, and requires substantial review of intelligence for sufficient evidence of the source of the attack. The U.S. Director of National Intelligence has stated that the primary indicators for "timely, accurate attribution" are: tradecraft, infrastructure, malware, intent, and external sources (such as the media and industry).[47]

The United States may only engage in operations that qualify as countermeasures in response to an adversary's breach of international legal obligations owed to the United

[41] *Id*. at 130 ("Countermeasures are not intended as a form of punishment for wrongful conduct, but as an instrument for achieving compliance with the obligations of the responsible State[.]").

[42] *See* Tallinn Manual at 113 ("Since the responsible State has itself engaged in an internationally wrongful act, the cyber countermeasure is lawful; as a matter of law, the State is the object of the countermeasure, which is designed to put an end to that State's wrongful activity.").

[43] *See* Articles on Responsibility at 130 (discussing "the temporary or provisional character of countermeasures.").

[44] *See* Eric Jensen & Sean Watts, *A Cyber Duty of Due Diligence: Gentle Civilizer or Crude Destablizer*, 95 TEX. L. REV. 1555, 1564 (2017).

[45] *See Id.*; Tallinn Manual at 113.

[46] Michael N. Schmitt, *Peacetime Cyber Responses and Wartime Cyber Operations Under International Law: An Analytical Vade Mecum*, 8 HARV. NAT'L SEC. J. 239, 255 (2017) (internal quotation marks and citations omitted).

[47] OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE, A GUIDE TO CYBER ATTRIBUTION (Sept. 14, 2018) at 2-3.

States.[48] Such a breach would occur if another state usurped "inherently governmental functions," such as by initiating cyber operations that prevent a government from collecting taxes or conducting elections.[49] Moreover, the international legal principle of non-intervention[50] prohibits a state from intervening, through coercion, in another state's "internal or external affairs," including the "choice of a political, economic, social, and cultural system, and the formulation of foreign policy."[51] The United States has a strong argument that Russia's unrelenting attempts to interfere in U.S. elections violates both principles,[52] though experts are divided as to the strength of these arguments as applied to individual components of the Russian efforts.[53] In short, even if a nation's actions against the United States fall far short of the armed attack threshold, they may well entitle the United States to exercise countermeasures to prevent future interference, particularly in light of the tenacious nature of the threats that target the very essence of U.S. democracy.[54]

To the extent that the United States determines that another country has violated an international legal obligation, what countermeasures is it entitled to exercise? U.S. countermeasures that leverage the country's positioning must be proportionate, which, according to the Articles on Responsibility, means that they "must be commensurate with the injury suffered, taking into account the gravity of the internationally wrongful act and the rights in question."[55] When determining whether a cyber countermeasure is proportionate, the United States should consider "the injury suffered (i.e., the extent of harm), the gravity of the wrongful act (i.e., the significance of the primary rule breached), the rights of the injured and responsible State (and interests of other States) that are affected, and the need to effectively cause the responsible State to comply with its obligations."[56] For example, if the United States detects that a country has made a few feeble attempts to infiltrate the election registration databases in a single U.S. town, it very well may be entitled to engage in countermeasures to prevent irreparable harm to the electoral system. However, in light of the relatively toothless nature of the aggressor's attempts to harm the U.S. electoral system, it likely would

[48] Tallinn Manual at 111.

[49] *Id*. at 21-22.

[50] *Id*. at 312 ("A State may not intervene, including by cyber means, in the internal or external affairs of another State.").

[51] Nicaragua v. United States, 1986 I.C.J. 14 (1986) at para. 205; see also Tallinn Manual at 315 ("Thus, this Rule prohibits coercive cyber acts by a State that are intended to eliminate or limit another State's prerogative on these matters.").

[52] *See* Steven J. Barela, *Zero Shades of Grey: Russian-Ops Violate International Law*, JUST SECURITY (March 29, 2018) ("A greater appreciation of the expansive costs, planning and aims of Russia's intervention helps bolster my judgment of coercion by exposing the massive 'scale' and 'reach' of the operation.").

[53] *See* Jens David Ohlin, *Did Russian Cyber Interference in the 2016 Election Violate International Law?*, 95 TEX. L. REV. 1579, 1587 (2017) ("the technical requirements for an illegal intervention might not apply to the Russian intervention, depending on how one understands the concept of coercion.").

[54] *See* Eric Jensen, *Countering Russian Election Hacks*, JUST SECURITY (Nov. 5, 2018) ("These self-help responses to Russian intervention could include cyber measures that would otherwise be unlawful but are designed to bring Russia back into compliance with international law.").

[55] Articles on Responsibility at 134.

[56] Tallinn Manual at 128.

be disproportionate for the United States to engage in a countermeasure that causes widespread Internet outages in the adverse country.

To be sure, the United States still would have significant breathing room to implement countermeasures. If a country continuously attempts to violate U.S. sovereignty, the United States would have a strong argument that it is entitled to take proportionate countermeasures to establish a position to be able to degrade the adversaries' ability to cause further harm. Even under the proportionality restriction, the United States would have substantial leeway to exercise and leverage positioning operations. The injury suffered – the threat to the legitimacy of the U.S. democratic system – and the gravity of the harms to democracy would justify efforts to prevent the adversary from carrying out future systematic campaigns. If, for instance, the United States identified a state that was routinely testing election registration databases, the United States arguably could take targeted actions to halt the aggressor's cyber capabilities without violating the countermeasures proportionality rule. The proportionality rule does *not* mean that the United States must respond by interfering with the aggressor's electoral system;[57] in fact, the more appropriate and effective response under the law of countermeasures would target the operators, systems, and networks that have been attacking U.S. voting systems.

B. Warning

Defend Forward calls for the United States to gather information about adversaries' cyber capabilities and planning. "Warning" involves operations that seek to better understand the cybersecurity threats that the United States faces. The United States may gather information about particular capabilities, allowing it to better structure U.S. defenses. The United States may also monitor adversaries in real time to understand when and how the United States may face significant threats. These warning operations hinge upon the United States' ability to access the communications networks of another country, raising concerns about espionage[58] or sovereignty.

To be sure, some operations within the "warning" function of Defend Forward are not necessarily espionage, such as making better use of open-source information about threats, or receiving threat information from allies. The use of public information for warning of cyber threats does not raise concerns under international law.[59]

[57] *See* Schmitt, *supra* note 30 at 726. ("Proportionality does not imply reciprocity; there is no requirement that the injured State's countermeasures breach the same obligation violated by the responsible State. Nor is there any requirement that the countermeasures be of the same nature as the underlying internationally wrongful act that justifies them.").

[58] *See* Darien Pun, *Rethinking Espionage in the Modern Era*, 18 CHI. J. INT'L L. 353, 357 (defining "espionage" as "the unauthorized intentional collection of information by states.").

[59] *See* Russell Buchan, *The International Legal Regulation of State-Sponsored Cyber Espionage*, in INTERNATIONAL CYBER NORMS: LEGAL, POLICY & INDUSTRY PERSPECTIVES, Anna-Maria Osula and Henry Rõigas (Eds.) (2016) at 85 ("one must distinguish between intelligence-gathering from publically available sources and intelligence-gathering from private, unauthorised sources, namely espionage.").

To the extent that U.S. operations constitute espionage, international legal concerns may arise, but perhaps not to the same extent as positioning. There is no prohibition on espionage *per se*.[60] This is consistent with the U.S. Defense Department's view that "unauthorized intrusions into computer networks solely to acquire information" will be treated as "traditional intelligence and counter-intelligence activities under international law."[61] Some operations for gathering information from known cyber adversaries, such as the use of honeypots to trace the source of attacks, are commonly accepted as espionage that conforms to international law.[62]

Although there is no prohibition of cyberespionage *per se*, the United States may encounter some outer-bound restrictions on particular operations. Imagine, for instance, that the United States exploits a vulnerability on the Russian government's systems to learn about its plans to interfere in the 2020 U.S. elections, and in doing so, accidentally deletes large quantities of important data from the Russians' systems. The majority view in *Tallinn Manual 2.0* suggests that if this damage is sufficient to constitute a violation of the United States's international legal obligations, the United States could not avoid responsibility merely because the damage was connected to an espionage operation.[63] Accordingly, a Defend Forward operation carried out for the purpose of gathering information must be performed with great care to ensure that the operation does not cause significant harm to data, networks, or systems.

The "warning" function, as described by U.S. Cyber Command, involves leveraging information that is useful to prepare the United States to better defend against cyber threats posed by other states.[64] The United States might still attempt to ensure that these warning operations do not involve the mass surveillance of the public and government officials that has drawn criticism from some as crossing the boundaries of international law.[65]

To the extent that a warning action crosses the line from legal espionage to a cyber operation that violates a legal obligation such as sovereignty or non-intervention, the

[60] *See* Christopher Yoo, *Cyber Espionage or Cyber War?: International Law, Domestic Law, and Self-Protective Measures*, in CYBERWAR: LAW AND ETHICS FOR VIRTUAL CONFLICTS 175-194 (Oxford University Press 2015) ("In the absence of any clear principles, with the exception of a handful of exceptions such as interference with diplomatic communiques, espionage remains the province of domestic law and falls outside the province of jus ad bellum and jus in bello."); Tallinn Manual at 169.

[61] OFFICE OF GENERAL COUNSEL, DEPARTMENT OF DEFENSE, DEPARTMENT OF DEFENSE LAW OF WAR MANUAL (June 2015, updated December 2016) at 1016.

[62] *See* Tallinn Manual at 173.

[63] *See Id*. at 170-72 ("The majority of the Experts agreed that although acts of cyber espionage may not be unlawful standing alone, they can nevertheless constitute an integral and indispensable component of an operation that violates international law."). Note that the minority view contends that "two aspects of the operation must be assessed separately." *Id*.

[64] *See* U.S. Cyber Command, *supra* note 19 at 5.

[65] *See* Daniel Trotta, *At U.N., Brazil's Rousseff blasts U.S. spying as breach of law*, REUTERS (Sept. 24, 2013) ("Brazilian President Dilma Rousseff used her position as the opening speaker at the U.N. General Assembly to accuse the United States of violating human rights and international law through espionage that included spying on her email.").

United States might still justify the act as a countermeasure. As described above in Section 3.A, provided that another state has violated an international legal obligation to the United States, the United States may engage in proportionate countermeasures aimed at ceasing the unlawful behavior. Accordingly, even if the United States conducts its information-gathering in a manner that moves beyond legally acceptable espionage, it may still justify the operation as a countermeasure provided that the legal prerequisites are met.

C. Influence

The "Influence" prong of Defend Forward includes actions that the United States employs in an attempt to discourage other states from acting against it in cyberspace. However, "Influence" could also include more active methods to dissuade adversaries. Some influence operations do not raise concerns under international law. For instance, the United States could resort to sanctions against a state in response to an unlawful cyber action, as it did against North Korea after the Sony hack.[66] Likewise, in 2016 the United States closed Russian compounds in the United States and expelled diplomats in response to the election interference.[67] Such actions could deter future hostile cyber actions against the United States through cost imposition.[68] Although such measures could raise political and diplomatic difficulties, they are not problematic under international law, as they constitute retorsion, which is "'unfriendly' conduct which is not inconsistent with any international obligation of the State engaging in it even though it may be a response to an internationally wrongful act."[69]

Retorsion would continue to be a key part of Defend Forward influence operations. For instance, drawing on historical examples of U.S.-Soviet relations, Seth G. Jones concluded that one key component of the U.S. response to Russia's election interference requires "blunt and regular U.S. warnings to Russian leaders, both in public and private, that their information warfare campaign will be met with an equally forceful response."[70] The United States has a good deal of flexibility in developing responses that qualify as retorsion, as they are not subject to the same legal constraints as countermeasures.

The United States also might attempt to specifically influence particular cyber operators

[66] *See* Issie Lapowsky, *What We Know About the New U.S. Sanctions Against North Korea In Response to Sony Hack*, WIRED (Jan. 2, 2015).

[67] *See* Mark Mazetti and Michael S. Schmidt, Two Russian Compounds, Caught Up in History's Echoes, N.Y. TIMES (Dec. 29, 2016).

[68] *See* Eric Lorber & Jacquelyn Schneider, *Sanctioning to Deter: Implications for Cyberspace, Russia, and Beyond*, WAR ON THE ROCKS (April 14, 2015).

[69] Articles on Responsibility at 128; *see also* Schmitt, *supra* note 46 at 258 ("The expulsion of diplomats and imposition of economic sanctions following allegations of Russian government hacking intended to interfere with U.S. elections qualified as retorsion."); Troy Anderson, *Fitting a Virtual Peg into a Round Hole: Why Existing International Law Fails to Govern Cyber Reprisals*, 34 ARIZ. J. INT'L & COMP. L. 135 (2016) (listing examples of retorsion).

[70] Seth G. Jones, *Going on the Offensive: A U.S. Strategy to Combat Russian Information Warfare*, CENTER FOR STRATEGIC AND INTERNATIONAL STUDIES BRIEFS (Oct. 1, 2018).

who have targeted the United States. For instance, an October 2018 article in the *New York Times* reported that U.S. Cyber Command had identified and directly messaged Russians who were involved in election propaganda operations.[71] The United States reportedly informed the Russians "that American operatives have identified them and are tracking their work, according to officials briefed on the operation," according to the *Times* report, and U.S. defense officials anonymously told the newspaper that the communications did not involve threats.[72] Although the communications are more tailored to specific operators rather than issuing a government-wide notification to Russia, it is unlikely that sending a notification to Russian cyber operators who are conducting information warfare on the United States violates Russia's sovereignty. Moreover, even if such communications infringed Russia's sovereignty or another legal obligation, the limited scope and severity fall well within the range of acceptable countermeasures aimed at terminating attempts to interfere in U.S. democracy.

4. CONCLUSION

Experts have engaged in important and significant debate about whether Defend Forward is a strategically wise choice for the United States.[73] While the normative debate about what the United States *should* do in cyberspace is vital, this paper has focused on what the United States *could* do within existing legal limits to inhibit continuous cyber campaigns against the United States that fall below the threshold of armed attacks. In sum, international law provides the United States with significant flexibility to "defend forward". To be sure, Defend Forward is subject to several legal limits, particularly when it comes to positioning and degradation; but even within these limits, the United States can conduct cyber operations that are far more active than the U.S. active defense concept of years past.

[71] Julian E. Barnes, U.S. Begins First Cyberoperation Against Russia Aimed at Protecting Elections, N.Y. TIMES (Oct. 23, 2018).

[72] *Id*.

[73] See, e.g., Josephine Wolff, *Trump's Reckless Cybersecurity Strategy*, N.Y. TIMES (Oct. 2, 2018).

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

The Rise of the Regionals: How Regional Organisations Contribute to International Cyber Stability Negotiations at the United Nations Level

Nikolas Ott[*]
Transnational Threats Department,
Co-ordination Cell
Organization for Security and
Co-operation in Europe (OSCE)
Vienna, Austria
nikolas.ott@osce.org

Anna-Maria Osula, PhD[**]
Guardtime / TalTech / Masaryk
University
Tallinn, Estonia
annamaria.osula@guardtime.com

Abstract: While States did not reach consensus on the 2017 report by the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (UN GGE), the UN remains a core platform for diplomatic deliberations on international law, norms and principles for responsible state behaviour.

At the same time, regional organisations play an increasingly important role in stabilising State relations in cyberspace. Their relevance is also recognised in the new UN GGE mandate for 2019-2021. For the first time, the UN GGE negotiations include a formal way of embracing regional cyber expertise, knowledge and concerns, albeit they are ambivalent about how the envisaged input will be incorporated into the UN GGE process.

The paper argues that regional organisations should and are willing to increase their substantial input to the global debates on international cyber stability. Specifically, we analyse the benefits of the work of the Organization for Security and Co-operation in Europe (OSCE), the Organization of American States (OAS) and the Association of

[*] The opinions expressed in this article are those of the authors alone and are not representing the official policy of any organisation or other entity.

[**] Dr Anna-Maria Osula's contribution to this paper is based on research supported by Masaryk University project no. CZ.02.1.01/0.0/0.0/16_019/0000822 (C4E).

Southeast Asian Nations (ASEAN), undertaken in the context of Confidence-Building Measures (CBMs). In addition to global platforms, we see great potential in inter-regional collaboration.

Moreover, the paper points out a number of suggestions which would enhance the inclusion of regional organisations' efforts into UN GGE; and potentially, also into the Open-Ended Working Group (OEWG) negotiations. More effective norm development and CBM implementation can be achieved by carefully assessing the pros and cons of various venues and formats as well as taking advantage of existing synergies between UN initiatives and regional CBM and capacity-building initiatives. Regional organisations have better insights into national or regional priorities; while domestic implementation frameworks may be developed by regional organisations for faster CBM and norm implementation procedures, and possibly allow for additional funding for priority areas. Regional roadmaps should be developed for more effective norm and CBM development, while joint implementation efforts could foster the global uptake of norms. Furthermore, regional organisations may serve as incubators for new ideas and share valuable experience of lessons learned.

Keywords: *UN GGE, OSCE, OAS, ASEAN, regional organisations, cyber security, cyber security strategies, capacity-building, confidence-building measures, cyber norms*

1. INTRODUCTION[1]

The failure to reach consensus on the 2017 report by United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (UN GGE) reflects the widening gap between States' visions on how to achieve a secure and stable cyberspace. Simultaneously, recent incidents highlight how States are further developing and increasingly deploying destructive cyber capabilities. Combined with the ongoing dispute over norms, rules, and principles for responsible State behaviour, there is an increasing risk of unintended military escalation. Therefore, a new perspective toward stabilising cyberspace is necessary.

It may be argued that due to the cross-border nature of cyber threats, regional solutions become less relevant. However, this article posits quite the opposite: namely, that regional governmental organisations[2] play a crucial role in tackling

[1] The authors are grateful to the anonymous reviewers for their comments. Special thanks also goes to Christoph Berlich, Ingmar Snabile, Henry Rõigas, Jessica Zucker and Kerry-Ann Barrett for their comments and feedback throughout the drafting process.

[2] This paper focuses exclusively on regional (inter-)governmental organisations and therefore uses the term 'regional organisations' as a shorter substitute thereof.

concerns related to cyber security. Their active input on the international level has the potential to contribute to 1) more effective and targeted norm development by taking advantage of existing synergies between the UN and regional organisations; 2) faster implementation procedures on the regional and national level through targeted and customised support; 3) more coherent inter-regional co-ordination of agreed stability efforts through inter-regionally co-ordinated, but regionally implemented roadmaps and frameworks; and 4) capacity-building and awareness raising. Thus, further incorporation of regional voices in reaching a global agreement on the content, interpretation and implementation of the norms of State behaviour, confidence-building measures (CBMs) and capacity-building is essential.

However, their presence at global venues has so far been limited. This is mostly due to regional organisations having a specific mandate tailored toward activities within their respective regions. This limits the extent to which they may engage in other international fora and partly explains why regional organisations are rarely present at the international negotiating table.[3] In fact, the UN GGE 2019-2021 is the first UN entity venue which now includes a formal way of embracing regional cyber expertise, knowledge and concerns.[4] This development should be applauded and will hopefully mark a trend of further inclusion of regional organisations and their Member States' concerns and suggestions. However, besides mentioning the additional consultations with regional organisations in the resolution, it remains unclear how the envisaged regional organisations' input will be incorporated into the UN GGE negotiations. Furthermore, there is no indication on whether this consultation process will lead to a regular substantive exchange between the global and regional levels. There are also doubts regarding overcoming the different views which stalled progress on the previous UN GGE consensus report.

Against this background, the paper investigates mechanisms for further involving regional organisations in cyber security policy deliberations within the UN. The paper analyses selected regional organisations' activities and documents related to norm-building and CBMs. In order to narrow our scope, we focus on selected regional organisations' prominent role in agreeing upon and implementing CBMs and discuss how these initiatives could better support ongoing work on norm-building.

Our paper is structured as follows. After a brief introduction to the current UN GGE process and status quo, it analyses CBM-related developments undertaken at regional venues such as the Organization for Security and Co-operation in Europe (OSCE), the Organization of American States (OAS) and the Association of Southeast Asian Nations (ASEAN). We also reference some capacity-building efforts and norms

[3] The European Union is an exception among regional organisations given its unique competencies and governance model. Therefore, the authors have decided to exclude the EU as a case study from this article.

[4] United Nations, Advancing responsible State behaviour in cyberspace in the context of international security, Resolution adopted by the General Assembly on 22 December 2018, A/RES/73/266, p 4, available at: https://undocs.org/A/RES/73/266.

discussions when they relate to regional CBM activities in the respective regions. We then outline opportunities through regional organisations' efforts on CBMs as well as the increasing role and inter-connectedness of regional organisations. After that, cross-cutting benefits of inter-regional collaboration are discussed. Finally, we conclude by proposing practical options for further including representatives of regional organisations into global processes.

2. UN GGE STATUS QUO

The UN GGE is the most reputable platform for agreeing international norms for States in cyberspace. Since 1998, when the Russian Federation first introduced a draft resolution on information security in the First Committee of the UN General Assembly,[5] the UN Secretary-General has issued annual reports with the views of UN Member States to the General Assembly.[6] Additionally, UN GGEs have been formed in 2004/5, 2009/10, 2012/13, 2014/15, and 2016/17, with a total of three consensus reports (in 2010, 2013 and 2015) examining the existing and potential threats from cyberspace, and possible co-operative measures to address them.[7]

In the latest development, in November 2018, the UN First Committee (Disarmament and International Security) approved two separate proposals to create working groups which would develop rules for responsible State behaviour in cyberspace. These were later adopted by the UN General Assembly. The first initiative, proposed by the Russian Federation, was to form an open-ended working group (OEWG) in 2019, "acting on a consensus basis to further develop the rules, norms and principles of responsible behaviour of States".[8] The second suggestion, tabled by the United States (US), was to continue the previous UN GGE efforts in order to study "possible cooperative measures to address existing and potential threats in the sphere of information security, including norms, rules and principles of responsible behaviour of States".[9]

The tension between these two proposals is evident. On the one hand, the US claimed that the Russian proposal "imposes a list of unacceptable norms and language that

[5] United Nations, Resolutions adopted by the General Assembly, 4 January 1999, A/RES/53/70.

[6] United Nations Office for Disarmament Affairs, Developments in the field of information and telecommunications in the context of international security, available at: https://www.un.org/disarmament/topics/informationsecurity/.

[7] The UN GGE convened in 2009 reached no consensus report. However, reports were published in 2010 (A/65/201), 2013 (A/68/98*) and 2015 (A/70/174). The UN GGE convened in 2016 did not reach a consensus report. UNODA fact sheet, available at: https://unoda-web.s3-accelerate.amazonaws.com/wp-content/uploads/2015/07/Information-Security-Fact-Sheet-July2015.pdf.

[8] United Nations, Developments in the field of information and telecommunications in the context of international security, Resolution adopted by the General Assembly on 5 December 2018, A/RES/73/27, available at: https://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/73/27.

[9] Footnote 4.

is broadly unacceptable to many States”,[10] with other commentators adding that the text “departed from previous year’s versions and included excerpts from the Group of Governmental Experts reports in a manner that distorted their meaning and transformed the draft resolution”.[11] On the other hand, the Russian Federation argued that the working group proposed by the US would take the “international community backwards and result in a complete waste of resources, also being the product of extremely narrow interests of Western countries, especially the United States”.[12]

One of the focal issues in this debate and a point of critique towards previous UN GGE processes is the selection of participating States. The number of countries involved in the UN GGE process has, over time, risen from 15 to 25, which reflects general aspirations of including a wider range of States, and eventually a hope for bigger buy-in to the agreed principles. At the same time, more members may also mean lengthier discussions and increased difficulties in reaching a consensus.

Both previously mentioned initiatives proposed to the UN General Assembly in 2018 touch upon including further stakeholders. The US proposal specifically requested the UN GGE meetings to be preceded by two two-day, open-ended, informal consultative meetings, so that all Member States could share their views, which the UN GGE Chair would then convey to the group of governmental experts for consideration.[13] In the same vein, the proposed OEWG has promised to take the negotiating process to a “higher level that is more inclusive, open and democratic”[14]; and has also asserted the possibility of holding inter-sessional consultative meetings with representatives of business, non-governmental organisations and academia, to share views on the issues within the group’s mandate.[15]

3. UN GGE AND REGIONAL ORGANISATIONS

It is against this background that our article will look into the role of regional organisations in shaping the international norms for States in cyberspace. We will examine the UN GGE reports published in 2010, 2013 and 2015, to analyse how the role of regional organisations has developed.

Regional organisations and initiatives have always been an integral part of the reports. All three reports recognise the valuable work undertaken by regional entities;[16] we can

[10] United Nations, First Committee approves 27 texts, including two proposing new groups to develop rules for States on responsible cyberspace conduct. Meeting coverage, GA/DIS/3619, 8 November 2018, available at: https://www.un.org/press/en/2018/gadis3619.doc.htm.

[11] Id.

[12] Id.

[13] Footnote 4, p 5.

[14] Footnote 10.

[15] Footnote 8, p 5.

[16] e.g. UN A/65/201 (2010) p 13; UN A/68/98* (2013) p 4, 14; UN A/70/174 (2015) p 35.

observe an increasingly substantial role being foreseen for the regional organisations.

This can be best seen in the UN GGE report of 2015, which finely outlined the areas where different actors should provide input in achieving international peace and security in cyberspace. The report established a detailed four-pillar system for guaranteeing cyber stability between States, made up of: a) the applicability of international law; b) norms, rules and principles for the responsible behaviour of States; c) CBMs; and d) capacity-building enhancing international co-operation.[17]

For example, similar to the conclusions adopted in 2013, the 2015 report recognised the importance of regional organisations in developing and implementing CBMs such as exchanging views and information, providing more transparency, enhancing common understandings and intensifying cooperation.[18] Equally relevant were regional efforts in capacity-building, such as securing ICT use and ICT infrastructures, strengthening national legal frameworks, law enforcement capabilities and strategies; combatting the use of ICTs for criminal and terrorist purposes, and assisting in the identification and dissemination of best practices.[19]

The 2015 report noted separately that the "development of regional approaches to capacity-building would be beneficial, as they could take into account specific cultural, geographic, political, economic or social aspects and allow a tailored approach".[20] Also, both the 2013 and 2015 reports clearly point out that the UN should encourage regional efforts,[21] and recommend regular dialogue through regional forums.[22] In 2015, the report puts specific focus on increased co-operation at regional and multilateral levels to "foster common understandings on the potential risks to international peace and security".[23]

The most significant development in engaging regional efforts within the UN GGE process was put forward through the US proposal for a new UN GGE in 2018. The Office for Disarmament Affairs of the Secretariat was invited to collaborate on behalf of UN GGE members and through existing resources and voluntary contributions, with relevant regional organisations, such as the African Union (AU), the European Union (EU), the OAS, the OSCE and the ASEAN, via a series of consultations: with the aim of sharing views on the issues within the group's mandate in advance of its sessions.[24]

[17] For more information on the general purpose and conceptual underpinnings of CBMs as well as linkages between the four pillars, see Patrick Pawlak, "Confidence-Building Measures in Cyberspace: Current Debates and Trends", in *International Cyber Norms: Legal, Policy & Industry Perspectives*, Anna-Maria Osula and Henry Rõigas (Eds.), NATO CCD COE Publications, Tallinn 2016.

[18] e.g. UN A/68/98* (2013) 26a, 26b, 29; UN A/70/174 (2015) 16b-16d, 17, 18.

[19] UN A/68/98* (2013) p 32a.

[20] UN A/68/98* (2013) p 22.

[21] e.g. UN A/68/98* (2013) p 13; UN A/70/174 (2015) p 35.

[22] UN A/68/98* (2013) p 29; UN A/70/174 (2015) p 18.

[23] UN A/70/174 (2015) p 30b.

[24] Footnote 4, p 4.

This can be interpreted as an acknowledgment by States that the UN GGE process needs to be more inclusive and can benefit from stronger engagement of regional expertise. At the same time, the envisaged procedures are proof of the readiness of regional organisations to play a greater role in enhancing confidence between States as well as global norm- and national capacity-building. Indeed, as will be illustrated in the remainder of this article, there is a clear interest of regional organisations in contributing to enhanced trust and confidence among States, as well as reaching an understanding on acceptable and unacceptable State behaviour in cyberspace.

4. OPPORTUNITIES THROUGH REGIONAL ORGANISATIONS' EFFORTS ON CONFIDENCE-BUILDING MEASURES

The inter-connectedness of the four-pillar approach presented in the UN GGE 2015 report has provided the groundwork for increased involvement of regional organisations. These four pillars as a whole can be understood as cyber stability mechanisms which are only effective if they reinforce each other. For example, norms of responsible State behaviour require to be put into practice to ensure buy-in. CBMs serve exactly this purpose by translating broader legal concepts into more concrete, straightforward actions. As the following chapter will extensively outline, regional organisations are also uniquely equipped to develop and implement CBMs which are not directly linked to norms, rules and principles for responsible State behaviour; but instead are more pragmatic and practical by design, thereby developing the foundational groundwork for enhanced communication, transparency and collaboration. Moreover, CBMs only serve their purpose to the fullest extent if they are implemented, which requires the capacity to do so. The following paragraphs will outline how CBMs are connected to and reinforce the other pillars; and why this is important in securing the success of global agreements.

4.1. THE MUTUALLY REINFORCING ROLE OF CBMS IN GLOBAL NORM-BUILDING

While developing norms, rules and principles for the responsible behaviour of States is vital, States need to have confidence that others will adhere to the same rules. This might sound trivial, but it requires a high level of co-operation among States. Given their more practical and concrete design, CBMs serve as pragmatic mechanisms in crisis situations. They can therefore be employed as measures to address norms or rules violations. CBMs are thus critical components of any cyber stability mechanism. Nevertheless, it is important to note that even the most advanced set of CBMs will not stop an intentional conflict; but they can stop an unintentional one by stopping or slowing down the spiral of escalation.

While norms and responsible State behaviour are discussed on the global level, CBMs tend to be developed on a regional or national level. This difference makes a lot of sense when reviewing the purpose of norms of responsible State behaviour and CBMs respectively. Ideally, norms of responsible State behaviour should not be subject to extensive interpretation, while CBMs leave more room for adjustment and allow for the inclusion of already existing regional or national procedures. This therefore allows for greater customisation and adjustment for regional needs. Regional organisations such as the OSCE, OAS and ASEAN Regional Forum (ARF) have engaged in this path and developed or are developing their respective sets of cyber/ICT security CBMs. In comparison with the EU, these three regional organisations bring together States that sometimes have difficult relations.[25] This is an important characteristic, as cyber stability needs to be built between non-like-minded States, not just geopolitical allies. Furthermore, in the context of the UN GGE process, if certain proposals are already supported or even initiated by regional organisations, there would automatically be a bigger buy-in during the UN GGE process in finding a consensus.

In addition to proposing and agreeing to norms, regional organisations benefit from their accumulated political capital in implementing practical measures. This aligns perfectly with the purpose of CBMs and helps drive their operationalisation forward. Third, regional organisations can consult, learn from and bridge different cultural and political approaches to cyber/ICT security. These three characteristics provide an excellent platform for regional organisations to address global cyber security challenges through explicitly regional means.

Additionally, there is a shared interest among nations in keeping the diplomatic process on cyber stability measures alive. Having multiple platforms across regions will help to test, for example, how States may practically implement norms. However, even though the cyber CBMs of the 21st century may share the same name as arms control CBMs of the Cold War era, their purpose and design is quite different;[26] 21st century CBMs are about "building areas of common understandings and practical cooperation among nations, including preparations for crisis management".[27] Large-scale cyber security incidents tend to spread fast, are normally trans-national; and most of the time, difficult to predict or anticipate. If States are to deal with such features, established practice, trust in each other and confidence that others will come to their support is needed.

[25] OAS can be considered as the most 'like-minded' group among the three of them.

[26] James A. Lewis, Confidence Building Measures in Cyberspace, Presentation to the Inter-American Committee Against Terrorism (CICTE) of the Organization of American States, Center for Strategic and International Studies, February 26, 2016, p 1, available at: https://www.oas.org/en/sms/cicte/Documents/2016/Speeches/JAMES%20LEWIS%20CSIS.pdf.

[27] Id.

For this very reason, if one considers norms as means to establish and enhance trust and confidence amongst nations, it seems obvious that a discussion on norms needs to be complemented with practical considerations that foster an environment of collaboration and support amongst nations. This can be achieved by implementing agreed norms and CBMs into practical considerations that have a positive impact on nations' relations and interactions. Only through practice will nations eventually reach a level of trust and confidence, leading them to move negotiations on more delicate cyber security issues forward.

All three regional organisations discussed here have now adopted some CBMs and are currently discussing additional ones.[28] Member States have come a long way towards agreeing on these different sets of CBMs; but in order to put them in practice, national policy structures and capacities need to be in place. This process is commonly referred to as implementation and requires commitment from involved States, and support of external experts and consultants.

Given current emphasis on implementation across the regions, it is important to critically review how it can be most effective and achieve the desired results. A significant component of successful implementation involves proper guidance and assistance by a neutral actor with sufficient cyber security expertise, as well as knowledge about the respective nation. Given their long-standing engagement in the respective regions, the OSCE, the OAS and the ASEAN are uniquely equipped to provide customised support and guidance on the regional and sub-regional levels. Moreover, regional organisations have been a perfect platform for bridge-building exercises[29] like this for quite some time. However, targeted capacity-building needs to be provided on the national level to ensure proper engagement in CBMs. Workshops are one way of solving this issue; but raising the implementation rate of cyber CBMs requires a whole-of-government approach.

Capacity-building efforts on the working level might only have a small impact on the CBM implementation process due to the lack of awareness amongst high-level politicians and policymakers. While cyber security is widely covered in many media outlets these days, there still seems to be a certain degree of scepticism among high-level politicians and policymakers about the policy component of cyber security. Moreover, given that cyber security is a cross-cutting issue, normally addressed by several ministries, sometimes division of labour is unclear or not clearly defined. Most nations have national cyber security strategies or other strategy documents that explicitly address these issues. This is a starting point for any international effort to further enhance cyber stability, such as CBMs or norms of responsible State behaviour.

[28] The OSCE is an exception here, as the set of 16 CBMs is already quite advanced there. Discussions on a third set are therefore not a priority at this point.

[29] See following sub-chapter for a series of examples.

One way to facilitate enhanced implementation would therefore consist of its inclusion and clear reference in national strategy documents, such as cyber security strategies or defence strategies. This has the positive side-effect of helping nations better read each other, which already constitutes a confidence-building activity *per se*. Some regional organisations, such as the OAS, have been extensively involved in the development of national cyber security strategies. Synchronising such activities with the UN GGE process and other regional organisations would provide ample potential to further increase the impact of UN GGE reports, as well as harmonise national, regional and international efforts on cyber/ICT security.

The following sub-chapters will provide a summary of the OSCE, OAS and ASEAN/ARF CBM- and norm-related efforts, with a view to subsequently outlining how they connect to each other, as well as to the global discussion on the UN level.

4.2. ORGANIZATION FOR SECURITY AND CO-OPERATION IN EUROPE (OSCE)

The OSCE has engaged in cyber/ICT security CBMs since 2013; and has passed two sets of CBMs, and two Ministerial Council Decisions[30] on cyber/ICT security. It continues to be a platform used by nations with significantly diverging interests due to its focus on practical measures rather than international policy or law components, which are traditionally covered by the UN. Thus, despite the ongoing political tensions between participating OSCE States, cyber/ICT security continues to be addressed by it, most recently through a series of sub-regional capacity-building and awareness raising workshops.[31] Just like the CBMs as a whole, these events are aimed at reducing tension between States by enhancing transparency, fostering collaboration and building trust.

As a first step, the OSCE set up an Informal Working Group in 2012.[32] This provided a platform to engage in structured, but still informal, discussions on CBMs. The first set of OSCE CBMs (2013) established official Points of Contact (PoC) and communication lines to prevent possible tensions resulting from cyber activities.[33] The second set (2016) focussed on further enhancing co-operation between

[30] OSCE, Ministerial Council Decision No. 5/17 in 2017, available at: https://www.osce.org/chairmanship/361561 and Ministerial Council Decision No.5/16 in 2016 - available at: https://www.osce.org/cio/288086.

[31] OSCE, Press release: OSCE organizes sub-regional training event on cyber/ICT security in Astana, 12 December 2017, available at https://www.osce.org/secretariat/362201; OSCE, Press release: OSCE co-organizes sub-regional training course in Bucharest on role of information and communication technologies in context of regional and international security, 28 June 2018, available at: https://www.osce.org/secretariat/386139; OSCE, Event discription: Sub-regional training on the role of ICTs in the context of regional and international security, available at: https://polis.osce.org/subregional-training-role-icts-context-regional-and-international-security.

[32] OSCE, Permanent Council Decision No. 1039 in 2012, available at: https://www.osce.org/pc/90169.

[33] OSCE, Permanent Council Decision No. 1106 in 2013, available at: https://www.osce.org/pc/109168.

participating States: including, for example, effective mitigation of cyber-attacks on critical infrastructure which could affect more than one participating State.[34] The 16 voluntary CBMs can be broadly categorised in three clusters: 1) *Posturing* CBMs, which allow States to "read" another State's posturing in cyberspace in order to make cyberspace more predictable; 2) *Communication* CBMs, which offer opportunities for timely communication and co-operation, including to defuse potential tensions; and 3) *Preparedness* CBMs, which promote national preparedness and due diligence to address cyber/ICT challenges.

Subsequently, the OSCE's focus has shifted from developing additional CBMs towards ensuring that all States properly implement the existing ones through practical support. This includes the use of the OSCE Communications Network "to address security of and in the use of information and communication technologies […] upon the identification of contact centres/points for cyber/ICT security-related communications within capitals".[35] Having two sets of CBMs and an extensive mandate to drive implementation forward, OSCE is focussing its efforts more than ever on making its CBMs operational through increased targeted support and capacity-building for OSCE participating States. This is highly connected to global discussions within the UN, as norms of responsible State behaviour need to be encouraged, supported and fostered through the increased implementation of the CBMs.

The OSCE has launched numerous projects to enhance CBMs. Several of these initiatives can be seen as complementing and taking forward the work being done at the UN GGE. Others may even generate ideas which have yet to be covered by UN GGE reports. For example, as a recent effort to increase ownership and targeted implementation, the OSCE launched an "adopt a CBM initiative" within the Informal Working Group in late 2017.[36] States that formally 'adopt' a CBM bring forward proposals on how to advance its respective implementation, use or impact within the OSCE community. Another development features scenario-based discussions, where government officials are exposed to the practical application of CBMs and norms of responsible State behaviour.[37]

Similarly, since 2017, the OSCE has organised sub-regional training for policymakers, technical experts and private sector representatives; and provided small-scale simulations for PoCs to review how much time participating States require to reply to a request for assistant and/or provide information to an issue at hand. There is

[34] OSCE, Permanent Council Decision No. 1202 in 2016, available at: https://www.osce.org/pc/227281.

[35] OSCE, FSC.DEC/5/17, Use of the OSCE Communications Network to Support Implementation of Permanent Council Decisions No. 1039, No. 1106 and No. 1202, 19 July 2017, FSC.DEC/5/17, available at: https://www.osce.org/forum-for-security-cooperation/331821?download=true.

[36] Velimir Radicevic, Preventing cyberwar: the role of confidence-building measures and associated OSCE efforts, 3 December 2018, Presentation at the Institute for Higher National Defence Studies.

[37] OSCE, Press release: New technological features, policy engagement and public-private partnerships as ways to lower risks of cyber conflicts in focus at Rome Conference, 28 September 2018, available at: https://www.osce.org/chairmanship/397853.

also a separate project to promote operationalisation of the network of policy and technical PoCs by enhancing its functioning, both as a crisis communication network and a platform for co-operation. For the purpose of creating more transparency, OSCE also organises, among other activities, a series of bilateral country visits for PoCs of non-like-minded States. The visits aim to help bridge the largest divides between States in the OSCE area in terms of trust, threat perceptions, approaches to cyber/ICT security, capacities and strategic priorities; and explore commonalities and avenues of co-operation.

Furthermore, with the purpose of promoting, assisting and fostering the implementation process of existing cyber/ICT CBMs, in 2016, the OSCE launched a project that aims to identify and prioritise national implementation challenges. Within this project, it facilitates the creation of national implementation roadmaps and customised capacity-building assistance plans in co-operation with partners such as the Global Forum on Cyber Expertise (GFCE). The latter will include mapping current capacity-building initiatives by other international entities, which could also address CBM implementation challenges on the national and regional levels and therefore complement pertinent OSCE activities.

4.3. ORGANIZATION FOR AMERICAN STATES (OAS)

The OAS uses its Inter-American Committee against Terrorism (CICTE) and the Cyber Security Program to drive its work on cyber security forward. The OAS's mission is to "build and strengthen cyber-security capacity in the Member States through technical assistance and training, policy roundtables, crisis management exercises, and the exchange of best practices related to information and communication technologies".[38] Among the main objectives of the Secretariat are to "establish national 'alert, watch, and warning' groups, also known as Computer Security Incident Response Teams (CSIRTs)".[39]

The OAS has always had a strong emphasis on capacity-building: for example, through supporting the development of cyber security strategies. It has facilitated more than 30 Cyber Maturity Model deployments by the Oxford University Global Cyber Security Capacity Centre among its Member States.[40] Recently, it has shifted its capacity-building efforts towards more specific topics. For example, similarly to the OSCE, the OAS has also engaged in a series of sub-regional workshops on industrial control systems and critical infrastructure in the electricity sector, on the protection of

[38] OAS, Cyber Security, 2019, available at: https://www.oas.org/en/topics/cyber_security.asp.

[39] Id. At the 2004 OAS General Assembly, the Member States approved Resolution AG / RES. 2004 (XXXIV-O/04), "A Comprehensive Inter-American Strategy to Combat Threats to Cybersecurity: A Multidimensional and Multidisciplinary Approach to Creating A Culture of Cybersecurity".

[40] Oxford Martin School, CMM Assessments Around the World, August 2018, available at: https://www.sbs.ox.ac.uk/cybersecurity-capacity/content/cmm-assessments-around-world.

critical infrastructures, cyber security and border protection;[41] as well as workshops on the applicability of international law cyber operations in the Americas.[42]

Moreover, having recognised the importance of regional implementation of the UN GGE reports through practical means, in 2017, the CICTE decided to establish a working group on co-operation and CBMs in cyberspace.[43] In 2018, a draft set of "Cyber CBMs for the Inter-American System"[44] was adopted by the CICTE and the OAS General Assembly with a proposed plan of action to establish additional measures.[45] Each OAS Member State will, as a first step, be asked to determine a national focal point, who will act as a first responder on the policy level should an incident concerning cyber security threaten relations between States. Moreover, going forward, OAS Member States will commence sharing information on national cyber policies, strategies and doctrines in a more formalised way.

4.4. ASSOCIATION OF SOUTHEAST ASIAN NATIONS (ASEAN) AND THE ASEAN REGIONAL FORUM

As the ASEAN's regional emphasis has been on economic progress and development, it launched its international cyber security efforts with an emphasis on international co-operation and harmonisation of policies, particularly with regard to cyber crime.[46] Given the increase in small and medium-sized enterprises (SMEs) which work mostly online, governments seem to have felt an increasing responsibility to secure their operational environment; hence the emphasis on cyber crime. Similarly, efforts undertaken to protect critical infrastructures can be understood as an attempt to protect the increasing amount of services provided online within the region.

[41] OAS, Sub-Regional Workshop on Industrial Control Systems and Critical Infrastructure in the Electric Sector, 2017 available at: https://www.sites.oas.org/cyber/EN/Pages/Events/eventsdet.aspx?docid=102; OAS, Subregional Workshop on Protection of Critical Infrastructures: Cybersecurity and Border Protection, 2017, available at: https://www.sites.oas.org/cyber/EN/Pages/Events/eventsdet.aspx?docid=99.

[42] The legal courses are jointly organised by the Secretariat of the CICTE and the Ministry of Foreign Affairs of the Netherlands. See OAS, The Hague Process: Courses on the International Law Applicable to Cyber Operations, 2017, available at: https://www.sites.oas.org/cyber/EN/Pages/Events/eventsdet.aspx?docid=90; Autoridad Nacional para la innovación gubermental, Panama, November 2018, available at: http://innovacion.gob.pa/noticia/3231.

[43] OAS, Inter-American Committee against Terrorism, Establishment of a Working Group on Cooperation and Confidence-Building Measures in Cyberspace, OEA/Ser.L/X.2.17, CICTE/RES. 1/17, 10 April 2017, available at: http://scm.oas.org/doc_public/ENGLISH/HIST_17/CICTE01114E07.doc.

[44] CICTE/GT/MFCC-7/17 rev.2, Inter-American Committee Against Terrorism (CICTE): Regional confidence-building measures (CBMs) to promote cooperation and trust in cyberspace, available at: http://scm.oas.org/doc_public/ENGLISH/HIST_18/CICTE01179E05.doc.

[45] The proposed text was approved in May 2018 by the Inter-American Committee against Terrorism: CICTE/RES.1/18, Inter-American Committee Against Terrorism (CICTE): Regional confidence-building measures (CBMs), to promote cooperation and trust in cyberspace, OEA/Ser.L/X.2.18 and in June 2018 by the OAS General Assembly through Resolution AG/RES. 2925 (XLVIII-O/18): http://scm.oas.org/doc_public/ENGLISH/HIST_18/AG07745E03.doc.

[46] NATO CCD COE, ASEAN Regional Forum Reaffirming the Commitment to Fight Cyber Crime, INCYDER, 20 July 2013, available at: https://ccdcoe.org/asean-regional-forum-reaffirming-commitment-fight-cyber-crime.html.

However, given the lack of agreement on the UN GGE 2017 report under its Singaporean Chairmanship, discussions within the ASEAN have increasingly looked at how it could move discussions on the four UN GGE pillars forward in its own region.[47] This also resembles a shift from compartmentalised cyber security efforts to a more strategic conversation on the challenges posed.

As a result, through a series of ministerial meetings, norms and CBMs rose to the top of the cyber security agenda, resulting in a formal endorsement of the 11 norms recommended by the UN GGE 2015 report during the ASEAN Ministerial Conference on Cybersecurity (AMCC) in September 2018.[48] As Elina Noor rightly points out, "The seeds of a more strategic conversation on positioning ASEAN within the norm-setting agenda in cyberspace have now finally been sown".[49] Shortly afterwards, ASEAN ministers formally affirmed the AMCC outcome and "noted the agreement by the relevant Ministers: (a) on the need for a formal ASEAN cybersecurity mechanism to coordinate cyber policy […]."[50] As a next step, the ASEAN Network Security Action Council will "prepare a proposal for a formal ASEAN cybersecurity coordination mechanism for consideration by relevant ASEAN sectoral bodies. [ASEAN Ministers] agreed that in the meanwhile, the AMCC should continue to serve as the interim and non-formal ASEAN platform for cybersecurity".[51]

These developments were accompanied by the Sydney Recommendations on Practical Futures for Cyber Confidence Building in the ASEAN region, which outlined how cyber confidence building can be moved forward.[52] At present, five CBMs are being discussed in the ASEAN-ARF Inter-sessional group and will probably resemble similar pathways taken by the OSCE and the OAS.[53]

4.5. INCREASING ROLE AND INTERCONNECTEDNESS OF REGIONAL ORGANISATIONS

Previous sub-chapters have outlined three regional organisations' efforts in shaping

[47] Caitríona Heinl, Can ASEAN Continue to Improve Cybersecurity in the Region and Beyond? March 22, 2018, available at: https://www.cfr.org/blog/can-asean-continue-improve-cybersecurity-region-and-beyond.

[48] CSA Singapore, Singapore International Cyber Week 2018 - Highlights and Testimonials, September 20, 2018, available at: https://www.csa.gov.sg/news/press-releases/sicw-2018---highlights-and-testimonials.

[49] Elina Noor, ASEAN Takes a Bold Cybersecurity Step, *The Diplomat*, October 4 2018, available at: https://thediplomat.com/2018/10/asean-takes-a-bold-cybersecurity-step/.

[50] ASEAN, Chairman's Statement of the 33rd ASEAN Summit, Singapore, November 2018, available at: https://asean.org/storage/2018/11/33rd_ASEAN_Summit_Chairman_s_Statement_Final.pdf.

[51] Id.

[52] Sydney Recommendations on Practical Futures for Cyber Confidence Building in the ASEAN region, September 2018, available at: https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2018-09/Sydney%20 recommendations_Cyber-ASEAN.pdf?kwrNP4FHCYxE9oGVhxzchUvF3rx11hoG.

[53] ARF Inter-Sessional Meeting on Security of and in the Use of Information and Communication Technologies and 1st ARF-ISM on ICTs Security: https://www.mofa.go.jp/press/release/press4e_002011.html.

the landscape of cyber security-related norms and practical CBMs. There are two main conclusions we can draw from this.

Firstly, the aforementioned regional organisations initially focused only on certain topics related to cyber security; and have thereby been keeping their work narrow and not as broad as the UN GGE reports. Recently, all have gradually expanded their scope into additional UN GGE pillars, recognising that one-sided emphasis only works for a limited amount of time. In fact, some regional organisations may have the mandate to focus on areas not covered by the UN GGE, such as Internet infrastructure, content management, freedom of expression, privacy protection, digital economy and introduction of new technologies. All in all, regional organisations eventually seem to have acknowledged that their initially limited efforts can become more substantiated if multiple, or ideally all four, pillars are addressed within each region. Since the OSCE, OAS and ASEAN are coming from different perspectives and originally had different foci, their comprehensive approach, covering most if not all four pillars, provides ample opportunity to support each other's efforts, as will be demonstrated in the next chapter.

Secondly, even though regional organisations have expressed their appreciation of the proposed norms, there appears to be some concern over the lack of consensus following the most recent UN GGE efforts. Commentators have suggested that regional organisations such as the ASEAN should not wait for the UN GGE to be reconvened: even if consensus will be achieved and additional norms agreed to, this will take time.[54] Instead, as detailed already, it has been proposed that the ASEAN should start working on implementing these norms and possibly shaping new ones "in ways that correspond to ASEAN Member States' needs and contexts, and can take the proactive role instead of waiting for larger States to dictate the rules of the road".[55] This clearly points to the interest as well as capacity to push towards more tailor-made solutions on the regional level. At the same time, it raises the question of whether the UN GGE is the most suitable global platform for regional organisations to harmonise their efforts and make themselves heard internationally.

Therefore, before concluding on how to better incorporate their views into global discussions, the following chapter will also look at how regional organisations may benefit from enhanced inter-regional activities.

[54] Benjamin Ang, Next steps for cyber norms in ASEAN, 2018, https://www.rsis.edu.sg/wp-content/uploads/2018/10/CO18174.pdf.

[55] Id.

5. CROSS-CUTTING BENEFITS OF INTER-REGIONAL AND GLOBAL OPPORTUNITIES

Instead of discussing possible new international platforms for developing cyber norms, focus should remain on maximising the impact of what is already agreed upon and established. In order to achieve that, national, regional and global efforts need to be linked in a coherent way and practical efforts focused on implementation should receive priority. Moreover, discussions need to become more nuanced, streamlined and channelled into the right structures. Only by doing so can States focus on implementing and operationalising agreed norms and CBMs.

Firstly, when discussing additional norms to be added to the framework of the already agreed UN GGE 2015 report, it would help to reflect on which topics are actually crucial for the maintenance of peace and stability among States at this point. Secondly, it is also key to parse out the vast number of topics within the cyber security umbrella and identify fitting fora for each issue. Global institutions like the UN, regional organisations like the OSCE, like-minded entities and fora that facilitate dialogue among non-like-minded States all have their value. Maximising the effect and impact of existing platforms by using the right platform for the respective topic at hand is key. When it comes to linking regional and global efforts, the UN, specifically the UN GGE but potentially the new OEWG too, provides room for such co-operation.

In the following sub-paragraphs, we will highlight elements of inter-regional and global platforms which we believe would benefit from the greater inclusion of regional organisations.

5.1. INTER-REGIONAL DEVELOPMENTS

As outlined above, the OSCE, OAS and ASEAN are the key actors worldwide to enhance international cyber stability through their cyber/ICT CBM catalogue, capacity-building efforts, international co-operation and dialogue. When applying a global lens, each of them is just one out of several regional organisations that are trying to foster regional co-operation and offer policy advice on cyber/ICT security-related issues within their area of operations. In order to better understand similarities, differences and room for additional collaboration, there is significant potential for an inter-regional initiative that aims at establishing knowledge and best practices exchange amongst regional organisations working on cyber/ICT security issues.

A sustainable network with other regional organisations developing cyber/ICT CBMs as well as capacity-building initiatives would be beneficial in several aspects. Such an inter-regional approach would facilitate gaining specific insights into related cyber/ICT security initiatives by other international organisations, as well as identify common interests and maximise the impact of potential overlapping initiatives by collaborating or planning joint workshops, training, conferences etc. Developing working-level connections among the regional organisations working on cyber/ICT security CBMs would facilitate co-operation and communication. Exchanging best practices and specific knowledge about regional characteristics, governmental structures or policy challenges related to cyber/ICT security issues would provide good grounds for furthering trust and collaboration. Equally relevant would be to explore the possibilities of joint CBM implementation initiatives in States that are part of several regional organisations engaged in cyber/ICT security initiatives; and identify possibilities of further linking capacity-building initiatives with CBMs.

One option for such inter-regional cooperation would be the Global Forum on Cyber Expertise (GFCE). The launch of the GFCE was a result of the 2015 Global Conference on Cyber Security. Initially created by the Dutch government, the GFCE is now a "global platform for countries, international organisations and private companies to exchange best practices and expertise on cyber capacity building".[56] By its very design and mandate, the GFCE is an ideal platform for an international best practice exchange, collaboration and co-operation. In the mid- to-long run, this initiative could establish a sustainable hub for constructive exchange amongst regional organisations and facilitate resource and capacity sharing, information exchange and long-term co-operative projects and initiatives, while avoiding unnecessary duplication amongst regional organisations.

States of involved regional organisations would also benefit from this initiative since this platform is likely to reduce duplication and enhance global awareness of capacity needs across regions. Moreover, more effective inter-regional co-operation is likely to create improved distribution of resources amongst regional organisations and streamline cyber stability efforts across regions. Helping regional organisations better co-ordinate amongst themselves could also help States with their own international cyber/ICT policy initiatives, as most cyber/ICT security related initiatives are highly intertwined and connected across regions or even globally: and thus gain effectiveness from initiatives that are already harmonised between regional organisations. Equally, additional support from selected States through the GFCE could ensure political buy-in, increase the impact of this initiative and generate interest in operationalising this network for enhancing pertinent national capacities.

Such a platform could also support the effective implementation of the CBM catalogues of the OSCE, OAS and ASEAN by supplementing regional organisations'

[56] GFCE, about page, available at: https://www.thegfce.com/about.

efforts with additional capacity-building and awareness-raising efforts among GFCE members.

Another promising inter-regional development was a workshop organised in Geneva in January 2019 by the Center for Security and International Studies, and the United Nations Institute for Disarmament Research, on “The Role of Regional Organizations in Strengthening Cybersecurity and Stability”.[57] While this did not result in the establishment of a formal inter-regional body for exchange, the workshop itself was already a welcome development: for the first time, it provided representatives of regional organisations with the opportunity to constructively engage with UN officials, and discuss in concrete terms how regional contributions and expertise could best be integrated into UN-level discussions. As all regional organisations mentioned in the UN GGE mandate were present in the room, it also allowed them to discuss amongst themselves how they could best co-ordinate their input across regions.[58]

During the discussions, there seemed to be overall agreement amongst participants that regional organisations have been the enablers of capacity-building, awareness raising and CBM development. As a result, regional organisations have significant untapped potential to contribute to international cyber security policy negotiations. Such efforts would not seek to replace UN-level discussions, but to complement, support and incorporate regional perspectives into the discussions. It was reiterated that regional organisations have a unique advantage in launching certain activities, as they have a better grasp of regional developments and national preferences, which play a vital role in implementing norms and CBMs.

5.2. GLOBAL PLATFORMS

Global efforts such as the UN GGE are clearly interconnected with the work of regional organisations. When looking at the UN GGE 2015 report, many of the 11 norms and principles are already closely connected to existing capacity-building or CBM efforts. In fact, several studies have confirmed both the influence of the UN GGE on regional CBMs, and the potential of regional measures to complement the UN GGE measures.[59] However, what is missing is a clear structure and framework for enhancing the positive, mutually reinforcing impact. Clarifying how such parallel

[57] See UNIDIR Press Release, The 2nd International Security Cyber Issues Workshop Series: The Role of Regional Organizations in Strengthening Cybersecurity and Stability, available at http://unidir.org/programmes/security-and-technology/the-2nd-international-security-cyber-issues-workshop-series-the-role-of-regional-organizations-in-strengthening-cybersecurity-and-stability.

[58] Overview of the Group of Governmental Experts and Open-ended Working Group Processes, presentation by Gillian Goh, Political Affairs Officer and Cyber Team Leader, UN Office of Disarmament Affairs, available at: http://unidir.org/files/medias/pdfs/overview-of-the-group-of-governmental-experts-and-open-ended-working-group-processes-eng-0-786.pdf.

[59] See, e.g., footnote 17, pp.129-153; DiploFoundation, Towards a secure cyberspace via regional co-operation, 2017, available at: https://www.diplomacy.edu/sites/default/files/Diplo%20-%20Towards%20a%20secure%20cyberspace%20-%20GGE.pdf.

efforts can be harmonised and brought together should be part of the discussions within the newly formed UN GGE and OEWG.

Even though traditionally the UN GGE process does not directly involve non-State actors, more formalised input from regional organisations could benefit the overall process by presenting a consolidated view of its members, support the implementation of the agreed principles and enforce capacity-building efforts and awareness raising. Despite the lack of an explicit reference to regional organisations in its mandate, the UN OEWG should also consider how to engage with regional organisations. Overall, when designing the processes for further including regional organisations' efforts at the UN level, we suggest keeping in mind the following proposals.

a) Choosing the Right Venue and Format

The two somewhat overlapping proposals for taking forward the norms-building process at the UN level (described in Chapter 2) pose a dilemma to all involved stakeholders, ranging from States to regional organisations, which have previously been directly or closely involved following UN GGE reports. Which of the two working groups should be given more attention? Which one develops more relevant information for regional organisations? While these questions cannot be answered yet, only the UN GGE mandate explicitly invites regional organisations for consultations. We therefore suggest embracing this invitation, while also clarifying how regional organisations can contribute to discussions within the OEWG. For the benefit of the complementarity of efforts and the potential for convergence, regional organisations, even if they are explicitly mentioned in the UN GGE mandate, should try to identify means to actively engage with both groups.

However, given that at this stage regional consultations are only foreseen with the UN GGE, most of the following recommendations are more applicable to regional collaboration with it. Overall, close collaboration with regional organisations, mentioned in the UN GGE mandate, seems more practical, as the new GGE proposal follows a concrete timeline and specifically incorporates consultations with regional organisations. We therefore argue that it makes most sense for regional organisations that were explicitly mentioned in the UN GGE mandate to engage without reservations. On the other hand, even though the OEWG format does not foresee a strictly defined timeline,[60] it promises a multi-stakeholder approach,[61] therefore leaving room for the potential inclusion of consultations with regional organisations as well.[62]

[60] The OEWG's mandate asks for the submission of a report on the results of the study to the General Assembly at its 75th session, but leaves room for continued discussions after this deadline.

[61] Alex Grigsby, The United Nations Doubles Its Workload on Cyber Norms, and Not Everyone Is Pleased, 15 November 2018, available at: https://www.cfr.org/blog/un-doubles-its-workload-cyber-norms-and-not-everyone-pleased.

[62] Other entities which have not been invited to consultations with the UN GGE are facing an additional dilemma. They may be forced to focus their collaboration with the OEWG, as it addresses a wider range of stakeholders such as the private sector, non-governmental organisations and academia.

Moreover, the tentative meeting timeline[63] allows sufficient room for collaboration and information exchange between the UN GGE and the OEWG. This may be challenged by political differences, but could ideally result in a division of tasks or an assurance of avoiding overlap and/or contradiction between their respective reports.

b) Building on Existing Global-Regional Synergies

Our analysis of the ongoing efforts of regional organisations reveals a number of areas where there is a clear link between the UN GGE proposals and the work of regional organisations. For example, the limiting norm that "states should not knowingly allow their territory to be used for internationally wrongful acts using ICTs" has clear connections to national capacities to address malicious or criminal use of ICT infrastructure, an area where ASEAN has been particularly active over recent years, as described in the previous chapter. Moreover, the norm that "states should not conduct or knowingly support ICT activity that intentionally damages critical infrastructure" neatly aligns with multiple critical infrastructure protection efforts, such as OSCE CBM 15 or the OAS's capacity-building workshops.

Another example is the limiting norm that "States should not conduct or knowingly support activity to harm the information systems of another State's emergency response teams (CERT/CSIRTS) and should not use their own teams for malicious international activity", which directly relates to the OAS's capacity-building efforts; in particular, the development of CSIRTS among its members. Similar comparisons can be conducted for the good practices and positive duties included in the UN GGE's 2015 report.

These examples underline the large potential in systematically synchronising regional and global efforts. Building on already existing areas of collaboration will allow for more swift progress in the implementation of agreed UN GGE norms.

While previous CBMs agreed at UN level largely correspond to CBMs already agreed upon at regional level,[64] there is the possibility of additional CBMs being agreed in the UN. If the UN GGE or OEWG decide to propose additional CBMs, close collaboration with regional organisations would be beneficial for both sides, as mutually reinforcing efforts and regional expertise, needs and suggestions would most likely increase the impact, effectiveness and level of adoption of the UN-level CBMs.

[63] Footnote 58, slide 3.

[64] As Henry Rõigas and Tomáš Minárik outline: "The CBMs in the report largely correspond to those already adopted under the auspices of the OSCE in 2013. The key difference, however, is that, unlike the OSCE, the report does not establish or propose concrete cooperation channels". 2015 UN GGE Report: Major Players Recommending Norms of Behaviour, Highlighting Aspects of International Law, CCDCOE, available at: https://ccdcoe.org/incyder-articles/2015-un-gge-report-major-players-recommending-norms-of-behaviour-highlighting-aspects-of-international-law/.

c) Regional Organisations as Incubators for New Ideas

As outlined in the previous chapter, regional organisations have developed their own innovative ideas on how to address some of the most pertinent international cyber security policy challenges. These efforts have provided a positive contribution to international discussions on cyber security and remain a key component of effective implementation of globally accepted rules and norms. The OSCE's "adopt a CBM initiative" could be applied similarly to norms. Such targeted norm campaigns, driven by volunteer States, may provide new room for suggestions on how these norms can properly applied and implemented.

Also, unlike the OSCE, the UN GGE report does not "establish or propose concrete cooperation channels", since "the measures proposed in the report mainly relate to information exchange and developing international cooperation mechanisms between national entities dealing with ICT security".[65] Thus, the 2021 UN GGE report now has the potential to critically reflect on how existing co-operation channels can be made available for cyber security issues, or how the carefully constructed networks within different regions in the world could be connected.

d) Targeted Capacity-Building

As a positive example, the OAS's targeted capacity-building has helped its Member States to advance their national cyber security competencies significantly. While the OAS's efforts were constrained by its mandate, a dedicated UN capacity-building initiative, designed to help States that want to properly implement UN GGE reports but lack the resources to do so, would certainly contribute to a more coherent international cyber security policy landscape and eventually make cyberspace safer and more stable overall. With the OAS's existing expertise, the ASEAN Singapore Cybersecurity Centre of Excellence, the ASEAN Japan Cybersecurity Capacity Building Centre, and the OSCE's capacity-building workshop series, such a UN capacity-building initiative may be able to tap into regional areas of expertise and combine them in a way no regional organisation could by itself.

e) Not Re-Inventing the Wheel: Adding a Lessons Learned Instrument

When looking at potential focus areas for the newly created UN GGE, this paper argues that representatives should consider practical steps towards implementing previously agreed UN GGE reports. Especially after the lack of consensus for parts of the 2017 report, an initial focus on practical procedures could reduce the level of politically sensitive issues in the discussion while still making some meaningful progress on the issues at hand. Looking back at the overview of practical matters offered by regional organisations outlined in the previous chapter, this paper argues that global-regional collaboration within the UN GGE could easily include sharing lessons learned and

[65] Id.

experience from regional organisations. Through the lessons learned process, norms can be further developed and gaps in the existing international frameworks identified.

f) Regional Roadmaps and Joint Implementation Efforts

Another component of global-regional collaboration within the new UN GGE could involve regional roadmaps on the agreed measures, norms and initiatives. Having regional organisations take part in the preparation of concrete implementation roadmaps could have several benefits when looking at the potential impact of the new report. Instead of publishing its new report with no concrete implementation follow-up procedure, the UN GGE could involve regional organisations early on, to develop a customised workplan for each region. This could significantly speed up the implementation process, increase the coherence of norm implementation, facilitate the use of regional capacities and improve linkages between existing regional efforts and newly developed norms and initiatives within the UN GGE report. This paper therefore suggests that such roadmaps should be a component of the UN GGE 2021 report.

g) Involve More Funding

Another potential benefit of increased global-regional cooperation lies in project-based work and funding. If a certain initiative is included into the UN GGE process without including regional organisations' considerations, it might prove difficult for regional organisations to follow up if their mandate does not overlap with the initiative at hand. Having regional organisations be part of the framing procedure would prove helpful in preparing regional follow-up projects and attracting external funding for the new initiative. Moreover, if new initiatives within the UN GGE report overlap with regional organisations' mandates, it is likely that regional implementation would be less controversial and therefore States would probably be less reluctant to provide funding.

h) Enhanced Timing and Priorities

Lastly, another potential benefit through greater global-regional exchange relates to a more structured norms discussion in terms of timing and priorities. Regional organisations, especially those with national offices or extensive national capacity-building efforts, have extensive insights into national concerns and can therefore evaluate whether the proposed UN GGE priorities line up with national ones. Such a procedure might also have a positive impact on the implementation of the norm in the respective region. Knowing which norm lines up with national or regional priorities might prove useful to the UN GGE and allow it to develop certain norm implementation pilot projects in the respective regions.

However, even if formally hearing out regional organisations sounds good on paper, resolution A/C.1/73/L.37 leaves open how suggestions and concerns raised by regional organisations will be incorporated into the UN GGE deliberations. Besides this concern, we believe that our proposals should provide the stakeholder meetings, to be organised in 2019 by the United Nations Office of Disarmament Affairs, with sufficient concrete proposals on how to move the global-regional cooperation forward within the UN GGE.

6. CONCLUSION

This paper concludes by confirming that the UN GGE continues to have significant merit and is a much needed platform for enhancing international cyber stability negotiations. However, the deliberations and final report of the 2019-2021 negotiations could significantly benefit through increased collaboration with regional organisations. While the new UN OEWG provides room for private sector and NGO input, the new UN GGE mandate opens an entirely new opportunity for enhanced collaboration between the UN and regional organisations. This could lead to the development of a clear framework for enhancing the positive mutually reinforcing impact of global and regional efforts. This should also include a discussion on clarifying parallel efforts, which could be harmonised and brought together.

Another positive result of the increased exchange between the UN and regional organisations is that this opens up the possibility of expanding the scope of information, suggestions and expertise which is incorporated into UN GGE deliberations.

Furthermore, looking at the already agreed norms and principles, several areas of global-regional collaboration can be observed. There is large potential in systematically synchronising regional and global efforts. Regional organisations are already acting as an incubator for national implementation of UN GGE reports, and have developed their own innovative ideas on how to address some of the most pertinent international cyber security policy challenges.

Another potential benefit of increased global-regional cooperation lies in project-based work and funding. If a certain initiative is included in the UN GGE process without allowing for regional organisations' considerations, it might prove difficult for them to follow up if their mandate does not overlap with the initiative at hand. Having regional organisations be part of the framing procedure would prove helpful in preparing regional follow-up projects and attracting external funding for the new initiative.

When looking at the coherence between UN GGE reports and regional organisations' activities, this paper argues that there is significant potential in lining them up through a joint workplan, which could be annexed to the new UN GGE report. Such a workplan would provide the drafting process of the 2021 UN GGE report with the opportunity to critically reflect on how existing co-operation channels can be made available for cyber security issues and how carefully constructed networks within different global regions could be connected. Moreover, such a workplan may include regional roadmaps on the agreed measures, norms and initiatives of the new report. Having regional organisations take part in the preparation of concrete implementation roadmaps could significantly improve the implementation process and overall impact of the new report.

Another potential benefit of customised regional roadmaps relates to a discussion of more structured norms in terms of timing and priorities. Regional organisations, especially those with national offices or national capacity-building efforts, have extensive insights into national concerns and can therefore evaluate whether the proposed UN GGE priorities line up with national ones. Knowing which norm lines up with national or regional priorities might prove useful to the UN GGE and allow them to develop certain norm implementation pilot projects in the region in question. These would also have a positive benefit for concrete and practical norms implementation. The UN GGE can profit from the many years of regional experience in capacity-building and norm implementation.

Lastly, a dedicated UN capacity-building initiative, jointly developed with regional organisations and aimed at helping those States that want to properly implement UN GGE reports but lack the resources to do so, would contribute towards a more coherent international cyber security policy landscape; and eventually make cyberspace safer and more stable overall.

While the new UN GGE provides regional organisations with the chance to make themselves heard, this paper also argues for enhanced inter-regional collaboration amongst the most active of them. The OSCE, the OAS and the ASEAN are among the key actors worldwide seeking to enhance international cyber stability through their cyber/ICT CBM catalogue, capacity-building efforts, international co-operation and dialogue. In order to better understand similarities, differences and room for potential collaboration, there is significant potential for an inter-regional initiative which aims at establishing knowledge and best-practices exchange amongst regional organisations working on cyber/ICT security issues.

Such an inter-regional approach would facilitate gaining specific insights into related cyber/ICT security initiatives by other international organisations, identifying common interests and maximising the impact of potentially overlapping initiatives by collaborating or planning joint workshops, training, conferences, etc. Exchanging best practices and specific knowledge about regional characteristics, governmental structures or policy challenges related to cyber/ICT security issues would provide good grounds for furthering trust and collaboration.

Equally relevant would be to explore the possibilities of joint CBM implementation initiatives in States that are part of several regional organisations engaged in cyber/ICT security initiatives, and to identify possibilities of further linking capacity-building initiatives with CBMs.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Layered Sovereignty: Adjusting Traditional Notions of Sovereignty to a Digital Environment

Przemysław Roguski
Lecturer
Chair of Public International Law
Jagiellonian University
Kraków, Poland
przemyslaw.roguski@uj.edu.pl

Abstract: The question of how to define sovereignty in cyberspace is currently one of the most contentious issues in international law. The traditional understanding of sovereignty is based on the assumption of exclusive control over geographically defined territory. However, the global accessibility of computer networks eliminates distance and geography as limiting factors for the exercise of power by States (and non-State actors). This creates a security dilemma: while modern ICTs allow adversaries to challenge States' exclusive authority over 'their' cyberspace, traditional notions of sovereignty appear to limit the States' ability to actively respond to these challenges in foreign networks.

In this paper I argue for a 'layered' understanding of sovereignty in cyberspace. Recent international practice, including national legislation and court decisions relating to jurisdiction over transboundary activities, shows that while States stress the exclusive nature of authority and jurisdiction over the physical layer of cyberspace, the logical and social layers are open to transboundary assertions of jurisdiction. Applying these findings to the general concept of sovereignty in cyberspace, I argue that while the physical layer is covered by State sovereignty by virtue of the principle of territoriality, the logical and social layers of cyberspace may be open to the exercise of State authority based on a criterion of proximity, i.e. whenever the State can establish a genuine link with the digital objects or online personae over which authority is to be asserted.

Keywords: *sovereignty, cyberspace, jurisdiction, territory, Tallinn Manual*

1. INTRODUCTION

One of the functions of international law as a legal system is to allocate, delimit and protect spheres of competence of States.[1] These spheres of competence are tied to the concept of State sovereignty, which is one of the foundational principles of international law. In the classic, post-Westphalian system, sovereignty is understood as exclusive authority of the State over persons and things within a specified territory.[2] All three elements of this definition – the nature of power/authority, its exclusivity and its territoriality – have been challenged by the invention of interconnected global communications networks, in short: cyberspace. Because cyberspace creates a space for storage of and access to information, as well as social interaction regardless of the user's location and irrespective of distances, it creates the perception of a space not restricted by – or even detached from – geography. In other words, cyberspace is perceived as a-territorial.[3] Similarly, cyberspace constitutes a challenge to the nature and exclusivity of authority. The worldwide accessibility of online content poses questions as to the extent of State jurisdiction in cyberspace and creates the possibility of a multitude of overlapping jurisdictions.[4] Additionally, the ease of access to information and communications technology [ICT] and the interconnectedness of computer networks have led to a rising importance of technology companies, individuals and groups of individuals as actors in cyberspace.

In view of these challenges, the question of how sovereignty applies in (and to) cyberspace has been a topic of constant debate among experts, in academia and in the international community. While the 2013 and 2015 Reports of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security [GGE] have confirmed that sovereignty and the international norms and principles that flow from it apply to State conduct in cyberspace, they have left open the meaning and scope of sovereignty with respect to the cyber domain.[5] Since 2015 there has been little progress in this regard. The failure of the 2016-2017 GGE to adopt a consensus report[6] and, most recently, the adoption by the United Nations General Assembly [UNGA] of two

[1] Hermann Mosler, 'Völkerrecht Als Rechtsordnung' (1976) 36 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 6, 39, 48.

[2] Arthur Jennings and Robert Watts, *Oppenheim's International Law* (9th edn, Longmans 1992) para 117.

[3] Nicholas Tsagourias, 'The Legal Status of Cyberspace' in Nicholas Tsagourias and Russell Buchan (eds), *Research Handbook on International Law and Cyberspace* (Edward Elgar Publishing 2015) 22.

[4] Uta Kohl, 'Jurisdiction in Cyberspace' in Nicholas Tsagourias and Russell Buchan (eds), *Research Handbook on International Law and Cyberspace* (Edward Elgar Publishing 2015) 31ff.

[5] UN GGE, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, 24 June 2013, UN Doc. A/68/98 [hereinafter GGE Report 2013], para 20; UN GGE, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, 22 July 2015, UN Doc. A/70/174 [hereinafter GGE Report 2015], para 27.

[6] Michelle Markoff, *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*, 23 June 2017, <https://www.state.gov/s/cyberissues/releasesandremarks/272175.htm> [accessed 11.03.2019].

competing resolutions on the further study of international security dimensions of cyberspace, make clear that the international community is yet to achieve a common understanding on many issues regarding the application of international law in cyberspace, including sovereignty.[7]

Against this background I argue that the Westphalian concept of sovereignty needs to be adjusted to account for the peculiarities of cyberspace. First, I recapitulate the current definition of sovereignty and its connection to the concept of territory. Then I briefly turn to the *Tallinn Manual 2.0* conception of sovereignty and why, in my view, it is too restrictive. After that I discuss examples where the traditional notion of territoriality is challenged in cyberspace and argue that sovereignty in cyberspace should indeed be perceived differently from sovereignty over physical territory. Lastly, I propose to use an analogy to the layered structure of cyberspace to conceptualise how sovereignty operates in cyberspace.

2. THE RELATIONSHIP BETWEEN SOVEREIGNTY AND TERRITORY IN CYBERSPACE

A. The Westphalian Concept of Sovereignty

Sovereignty is a foundational principle of public international law, with its origins going back to Jean Bodin, who understood it as the absolute and indivisible power of the sovereign to make and enforce laws binding his subjects.[8] In its classical form, it signifies *summa potestas, i.e.* the highest authority and the right to exercise its own judgment within a territory.[9] This authority within the State (internal sovereignty) refers to 'the State's exclusive right or competence to determine the character of its own institutions, to ensure and provide for their operation to enact laws of its own choice and to ensure their respect'.[10] By virtue of this sovereignty States have, *inter alia*, the right to: control access to their territory; exercise authority over all persons

[7] During the 73rd Session of the UN General Assembly both the US and Russia, together with their respective allies, introduced draft resolutions relating to the further study of norms on responsible State behaviour in cyberspace. The US-sponsored resolution establishes a new Group of Governmental Experts to continue the work of previous GGEs, while the Russia-sponsored resolution establishes an open-ended working group acting on a consensus basis to further develop the rules, norms and principles of responsible behaviour of States in cyberspace. Instead of negotiating a compromise between the two proposals, the UNGA decided to adopt them both: *Developments in the field of information and telecommunications in the context of international security*, 11 December 2018, UN Doc. A/Res/73/27; *Advancing responsible State behaviour in cyberspace in the context of international security*, 2 January 2019, UN Doc. A/Res/73/266.

[8] Jean Bodin, *Six Livres de la République* (Chez Jacques Du Puys, France, 1577); *See also* Daniel Lee, *Popular Sovereignty in Early Modern Constitutional Thought* (Oxford University Press 2016) 188.

[9] PCIJ, *Customs Régime between Germany and Austria (Protocol of March 19th, 1931)*, Advisory Opinion, 1931 PCIJ Series A/B No 41, sep. opinion Judge Anzilotti at para 13.

[10] Nkambo Mugerwa, 'Subjects of International Law' in Max Sorensen (ed), *Manual of Public International Law* (Macmillan 1968) 253.

and things within their territory as well as over their citizens at home and abroad; enact and enforce laws; and determine the State's political and economic system.[11]

The second requirement of sovereignty (in the Westphalian sense), closely linked to the notion of authority, is territory.[12] In its basic meaning, territory is first and foremost a geographical and spatial construct[13] relating to a physical area of the globe.[14] However, in relation to the concepts of statehood and sovereignty, territory ceases to be only a geographical description and instead becomes a legal and political construct.[15] In its interaction with authority, territory is not only the object of sovereignty, but also the spatial framework in which power and authority are manifested.[16] Competences of a State which flow from its sovereignty, such as jurisdiction, are manifest in largely territorial terms.[17] Moreover, it also functions as the 'container' for sovereignty, limiting its reach by drawing legal and political borders.[18] Territory's importance is such that even the notion of statehood is dependent on the nexus between a population which within a specified geographical space forms a community possessing an effective government.[19] The exclusivity of control over territory as a paramount condition for peace and stability[20] is thus protected against violations through the use of force (Art. 2(4) UN Charter), intervention into internal affairs,[21] as well as any other exercise of power within the territory of another State without that State's consent or the existence of a permissive rule.[22]

B. The Peculiarities of 'Territory' in Cyberspace

Given that the traditional understanding of sovereignty rests upon the exercise of authority within a geographical space, the question immediately arises how it can be applied to cyberspace – a global network of computers, including the information stored therein and the interactions between its users,[23] which is often perceived as

[11] Samantha Besson, 'Sovereignty' in Rüdiger Wolfrum (ed), *Max Planck Encyclopaedia of Public International Law* (Oxford University Press 2011) para 118ff.

[12] Territory has even been described as 'perhaps the fundamental concept of international law', *see* Malcolm N Shaw, 'Territory in International Law' (1982) 1 Netherlands Yearbook of International Law 17, 62.

[13] Sara Kendall, 'Cartographies of the Present: "Contingent Sovereignty" and Territorial Integrity' (2016) 47 Netherlands Yearbook of International Law 83, 84.

[14] Shaw (n 12) 61.

[15] Tsagourias (n 3) 18.

[16] Christian Marxsen, 'Territorial Integrity in International Law – Its Concept and Implications for Crimea' (2015) 75 Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 7, 10.

[17] Daniel Bethlehem, 'The End of Geography: The Changing Nature of the International System and the Challenge to International Law' (2014) 25 European Journal of International Law 9, 14.

[18] Tsagourias (n 3) 17.

[19] See Art. 1 *Montevideo Convention on the Rights and Duties of States*, LNTS No. 3802.

[20] *Indo-Pakistan Western Boundary (Rann of Kutch)* (India v. Pakistan), Award, RIAA XVII 1, 571.

[21] ICJ, *Case Concerning Military and Paramilitary Activities in and Against Nicaragua* (Nicaragua v. United States of America); Judgment of 27 June 1986, ICJ Rep. 1986 p. 14, para 205.

[22] '[T]he first and foremost restriction imposed by international law upon a State is that, failing the existence of a permissive rule to the contrary, it may not exercise its power in any form in the territory of another State', PCIJ, *S.S. Lotus* (France v. Turkey.), Judgment, 1927 PCIJ Ser. A No. 10, at p. 18.

[23] Michael N Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare* (2013) 257.

a-territorial.[24] This problem, of course, is not new. Since the 1990s 'territorialists' and 'unterritorialists'[25] have debated whether cyberspace lies beyond the borders of existing States,[26] is akin to *res communis omnium*[27] or is subject to the jurisdiction of States because it operates on the basis of technical infrastructure within a specific geographic location.[28] As these debates are well-known, they need not be repeated for the purposes of this paper. Suffice it to recall that the distinctiveness of cyberspace is rooted in its 'layered' construction. The most popular models describe between three[29] and seven[30] layers,[31] which together create a space for interaction and communication characterised by three main features: interconnectedness, anonymity and ease of entry.[32] These features, in turn, contribute to the main distinction between cyberspace and traditional space: while the technical components which form the backbone of global computer networks have a unique physical location, their location is not perceived by the users of cyberspace. Rather, the impression of a distinct space is formed by the logical and social layers that construct a global platform for the exchange of information, services and activities, without regard for existing borders between States. Since the international community has declared the principle of State sovereignty to be applicable in cyberspace,[33] the question remains whether traditional principles and rules of sovereignty, such as the prohibition against violations of territorial sovereignty, extend to cyberspace unchanged or whether they need to be modified in order to account for the unique technical circumstances of cyberspace.

[24] Tsagourias (n 3) 22.

[25] Borrowed from Jennifer Daskal, 'Borders and Bits' (2018) 17 Vanderbilt Law Review 179, 181.

[26] John P Barlow, 'A Declaration of the Independence of Cyberspace' (1996) <https://wac.colostate.edu/rhetnet/barlow/barlow_declaration.html> [accessed 11.03.2019]; David R Johnson and David Post, 'Law and Borders - The Rise of Law in Cyberspace' (1996) 48 Stanford Law Review 1367, 1370; Yaroslav Radziwill, *Cyber-Attacks and the Exploitable Imperfections of International Law* (Martinus Nijhoff Publishers 2015) 91.

[27] Darrel C Menthe, 'Jurisdiction in Cyberspace: A Theory of International Space' (1998) 4 Michigan Telecommunications and Technology Law Review 69, 93–94.

[28] Jack Goldsmith and Timothy Wu, *Who Controls the Internet? Illusions of a Borderless World* (Oxford University Press 2006) 73.

[29] The three layer model, consisting of physical, social and logical layers has been first proposed by Yochai Benkler and is applied, with slight modifications, e.g. by the *Tallinn Manual 2.0* or the US military (which distinguishes between physical, logical and cyber-persona layers); see, respectively, Yochai Benkler, 'From Consumers to Users: Shifting the Deeper Structures of Regulation toward Sustainable Commons and User Access' (2000) 52 Federal Communications Law Journal 561, 561; Michael N Schmitt and Liis Vihul (eds), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017) 12 [hereinafter: *Tallinn Manual 2.0*]; Joint Chiefs of Staff, 'Cyberspace Operations, JP 3-12' (2018) I-2.

[30] The Open Systems Interconnection (OSI) Model divides the process of data transmission into seven layers/steps: physical, data link, network, transport, session, presentation and application, see James E Goldman, 'Network Concepts' in Jerry C Whitaker (ed), *Systems Maintenance Handbook* (2nd edn, CRC Press) 17–1; some authors group these into five layers (geographical, physical, logical, cyber persona, persona), Dieter Fleck and Terry D Gill, 'Military Cyber Operations' in Dieter Fleck and Terry D Gill (eds), *The Handbook of the International Law of Military Operations* (2nd edn, Oxford University Press 2015) 458.

[31] For the purposes of this paper I will apply the three-layer model as developed by Benkler and described by the Tallinn Manual 2.0.

[32] Ido Kilovaty, 'Cyber Warfare and the Jus Ad Bellum Challenges' (2014) 5 National Security Law Brief 91, 94.

[33] GGE Report 2013, para 20; GGE Report 2015, para 27.

C. The Tallinn Manual 2.0 approach to Sovereignty – the Primacy of Territorial Effects

The authors of the Tallinn Manual 2.0 seem to subscribe to the first view. They argue that 'the physical, logical, and social layers of cyberspace are encompassed in the principle of sovereignty'.[34] The most important feature is that 'cyber activities occur on territory and involve objects (…) over which States may exercise their sovereign prerogatives'.[35] In particular, even if cyber activities are conducted in such a way that they cross multiple borders, the acting individuals and entities remain subject to the jurisdiction of particular States.[36] In consequence, traditional notions of sovereignty are applied to conduct in cyberspace by way of a territorial analogy.[37] The primacy of territorial effects in the *Tallinn Manual 2.0* is best seen with regard to its approach to cloud computing. According to the *Manual*, operations against cloud infrastructure 'would generally not violate the sovereignty of other States that are affected by the operations unless the consequences that manifest in those States are of the requisite nature [*i.e.* with physical effects on the territory of the State – P.R.] as discussed in this Rule.'[38] Sovereignty over data stored abroad is rejected,[39] with an exception for government data under the 'inherently governmental functions' test.[40]

3. A 'LAYERED' APPROACH TO SOVEREIGNTY IN CYBERSPACE

A. Challenges to a Westphalian Understanding of Sovereignty in Cyberspace

Both the traditional understanding of sovereignty and recent State practice and *opinio iuris* are clear that sovereignty is primarily territorial. This means above all, as the *Tallinn Manual 2.0* points out, that States have the power to regulate ICT infrastructure, persons and activities located in their territory.[41] However, the *Tallinn Manual* underestimates the challenges to a territorial understanding of territoriality brought about by cloud computing, data partitionability and the mobility of ICT devices. The increasing use of cloud computing, understood as the 'storing by users of their infrastructure or content on remote servers',[42] allows companies and governments to move critical functions and services 'to the cloud' and run them from

[34] *Tallinn Manual 2.0*, 12.

[35] *Ibid*.

[36] *Ibid* 12–13.

[37] See, for example, for the rule prohibiting violations of territorial sovereignty: ibid 17; Wolff Heintschel von Heinegg, 'Legal Implications of Territorial Sovereignty in Cyberspace' in Chcosristian Czosseck, Katharina Ziolkowski and Rain Ottis (eds), *4th International Conference on Cyber Conflict* (2012); Michael N Schmitt and Liis Vihul, 'Respect for Sovereignty in Cyberspace' (2017) 95 Texas Law Review 1639.

[38] *Tallinn Manual 2.0*, 25.

[39] *Ibid* 16.

[40] *Ibid* 23.

[41] *Ibid* 14.

[42] Primavera De Filippi, Smari McCarthy, 'Cloud Computing: Centralization and Data Sovereignty' *European Journal for Law and Technology*, Vol. 3 No. 2, 2012.

ICT infrastructure usually grouped in large data centres located in a few key points around the globe,[43] often on the territory of another State. Due to a lack of technical restrictions for transborder data flows, data stored in the cloud can be partitioned, held in more than one location and moved between servers to reduce latency and facilitate access for customers. The move to the cloud regularly concerns communications and content data, but increasingly affects whole platforms and services in sectors such as banking[44] and even elements of critical infrastructure, such as remote terminal units, programmable logic controllers[45] or smart grid applications.[46]

If critical infrastructure such as industrial control applications or banking services, or governmental data and services, were to be stored in offshore data centres, the question arises as to the extent of each State's sovereignty. For instance, in case of a cyberattack against these data centres, would only the sovereignty of the State on whose territory the data centre is located be implicated, or would the de-territorialised sovereignty of the other State also be affected? Rather than conceptualising sovereignty in cyberspace exclusively by territoriality (in terms of location of ICT infrastructure), I would submit that there is emerging State practice to suggest that sovereignty in cyberspace may be understood as containing multiple spheres – or layers – of overlapping rights, responsibilities, and political authority.

1) Example 1: Asserting Jurisdiction Over Data Stored Abroad

Recent case law and legislation suggest that States treat remotely stored data and services as falling under their jurisdiction if they have a close connection to the territory of the regulating State. For instance, in *Google Spain*[47] the Court of Justice of the European Union (CJEU) held that the Data Protection Directive 95/46 grants an individual the right to request, under certain circumstances, that his or her personal data be no longer accessible through a search engine,[48] irrespective of the place where the actual data processing takes place, provided that the processing of personal data is carried out in the context of commercial activity on the territory of a Member State.[49]

In *Microsoft Ireland*, federal prosecutors sought and obtained a warrant for the search and seizure of information, including email, stored in a specified account hosted by Microsoft, to disclose the contents of e-mails of a suspect in an investigation related

[43] For the location of Amazon's data centres, see Richard Fox and Wei Hao, *Internet Infrastructure. Networking, Web Services and Cloud Computing* (CRC Press 2018) 475.

[44] Cary Springfield, 'The Impact of Cloud Computing on the Banking Sector' (*The International Banker*, 2018) <https://internationalbanker.com/banking/the-impact-of-cloud-computing-on-the-banking-sector/> [accessed 11.03.2019].

[45] Áine MacDermott and others, 'Hosting Critical Infrastructure Services in the Cloud Environment Considerations' (2015) 11 International Journal of Critical Infrastructures 365, 371.

[46] Bhaskar Prasad Rimal and Ian Lumb, 'The Rise of Cloud Computing in the Era of Emerging Networked Society' in Nick Antonopoulos and Lee Gillam (eds), *Cloud Computing. Principles, Systems and Applications* (2nd edn, Springer 2017) 14.

[47] CJEU, *Google v. Mario Costeja González*, Case C-131/12, Judgement of 13 May 2014.

[48] *Ibid*. para 98.

[49] *Ibid*. paras. 55-57.

to drug trafficking.[50] On appeal, the US Court of Appeals for the Second Circuit (CoA) reversed the Magistrate's order,[51] but lower courts in other Circuits did not join with the Court of Appeals for the Second Circuit and granted search warrants in cases relating to, among others, Yahoo and Google e-mail accounts.[52] The issue was resolved by the adoption of the Clarifying Lawful Overseas Use of Data (CLOUD) Act on 22 March 2018, which requires service providers subject to US jurisdiction to produce data under an SCA warrant regardless of the location of the server where the data is stored.[53]

In response to the CLOUD Act, the European Commission proposed a Regulation on European Production and Preservation Orders for electronic evidence in criminal matters (EPO Regulation).[54] While in its *amicus curiae* brief in the *Microsoft Ireland* case the Commission argued for an interpretation of domestic law 'mindful of the restrictions of international law and considerations of international comity' by giving due regard to the principle of territoriality,[55] it addressed the issue of transborder access to electronic evidence in much the same way as the United States in the CLOUD Act – by allowing access to data stored in a third State. In its explanatory summary the Commission clearly states that the draft Regulation deliberately 'moves away from data location as a determining factor, as data storage normally does not result in any control by the state on whose territory data is stored'.[56] This is so, because data is no longer stored locally but made available on cloud-based infrastructure that is accessible from anywhere and service providers use decentralised systems to store data in order to optimise load balancing, while also often copying content in several servers distributed globally to speed up content delivery.[57]

2) Example 2: Data Embassies and the De-territorialisation of Governmental Functions

The proliferation of cloud computing not only offers benefits to consumers and the private sector, but also opens opportunities for governments with respect to the performance of State functions. A quick survey shows that many State organs and

[50] US District Court (S.D. New York), *In Re Warrant to Search a Certain E-Mail Account*, 15 F.Supp.3d 466 (2014), 468.

[51] US Court of Appeals (2d Circuit), *Microsoft Corp. v. USA (In Re Search Warrant)*, 829 F.3d 197 (2016).

[52] US District Court, E.D. Pennsylvania, *In re Search Warrant No. 16-1061-M to Google*, 232 F. Supp. 3d 708 (2017); US District Court, E.D. Wisconsin, *In re: Information associated with one Yahoo email address that is stored at premises controlled by Yahoo, In re: Two email accounts stored at Google, Inc*., Case Nos. 17-M-1234, 17-M-1235, 21 Feb. 2017.

[53] Jean Galbraith, 'Congress Enacts the Clarifying Lawful Overseas Use of Data (CLOUD) Act, Reshaping U.S. Law Governing Cross-Border Access to Data' (2018) 112 American Journal of International Law 486, 487.

[54] European Commission, *Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters*, 17 Apr. 2018, Doc. COM(2018) 225 final [hereinafter: Draft EPO-Regulation].

[55] US Supreme Court, *United States v. Microsoft Corp*., No. 17-2, Brief of the European Commission on Behalf of the European Union as Amicus Curiae in Support of Neither Party, p. 6-7.

[56] Draft EPO-Regulation, Explanatory Memorandum, p. 13.

[57] *Ibid*. p. 14.

governmental agencies already employ cloud-based web services. For instance, the company Amazon offers hosting solutions and web-based applications to governmental customers which include, *inter alia*, the US Department of State, the Department of Homeland Security, the UK Justice Department, the Government of Singapore and Europol.[58]

An early example where, to increase resilience, certain governmental functions were temporarily performed from ICT infrastructure located in a third State occurred during the Russian attack on Georgia in 2008, when a US internet service provider hosted the website of the Georgian President to better protect it against defacement and DDoS attacks.[59] However, maybe the most prominent example so far of moving certain State functions into the cloud is the Estonian 'data embassy' in Luxembourg.[60] Based on an agreement with the Grand Duchy of Luxembourg, Estonia acquired dedicated data centre space in Luxembourg for the purpose of hosting Estonian data and information systems.[61] Inspired by the Vienna Convention on Diplomatic Relations,[62] the agreement grants data stored in the data centre the status of archives and declares them inviolable, thus exempt from search, requisition, attachment or execution.[63] It further stipulates that assets used for the storage of data and information systems enjoy sovereign immunity.[64] While Estonia and Luxembourg found a treaty solution to the storage of governmental data abroad, even without a treaty one can argue that international law contains mechanisms 'that support the extension of a sovereign's right to inviolability of its data to the internet and cloud storage'.[65] Examples such as these seem to suggest that States might regard governmental data stored abroad as covered by their sovereignty, even though it is not stored on their territory. While no examples of cyberattacks against data embassies are known as of today, I would suggest that a cyberattack crossing the threshold of sufficient harm might indeed be regarded as a violation of the sovereignty of a State, because the State might regard the attack as infringing its exclusive authority.

[58] Amazon, *Government, Education, and Nonprofits Case Studies*, <https://aws.amazon.com/solutions/case-studies/government-education/all-government-education-nonprofit/?nc1=f_ls> [accessed 11.03.2019].

[59] Jason Healey, 'When "Not My Problem" Isn't Enough: Political Neutrality and National Responsibility in Cyber Conflict' in Christian Czosseck, Rain Ottis and Katharina Ziolkowski (eds), *2012 4th International Conference on Cyber Conflict. Proceedings* (NATO CCD COE 2012) 24.

[60] E-Estonia, 'Estonia to open the world's first data embassy in Luxembourg', < https://e-estonia.com/estonia-to-open-the-worlds-first-data-embassy-in-luxembourg/> [accessed 11.03.2019].

[61] Loi du 1er décembre 2017 portant approbation du 'Agreement between the Grand Duchy of Luxembourg and the Republic of Estonia on the hosting of data and information systems', signé à Luxembourg, le 20 juin 2017, Annex, Doc. parl. 7185, [hereinafter: Data Embassy Agreement] <http://legilux.public.lu/eli/etat/leg/loi/2017/12/01/a1029/jo> [accessed 11.03.2019].

[62] Bartłomiej Sierzputowski, 'The Data Embassy Under Public International Law' (2019) 68 International and Comparative Law Quarterly 225, 234.

[63] Art. 6(2) Data Embassy Agreement.

[64] Art. 5 Data Embassy Agreement.

[65] Estonian Ministry of Economic Affairs and Communications, Microsoft Corp., 'Implementation of the Virtual Data Embassy Solution', <https://www.mkm.ee/sites/default/files/implementation_of_the_virtual_data_embassy_solution_summary_report.pdf> 14 [accessed 11.03.2019].

B. Layers of Sovereignty and the Criterion of Proximity

In the examples cited above, as well as in similar cases, we see a separation between the territory where the data is stored and the authority over the data. While the host State has jurisdiction over infrastructure and data located in it, the data usually does not affect its territory and therefore, as the EU Commission pointed out, it does not have an interest in regulating it.[66] The interest lies with the State on whose territory the services are offered and/or the users are located. This creates concurrent jurisdictions: one based on the principle of territoriality of the ICT infrastructure storing the data, the other on the territorial availability of the offered services and the nationality or domicile of the data owner. I would therefore argue that, similarly to the Law of the Sea,[67] we might conceptualise cyberspace as consisting of different zones – or layers – of decreasing sovereignty, depending on the proximity to the sphere of exclusive authority, which forms the core of sovereignty.

The criterion of proximity should not be thought of in geographical terms; rather, it is the degree of connectedness of the data to the sphere of exclusive State authority. Similar to the criterion of a 'genuine connection' in *Nottebohm and Barcelona Traction*,[68] used to determine whether a State can assert extraterritorial jurisdiction,[69] it describes the degree of the link between the data or service stored abroad and the State. Proximity therefore does not establish an absolute test, but rather a relative one, depending on the concrete situation and the interests of the States involved. The following criteria established in cases relating to the extraterritorial access to data,[70] factors to determine proximity might include in cases of overlapping sovereignty claims: the degree to which the territory of a particular State is affected, the interests of the affected States, the location and nationality of the data owner, the principal territory the data is accessed from and targeted at, and in case of services the nature and extent of the service provider's ties to the particular State.

C. Mapping Layers of Sovereignty on the Layers of Cyberspace

Based on the criterion of proximity, several layers of sovereignty can be distinguished.

1) Baseline Sovereignty – Exclusive Authority of the Territorial State over ICT Components of the Physical Layer

With regard to the physical layer of cyberspace, the proximity to the State is absolute through the criterion of territory. This reflects the international consensus on the applicability of international law in cyberspace, established by the UN Group of Governmental Experts in its 2013 and 2015 Reports, which found that State

[66] Draft EPO-Regulation, Explanatory Memorandum, p. 13.

[67] Jon D Carlson and others, 'Scramble for the Arctic: Layered Sovereignty, UNCLOS, and Competing Maritime Territorial Claims' (2013) 33 SAIS Review of International Affairs 21, 23.

[68] ICJ, *Nottebohm* (Liechtenstein v Guatemala), Judgment, (1955) ICJ Rep. 4 et seq; ICJ, *Barcelona Traction* (Belgium v Spain), (1970) ICJ Rep. 42.

[69] Cedric Ryngaert, *Jurisdiction in International Law* (2nd edn, Oxford University Press 2015) 156.

[70] Compare CLOUD Act, 18 U.S.C. §2703(3)(A)-(H).

sovereignty and rules of jurisdiction apply to ICT infrastructure located within State territory.[71] There is agreement on this point between most States, even those with otherwise differing views on cyberspace sovereignty such as the US[72] and China.[73] States regularly assert jurisdiction over components of the physical layer, for instance imposing regulatory standards or security requirements.[74] State authority over the physical layer components located on its territory is exclusive insofar as no other State is permitted under international law to prescribe and enforce rules regarding objects located within the territory of another State.[75] It may, however, be limited by international law if the exercise of exclusive authority over ICT infrastructure would cause harm to other States. If, for instance, States harbouring large Internet Exchange Points such as DE-CIX in Frankfurt or AMS-IX in Amsterdam were to exercise their authority to shut down these exchange points with the effect of disrupting internet traffic in neighbouring States, one might argue that this would violate the obligation not to knowingly harm the rights of other States,[76] as confirmed by the ICJ in *Corfu Channel*.[77]

2) Limited Authority over the Logical Layer

While the physical layer of cyberspace consists of ICT components and can thus be described in territorial terms, the logical layer, which consists of the codes and standards that drive physical network components and make communication and exchange of information between them possible,[78] is fundamentally a-territorial. Nevertheless, it is not free from considerations of sovereignty. The governance and allocation of critical resources making up the public core of the internet[79] – such as the allocation of IP addresses, domain names and the administration of root DNS servers – raises questions as to the extent of State authority over these functions. At present, these functions are being performed by the Internet Corporation for Assigned Names

[71] GGE Report 2013, para 20; GGE Report 2015, para 27.

[72] See Harold Hongju Koh, 'International Law in Cyberspace' (2012) 54 Harvard International Law Journal 1, 6; Brian Egan, 'Remarks on International Law and Stability in Cyberspace' (*Berkeley Law School, California November 10, 2016*).

[73] People's Republic of China, 'International Strategy of Cooperation on Cyberspace', Chapter II Principle 2, <http://www.xinhuanet.com//english/china/2017-03/01/c_136094371_2.htm> [accessed 11.03.2019].

[74] Compare e.g. 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, OJ L 194, 19.7.2016, p. 1–30.

[75] '[T]he first and foremost restriction imposed by international law upon a State is that, failing the existence of a permissive rule to the contrary, it may not exercise its power in any form in the territory of another State', PCIJ, *S.S. Lotus* (France v. Turkey.), Judgment, 1927 PCIJ Ser. A No. 10, at p. 18.

[76] On the no-harm rule in cyberspace see Katharina Ziolkowski, 'General Principles of International Law as Applicable in Cyberspace' in Katharina Ziolkowski (ed), *Peacetime Regime for State Activities in Cyberspace* (NATO CCD COE Publications 2013) 165.

[77] ICJ, *Corfu Channel Case* (United Kingdom v. Albania), Judgment, 1949 ICJ Rep. 4, 35.

[78] Joint Chiefs of Staff, 'Cyberspace Operations, JP 3-12' (2018) <https://fas.org/irp/doddir/dod/jp3_12.pdf> I-3 [accessed 11.03.2019].

[79] Dennis Broeders, 'Aligning the International Protection of "the Public Core of the Internet" with State Sovereignty and National Security' (2017) 2 Journal of Cyber Policy 366, 6 <https://doi.org/10.1080/23738871.2017.1403640>.

and Numbers (ICANN)[80] in a multi-stakeholder model of industry self-regulation.[81] Insofar as the US have transitioned control over key IANA functions to the global multi-stakeholder community, the authority of any State over the logical layer is limited to its role as one of the stakeholders. Under the current model, no State alone has sovereignty over the logical layer. However, States such as China and Russia fear that they do not have sufficient authority over core functions of those portions of globally connected networks located on their territory. It is for this reason that both China and Russia have made gaining control over internet governance a key part of their cyberspace strategies and included this principle as a key element of their definition of cyberspace sovereignty.[82] To this end, the Russian parliament has recently passed a bill aimed at creating a domestic Domain Name System, in order to be able to disconnect the Russian internet from the global internet exchange system.[83] Should Chinese and Russian efforts to replace the multi-stakeholder model with a multilateral model under the International Telecommunications Union[84] succeed, or should States choose to take over control over DNS servers and registries serving their territories, sovereignty over the elements of the logical layer necessary to run national networks would be restored.

3) Concurrent Sovereignty over Data Located on ICT Infrastructure in Another State

In cases concerning the sovereignty over data and services stored in the ICT infrastructure located in one State and offered in the territory of another State, it is appropriate to speak of concurrent sovereignty under the proposed model of 'layered sovereignty'. By virtue of the ICT infrastructure's location, the host State has a baseline sovereignty over the ICT infrastructure. However, concurrent sovereignty exists if the data stored within the ICT infrastructure is sufficiently proximate to the State asserting sovereignty. For instance, in the case of governmental data stored in data embassies, the layered model of sovereignty would permit two layers of sovereignty to exist: one of the territorial State over the ICT infrastructure, that is the physical layer, and another of the data holder State over the data, that is the logical (content) layer.

D. Practical Application

What, then, is the practical application of this theoretical model? In my view, there are two areas where a 'layered' conception of sovereignty might be useful. *First*, it would

[80] On the role of the Internet Corporation for Assigned Names and Numbers (ICANN) see Scott J Shackelford, 'Defining the Cyber Threat in Internet Governance', *Managing Cyber Attacks in International Law, Business, and Relations* (Cambridge University Press 2014) 20.

[81] Kal Raustiala, 'Governing the Internet' (2016) 110 American Journal of International Law 491, 501.

[82] Sarah McKune and Shazeda Ahmed, 'The Contestation and Shaping of Cyber Norms Through China's Internet Sovereignty Agenda' (2018) 12 International Journal of Communication 21, 3839.

[83] Katherine Landes, 'The "Iron Curtain" Is Close to Falling over the Russian Internet' (*International Policy Digest*, 2019) <https://intpolicydigest.org/2019/03/02/the-iron-curtain-is-close-to-falling-over-the-russian-internet/> [accessed 11.03.2019].

[84] Adam Segal, 'Holding the Multistakeholder Line at the ITU' *Council on Foreign Relations Blog (2014)*, <https://www.cfr.org/report/holding-multistakeholder-line-itu> [accessed 11.03.2019].

allow the allocation of sovereignty over data stored or a service offered from abroad, provided there is sufficient proximity between the data/service and the State asserting jurisdiction. Should this data/service fall victim to a cyberattack, such an attack might be qualified as a violation of sovereignty of the attacked State irrespective of the fact that the territory of that State has not been affected. This is because such a State might have an overwhelming interest in asserting authority over the data in question, for example if it is government data (in the case of data embassies) or if the attacked service is considered as critical infrastructure, is controlling critical infrastructure within the territory of that State or is otherwise of significant importance for essential interests of that State (e.g. banking services). In these cases, the State whose remotely stored data was attacked could resort to countermeasures or the plea of necessity to counter the action in question, irrespective of the rights of the territorial State, whose sovereignty over the ICT infrastructure might also be affected. *Secondly*, the criterion of proximity might be a useful tool to assess the proportionality of countermeasures or the existence of an essential interest of a State which has been affected through the cyberattack. The greater the proximity of the attacked data to the State, the greater its essential interest in protecting it against violations of sovereignty.

4. CONCLUSION

In conclusion, in the post-Westphalian system, geography – 'the physical space of a State' – is at the very core of the concept of sovereignty.[85] However, the advance of modern technology in the 20th and 21st centuries and especially the emergence of cyberspace, with its transboundary, geography-defying quality,[86] have led to a steady decline of the function of territory to exclude the activities of other entities within the boundaries of a State.[87] Therefore a strict application of traditional rules flowing from the principle of sovereignty, especially the rule of territorial sovereignty, would overemphasise the notion of territoriality and disregard the practical challenges to state authority emanating from cyberspace, leading to an imbalance in the rights and obligations of States in favour of the State on whose territory ICT infrastructure is located. A model of layered sovereignty, while at present a proposal *de lege ferenda*, would restore the balance between rights and obligations by adjusting for overlapping rights, responsibilities, and political authority in cyberspace.

[85] Bethlehem (n 17) 14.

[86] *Ibid.* 18.

[87] Shaw (n 12) 64–65.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

The Sound of Silence: International Law and the Governance of Peacetime Cyber Operations

Barrie Sander
Postdoctoral Fellow
School of International Relations
Fundação Getulio Vargas (FGV)
Sâo Paulo, Brazil
barrie.sander@graduateinstitute.ch

Abstract: In an age of cyber insecurity, anxieties about the silence of States concerning the applicability of international law to peacetime cyber operations have been growing. Concerns have focused on the reluctance of States to agree cyber-specific multilateral treaties and to publicly clarify the customary international rules applicable to hostile cyber operations. Taking these concerns as its point of departure, this paper argues for greater specificity in evaluating the silence of States in the cyber context by distinguishing between three distinct types of peacetime security threats: cyber attacks, cyber espionage, and cyber information operations. Cyber attacks and cyber espionage are technical security threats which involve breaking into and targeting information and communications technologies. The primary distinction between the two is in the nature of the payload to be executed; while a cyber attack's payload is destructive, a cyber espionage payload acquires information non-destructively. Cyber information operations are content-based security threats which involve harnessing the power of online information to cognitively target human intelligence. Relying on this typology, this paper highlights how State silences concerning the application of international law to peacetime cyber operations are not uniform, but vary in terms of their targets, scope and rationale depending on the particular security threat under examination. It is suggested that these variations not only reveal an important dimension of the politics of international law, but are also salient to how the silence of States in different cyber contexts may be evaluated. Contrary to the tendency to automatically cast State silences in a negative light, this paper reveals that silences can perform different and sometimes constructive functions that are yet to be fully acknowledged or appreciated.

Keywords: *State silence, peacetime cyber operations, international law, cyber attacks, cyber espionage, cyber information operations*

1. INTRODUCTION

In an age of cyber insecurity, international lawyers have been grappling with the challenge of identifying the extent to which international law applies to cyber operations.[1] The engagement of international lawyers with this question has evolved over time. Following the notorious cyber attack on Estonia in 2007, international lawyers were initially preoccupied by the prospect of cyber war – a concern reflected in the narrow focus of the first edition of the *Tallinn Manual*, which fastened its gaze on the law governing the use of force (*jus ad bellum*) and the law of armed conflict (*jus in bello*).[2] Over the course of the past decade, however, it has become increasingly apparent that the vast majority of hostile cyber operations neither cross the threshold required to constitute a prohibited use of force nor occur in the context of existing armed conflicts. In line with this realisation, the focus of international lawyers has gradually shifted towards a concern for interpreting the international legal rules applicable to so-called "below the threshold" peacetime cyber operations – an interest reflected in the expanded mandate of the second edition of the *Tallinn Manual*.[3]

Yet, for all the interpretive efforts of international lawyers, recent years have also witnessed growing concerns about the silence of States concerning the applicability of

[*] The author would like to thank Duncan Hollis, Russell Buchan, Kubo Mačák, Asaf Lubin, and the anonymous reviewers for their comments on earlier drafts of this paper. The author would also like to acknowledge the funding of Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), which enabled this research to be conducted. All errors remain the author's own.

[1] Barrie Sander, 'Cyber Insecurity and the Politics of International Law', *ESIL Reflections* (2017).

[2] *Tallinn Manual on the International Law Applicable to Cyber Warfare* (CUP, 2013).

[3] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (CUP, 2017) (*'Tallinn Manual 2.0'*). On the shift of focus to peacetime cyber operations, see generally, Kubo Mačák, 'From the Vanishing Point Back to the Core: The Impact of the Development of the Cyber Law of War on General International Law', in Henry Rõigas et al. (eds), *Defending the Core* (NATO CCD COE, 2017) 135.

international law to peacetime cyber operations.[4] According to conventional wisdom, this silence has manifested itself in a number of forms.

First, States have appeared resistant to agreeing cyber-specific multilateral treaties, a trend exemplified by the struggle of Microsoft to garner widespread support for its proposed Digital Geneva Convention. *Second*, States have been reluctant to publicly clarify the customary international rules applicable to peacetime cyber operations, a trend recently characterised by Dan Efrony and Yuval Shany as amounting to "a policy of silence and ambiguity" that is designed to preserve high levels of operational flexibility within the cyber domain.[5] This "wait and see" approach to cyber regulation recently came to the fore in the latest round of talks within the UN Group of Governmental Experts (GGE), which failed to agree a consensus report on the voluntary and binding norms applicable to cyber operations.[6] While recent developments – including the Paris Call for Trust and Security in Cyberspace[7] and the adoption of two resolutions by the UN General Assembly's first committee establishing an open-ended working group on cyber norms and a new UN GGE[8] – have demonstrated a willingness to continue the conversation, it remains to be seen how far States are able to achieve consensus beyond vague assertions about the applicability of international law to cyber operations.[9]

[4] See, for example, Brian J. Egan, 'International Law and Stability in Cyberspace', (2017) 35 *Berkeley Journal of International Law* 169, 172 ("States' relative silence could lead to unpredictability in the cyber realm"); Kubo Mačák, 'From Cyber Norms to Cyber Rules: Re-engaging States as Law-makers', (2017) 30 *Leiden Journal of International Law* 877 , 888 ("faced with states' silence, non-state actors have moved into the vacated norm-creating territory previously occupied exclusively by states"); and Dan Efrony and Yuval Shany, 'A Rule Book on The Shelf? *Tallinn Manual 2.0* on Cyber Operations and Subsequent State Practice', (2018) 112 *American Journal of International Law* 583, 648 (arguing that "a significant normative gap exists in relation to the regulation of interstate cyberoperations" because of "the combination of silence and ambiguity in state practice and their reluctance to articulate their official policy in cyberspace"). See, however, Nicholas Tsagourias, 'The Slow Process of Normativizing Cyberspace', (2019) 113 *AJIL Unbound* 71, 73-74 (arguing that the slow pace by which States are "translating overbroad principles of international law into rules and practice and […] translating practice into rules and principles […] is not peculiar to cyberspace" and, as such, "there is no reason to despair").

[5] Efrony and Shany, *supra* n.4, 588. See also Fleur Johns, 'War Without Words', (2019) 113 *AJIL Unbound* 67, 68 (observing how, according to Efrony and Shany, "[l]aw flows from language and its advance stalls in the quiet" and "international law's capacity to curtail or condition the exercise of military power, economic might, and tangible or intangible violence in the cyber domain is presumed to depend upon its capacity to saturate the vocabularies of those with means to deploy such power and to do so in visible, recordable ways").

[6] 'Dispute along cold war lines led to collapse of UN cyberwarfare talks', *The Guardian*, 23 August 2017.

[7] Arthur P.B. Laudrain, 'Avoiding a World War Web: The Paris Call for Trust and Security in Cyberspace', *Lawfare*, 4 December 2018.

[8] Alex Grigsby, 'The United Nations Doubles Its Workload on Cyber Norms, and Not Everyone Is Pleased', *Council on Foreign Relations*, 15 November 2018.

[9] The applicability of international law to cyber operations was famously confirmed by the UN GGE in its 2013 consensus report. A degree of progress was made in the UN GGE's 2015 report, which began to articulate binding international legal norms applicable in cyberspace. See generally, 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', U.N.Doc. A/68/98, 24 June 2013; and 'Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security', U.N.Doc. A/70/174, 22 July 2015 ('UN GGE 2015 Report').

Beyond the reluctance of States to engage meaningfully in the construction and clarification of the international legal rules applicable to cyber operations, a *third* development has been the growing tendency of States to embrace the language of non-binding voluntary norms to articulate responsible behaviour in cyberspace. This trend has been particularly visible in the work of the UN GGE, whose 2015 report recommended 11 norms for consideration by States.[10] According to Kubo Mačák, the emergence of a parallel track to develop voluntary norms of responsible State behaviour in cyberspace signifies "a trend of moving away from the creation of legal rules of international law in the classical sense".[11] *Finally*, the silence of States concerning international law in the cyber context has also been visible in their growing tendency to publicly attribute hostile cyber operations to other States without making reference to applicable international legal rules. In other words, while States have proven increasingly open to naming the involvement of other States in hostile cyber operations, they have often studiously avoided shaming them through recourse to the language of international law.[12]

While this account of the relationship between States, international law and peacetime cyber operations is not inaccurate, it is nonetheless incomplete. Taking this account as its point of departure, this paper argues for greater specificity in examining the silences of States concerning the relationship between international law and peacetime cyber operations. To this end, this paper distinguishes between *three distinct types of peacetime security threats* that have arisen in the cyber domain: cyber attacks, cyber espionage, and cyber information operations. *Cyber attacks* and *cyber espionage* are technical security threats which involve breaking into and targeting information and communications technologies. The primary distinction between the two is in the nature of the payload to be executed: while a cyber attack's payload is destructive, a cyber espionage payload acquires information non-destructively.[13] In contrast to these technical security threats, *cyber information operations* are content-based security threats which involve harnessing the power of online information to cognitively target human intelligence.[14] Although, in practice, a particular cyber operation may encompass more than one type of security threat, this typology offers a useful lens for examining the silences of States within the cyber domain.

[10] UN GGE 2015 Report, supra n.9, para. 13. For commentary, see Eneken Tikk (ed.), *Voluntary, Non-Binding Norms for Responsible State Behaviour in the Use of Information and Communications Technology: A Commentary* (UN Office for Disarmament Affairs, 2017). The Global Commission on the Stability of Cyberspace (GCSC) has also been conducting important work on global cybersecurity norms. See, for example, GCSC, 'Call to Protect the Public Core of the Internet', November 2017; GCSC, 'Call to Protect the Electoral Infrastructure', May 2018; and GCSC, 'Norm Package Singapore', November 2018.

[11] Mačák, *supra* n.4, 882.

[12] Martha Finnemore and Duncan B. Hollis, 'Beyond Naming and Shaming: Accusations and International Law in Cybersecurity', *SSRN*, 6 March 2019.

[13] Herbert Lin, 'Responding to Sub-Threshold Cyber Intrusions: A Fertile Topic for Research and Discussion', (2011) 12 *Georgetown Journal of International Affairs* 127, 129-130.

[14] Leonhard Kreuzer, 'Disentangling the cyber security debate', *Völkerrechtsblog*, 20 June 2018.

Relying on the typology, this paper reveals how State silences concerning the application of international law to peacetime cyber operations are not uniform, but vary in terms of their targets, scope, and rationale depending on the particular security threat under examination. In terms of *targets*, silences may pertain to the identification of hostile cyber operations, the existence and contours of international legal rules, issues of attribution, or measures adopted in response to cyber operations. In terms of *scope*, silences may be more prevalent amongst particular groupings of States or concerning the applicability of particular types of international legal norms. In terms of *rationale*, silences may be motivated by a range of concerns, including technical attribution challenges, geopolitical sensitivities, a desire for operational flexibility in cyberspace, or averting the risk of legitimising the repressive practices of other States.

The paper concludes that these variations not only reflect an important dimension of the politics of international law, but are also salient to how State silences in different cyber contexts may be evaluated. Contrary to the tendency to automatically cast State silences in a negative light, this paper reveals that silences can perform different and sometimes constructive functions that are yet to be fully acknowledged or appreciated.

2. PEACETIME CYBER ATTACKS

Peacetime cyber attacks are destructive cyber operations, encompassing acts undertaken by a State – or actors whose conduct is attributable to a State under international law – that uses cyber capabilities to alter, disrupt, degrade or destroy the computer systems or networks of a foreign State, or the information or programs resident in those systems or networks, which fall below the threshold required to constitute a prohibited use of force and occur outside the context of an armed conflict.[15] To the extent that information concerning peacetime cyber attacks has entered the public domain,[16] at least four types of silences are identifiable in the reactions of victim States.

First, victim States have sometimes been silent as to whether a particular incident resulted from an accident or a cyber attack. Kristen Eichensehr has referred to this type of silence as pertaining to the "what" attribution question, which involves determining what caused a particular incident.[17] For instance, when its centrifuges began spinning out control in 2008, it was not immediately apparent to Iran that its nuclear facilities had been subject to a cyber attack by the Stuxnet worm rather than failures of their own internal operating teams.[18]

[15] This definition draws on Lin, *supra* n.13, 129.

[16] On the limits of available open-source material that reveals both the existence of hostile cyber operations and State responses to them, see Efrony and Shany, *supra* n.4, 594-595 and 631-632.

[17] Kristen Eichensehr, 'Cyber Attribution Problems – Not Just Who, But What', *Just Security*, 11 December 2014.

[18] David E. Sanger, 'Obama Order Sped Up Wave of Cyberattacks Against Iran', *The New York Times*, 1 June 2012.

Second, in other instances, victim States have refrained from taking a public position in response to particular cyber attacks, remaining silent both in terms of whether attacks may be attributed to other States, as well as whether any response measures have been adopted. The series of cyber attacks involving the so-called Shamoon malware offers a clear illustration of this approach.[19] This malware was deployed against a range of Saudi Arabian and Qatari private and public sector targets between 2012 and 2017, resulting in the erasure of data from the hard drives of infected computers and significant network shutdowns. Yet, despite suspicions that the attacks were sponsored by Iran, to date the Shamoon operations have not been publicly attributed by Saudi Arabia or Qatar to any State or State-sponsored group, nor have there been any official non-covert operations in response.[20]

Third, in some contexts, victim States have responded by publicly attributing cyber attacks to other States whilst remaining silent about whether international law is applicable to the situation. A clear example of this approach may be found in the response of the US to the 2014 cyber attack against Sony Pictures Entertainment.[21] Conducted by a hacking group calling itself "Guardians of the Peace", the Sony operation involved, *inter alia*, the deployment of destructive malware which caused tens of millions of dollars of damage to Sony's computer infrastructure. In response, the US publicly attributed the cyber attack to North Korea and imposed a series of sanctions on ten individuals and three entities associated with the North Korean regime. In addition, the shutdown of North Korea's Internet network on Christmas Eve of 2014 is widely believed to have been a covert US response to the Sony hack. Yet, in terms of international law, US Secretary of State John Kerry was only willing to characterise the cyber attack as an operation that demonstrated North Korea's "flagrant disregard for international *norms*",[22] while US President Obama referred to the incident as "an act of cyber vandalism", a phrase without a clear legal connotation.[23]

A similar approach was adopted in response to the WannaCry cyber attack.[24] In 2017, WannaCry affected hundreds of thousands of computers across at least 150 States around the world. The WannaCry malware prevented Microsoft's Windows operating system from booting and encrypted all data stored on affected computers. In October

[19] See generally, Efrony and Shany, *supra* n.4, 620-624.

[20] Ibid., 623-624.

[21] See generally, ibid., 605-609; Clare Sullivan, 'The 2014 Sony Hack and the Role of International Law', (2016) 8 *Journal of National Security Law & Policy* 437; and Michael Schmitt, 'International Law and Cyber Attacks: Sony v. North Korea', *Just Security*, 17 December 2014.

[22] 'Condemning Cyber-Attacks by North Korea', *US Department of State Press Release*, 19 December 2014 (emphasis added).

[23] 'US may put North Korea back on state terror list after Sony 'cybervandalism'', *The Guardian*, 21 December 2014.

[24] See generally, Efrony and Shany, *supra* n.4, 626-628; Michael Schmitt and Sean Fahey, 'WannaCry and the International Law of Cyberspace', *Just Security*, 22 December 2017; Michael J. Adams and Megan Reiss, 'How Should International Law Treat Cyberattacks like WannaCry', *Lawfare*, 22 December 2017; Jack Goldsmith, 'The Strange WannaCry Attribution', *Lawfare*, 21 December 2017; and Kristen Eichensehr, 'Three Questions on the WannaCry Attribution to North Korea', *Just Security*, 20 December 2017.

2017, the UK publicly attributed the cyber attack to North Korea,[25] an assessment that was endorsed by Microsoft's President and Chief Legal Officer, Brad Smith.[26] In December 2017, US Homeland Security advisor Tom Bossert also publicly attributed WannaCry to North Korea, an assessment that was endorsed by several cybersecurity firms and five other States: the UK, Canada, Japan, Australia, and New Zealand.[27] Again, however, the vocabulary of international law was conspicuous by its absence. Bossert, for example, neglected to mention international law or to identify any particular response measures being taken against North Korea.[28] Similarly, while the UK Foreign Office Minister for Cyber, Lord Ahmad, confirmed that "international law applies online as it does offline", he stopped short of determining whether WannaCry itself violated international law.[29]

Finally, on at least one occasion States have responded to a pattern of hostile cyber operations – some of which amounted to cyber attacks – by publicly attributing them to another State and confirming that the pattern of operations constituted a violation of international law, whilst remaining silent as to which norms of international law in particular were violated. Specifically, in October 2018, the UK and its allies exposed a series of cyber operations conducted by the Russian military intelligence service against political institutions, businesses, media outlets, and an international sports agency.[30] Some of the operations amounted to cyber attacks, including, for example, a destructive cyber operation that targeted the Ukrainian finance, energy, and government sectors but which ultimately spread and affected other European businesses.[31] According to statements released by a number of States, this series of hostile Russian cyber operations violated both international law and non-binding norms of responsible behaviour in cyberspace. The UK's National Cyber Security Centre, for example, condemned the Russian campaign of cyber operations as a "flagrant violation of international law", while UK Foreign Secretary Jeremy Hunt claimed that "this pattern of behaviour demonstrates [Russia's] desire to operate without regard to international law or established norms and to do so with a feeling

[25] 'British security minister says North Korea was behind WannaCry hack on NHS', *The Independent*, 27 October 2017.

[26] 'North Korean government behind NHS cyber attack, says Microsoft boss', *ITV News*, 13 October 2017.

[27] 'Press Briefing on the Attribution of the WannaCry Malware Attack to North Korea', *White House Press Briefings*, 19 December 2017.

[28] Ibid.

[29] Foreign and Commonwealth Office and Lord Ahmad of Wimbledon, 'Foreign Office Minister Condemns North Korean Actor for WannaCry Attacks', *Press Release*, 19 December 2017.

[30] National Cyber Security Centre, 'Reckless campaign of cyber attacks by Russian military intelligence service exposed' *Press Release*, 4 October 2018. For commentary, see generally, Jeffrey Biller and Michael Schmitt, 'Un-caging the Bear? A Case Study in Cyber Opinio Juris and Unintended Consequences', *EJIL: Talk!*, 24 October 2018.

[31] National Cyber Security Centre, *supra* n.30.

of impunity and without consequences".[32] Although these statements made clear the UK's position that international law had been violated, they were nonetheless vague in three respects: first, they failed to distinguish between the different cyber operations attributed to Russia – merely noting that "the pattern" of cyber operations was in violation of international law and established norms; second, the statements failed to distinguish which cyber operations violated international law and which merely transgressed voluntary norms of responsible behaviour in cyberspace; and finally, the statements failed to specify which international laws and established norms in particular had been violated.

A range of reasons may explain these different forms of State silence in response to cyber attacks. State reticence to publicly identify, attribute or respond to cyber attacks may simply be a result of insufficient evidence either concerning the existence of an attack or concerning the attribution of the cyber operation to a suspected State. The challenges of attribution in the cyber domain are well documented, requiring *technical* attribution to identify the location and identity of the cyber infrastructure from which an operation originates, *political* attribution to identify the person behind the infrastructure, and *legal* attribution to identify a sufficient legal nexus between the persons behind the operation and a State. The complexity of attribution in the cyber context is compounded by a variety of factors, including the ability for cyber operations to be routed through multiple computer networks in different States and the use of "anti-attribution" mechanisms to hide the provenance of cyber operations.[33]

Even when attribution *is* possible, national security concerns may lead victim States to opt for silence; for example, to prevent their adversaries from finding out that they have been detected or to reduce the risks associated with publicly exposing the victim State's vulnerabilities and technological capabilities. In addition, victim States may have geopolitical interests in remaining silent in the face of a cyber attack, including reducing the risk of escalation or ensuring that ongoing diplomatic efforts with particular States in related issue areas are not negatively affected.[34] A lack of effective response measures may also motivate State silence in this context. As Jack Goldsmith and Stuart Russell explain: "Unless a nation is able to effectively redress a cyber

[32] Ibid. For similar statements, see 'Joint statement by Presidents Tusk and Juncker and High Representative Mogherini on Russian cyber attacks', Council of the EU, *Press Release*, 4 October 2018 ("We deplore such actions, which undermine international law and international institutions"); 'Attribution of a Pattern of Malicious Cyber Activity to Russia', Prime Minister of Australia, Minister for Foreign Affairs of Australia, *Media Release*, 4 October 2018 ("contrary to the consensus on international law and norms"); 'Canada identifies malicious cyber-activity by Russia', Global Affairs Canada, 4 October 2018 ("demonstrate a disregard for international law and undermine the rules-based international order"); 'Netherlands Defence Intelligence and Security Service disrupts Russian cyber operation targeting OPCW', Ministry of Defence of the Netherlands, 4 October 2018 ("undermine the international rule of law").

[33] Nicholas Tsagourias, 'Cyber Attacks, Self-Defence and the Problem of Attribution', (2012) 17 *Journal of Conflict & Security Law* 229, 234.

[34] Efrony and Shany, *supra* n.4, 632-637.

intrusion, it can be harmful or self-defeating to publicize it, since public knowledge of loss and the failure to respond effectively invite more attacks".[35]

The reluctance of States to confirm whether or how international law applies in the context of particular cyber attacks likely stems from additional factors. In particular, State silence in this context may reflect doubts about the adequacy and adaptability of the international legal framework to the cyber domain. In terms of *adequacy*, the limitations of self-help remedies available to victim States under the law of State responsibility, including the notification and proportionality conditions of countermeasures, may lead some States to conclude that there is little added utility in invoking international law in the cyber domain.[36] In terms of *adaptability*, State silence may reflect a lack of consensus within a particular government or disagreements *between* governments trying to formulate a coordinated response to a particular cyber attack over whether there has been an international legal violation and, if so, which norm of international law has been violated.[37] In the latter regard, it is entirely plausible that States may prefer to adopt a "wait and see" approach before publicly clarifying their international legal position, particularly given the rapidly-changing technological landscape in which cyber attacks are launched.

Differences in the technical capabilities of States to reliably attribute hostile cyber operations may also underpin the silence of certain States concerning the applicability and contours of international law in the context of cyber attacks. Evaluating the reasons behind the opposition of certain States to the applicability of countermeasures in the cyber context at the most recent round of UN GGE talks, Michael Schmitt and Liis Vihul point to the operational reality that "some States, such as Cuba, lack the technical wherewithal of more advanced States to reliably attribute hostile cyber operations and therefore will be less able to establish the necessary basis for resorting to […] countermeasures".[38] Similar concerns may conceivably underpin the reticence of certain States to have recourse to international law more generally when responding to cyber attacks.

Finally, State silence concerning which specific international legal norms have been violated by a given cyber attack may also stem from the conflicting internal interests of powerful States concerning how permissive they believe the international legal framework applicable in cyberspace should be. As Kubo Mačák explains, since powerful States are also some of the most vulnerable to hostile cyber operations, such

[35] Jack Goldsmith and Stuart Russell, 'Strengths Become Vulnerabilities: How a Digital World Disadvantages the United States in its International Relations', *Aegis Series Paper No. 1806* (Hoover Institution, 2018), 13. See similarly, Eichensehr, *supra* n.24 ("[A]nother possibility is that states do agree that WannaCry violated international law, but are making a policy choice not to call North Korea's actions a legal violation in order to avoid creating public expectations about the need for governments to respond").

[36] Efrony and Shany, *supra* n.4, 651.

[37] Eichensehr, *supra* n.24.

[38] Michael Schmitt and Liis Vihul, 'International Cyber Law Politicized: The UN GGE's Failure to Advance Cyber Norms', *Just Security*, 30 June 2017.

States tend to be confronted by a “glass house dilemma” when formulating their legal positions in the cyber domain, torn between an *offensive* desire for permissive rules that leave some operational flexibility for stone-throwing and a *defensive* desire for restrictive rules that protect the glass houses in which they reside.[39] It is this tension that likely explains the vagueness of the UK’s legal position concerning the hostile cyber operations conducted by the Russian military intelligence service. While it was in the UK’s defensive interests to interpret the applicable law to conclude that Russia’s cyber operations violated international law, it was in its offensive interests to remain silent and ambiguous about which specific international legal norms had been violated so as to leave operational leeway for the permissibility of its own hostile cyber operations in the future.[40]

3. PEACETIME CYBER ESPIONAGE

Peacetime cyber espionage is an information-gathering cyber operation, encompassing any act undertaken clandestinely or under false pretences by a State – or actors whose conduct is attributable to a State under international law – that uses cyber capabilities to copy information from closed as opposed to open sources of a foreign State, which falls below the threshold required to constitute a prohibited use of force and occurs outside the context of an armed conflict.[41] Historically, the predominant policy of States with respect to peacetime espionage operations has been one of criminalisation at the domestic level combined with silence as to their legality under international law. In the latter regard, while there is extensive State practice of espionage, which is widely accepted as a core national security function of the State, espionage operations have generally not been accompanied by government statements from which their legality or illegality under international law may be inferred.[42] According to this

[39] Kubo Mačák, ‘On the Shelf, But Close at Hand: The Contribution of Non-State Initiatives to International Cyber Law’, (2019) 113 *AJIL Unbound* 81, 82-84.

[40] In this regard, it is notable that the international legal norm that the UK could most easily have alleged Russia to have violated – sovereignty – had recently been characterised by the UK Attorney General as a general principle from which the UK could not currently extrapolate any “specific rule or additional prohibition for cyber activity beyond that of a prohibited intervention”. UK Attorney General, ‘Cyber and International Law in the 21st Century’, 23 May 2018, available online at: https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century (last accessed 5 January 2019). See similarly, Gary P. Corn and Robert Taylor, ‘Sovereignty in the Age of Cyber’, (2017) 111 *AJIL Unbound* 207, 208 (characterising sovereignty as “a principle of international law that guides state interactions, but is not itself a binding rule that dictates results under international law”). This position seems to be driven by an offensive desire to establish a broad zone of international legal permissibility within cyberspace. See, in this regard, Biller and Schmitt, *supra* n.30 (“because the criteria for engaging in a prohibited intervention or use of force are both demanding and ill-defined, the ‘sovereignty is not a rule’ position affords other States the flexibility to act in an ‘indiscriminate and reckless’ manner while claiming to operate within the boundaries of international law”).

[41] This definition draws on: *Tallinn Manual 2.0*, *supra* n.3, 168; Russell Buchan, *Cyber Espionage and International Law* (Bloomsbury, 2018), Chapter 1; and Asaf Lubin, ‘The Liberty to Spy’, *Harvard International Law Journal* (forthcoming).

[42] Katharina Ziolkowski, ‘Peacetime Cyber Espionage – New Tendencies in Public International Law’, in Katharina Ziolkowski (ED.), *Peacetime Regime for State Activities in Cyberspace: International Law, International Relations and Diplomacy* (NATO CCD COE, 2013) 425, 437-443.

traditional perspective, therefore, acts of espionage have generally been considered to be either permitted on the basis that they are not forbidden by international law, or *prima facie* in violation of general rules of international law but subject to a customary exception that regards those violations as permissible.[43]

With the advent of cyberspace, however, the espionage landscape has evolved. Cyber technologies have improved the efficiency of espionage operations, enabling cheaper, easier, and increasingly remote access to enormous volumes of information.[44] Significantly, the expansive nature of espionage missions in the digital age implicates non-State actors to an unprecedented degree, including, for example, through the bulk collection of personal data as part of State surveillance programmes.[45] The broader scope of cyber espionage operations has also coincided with their increased visibility, whether as a result of leaks, voluntary transparency on the part of States, or simply the heightened detectability of espionage programmes.[46] Responding to this new environment and to growing pressures from corporations, civil society groups, and the general public for greater regulatory constraints, States have begun to be more vocal about the international legal regulation of peacetime espionage operations. To evaluate these new practices, a distinction may usefully be drawn between international legal rules that aim to protect *the rights of States* and those that aim to protect *the rights of individuals*.[47]

Allegations that acts underlying cyber espionage operations violate international legal rules designed to protect *the rights of States* continue to be the exception. For example, in the wake of the 2013 Snowden disclosures concerning the surveillance practices of the US National Security Agency (NSA), the UK Government Communications Headquarters, and their allies, only a small minority of States declared such

[43] Iñaki Navarrete and Russell Buchan, 'Out of the Legal Wilderness: Peacetime Espionage, International Law and the Existence of Customary Exceptions', *Cornell International Law Journal* (forthcoming) (describing the "mainstream view about espionage" as holding that "while different forms of espionage violate different international legal rules, […] general and consistent practice of States acting out of a sense of legal obligation has carved out customary espionage "exceptions" (or "defenses") to those primary rules of international law"). For further discussion of the legality of traditional espionage, see generally, Ashley Deeks, 'An International Legal Framework for Surveillance', (2015) 55 *Virginia Journal of International Law* 291, 300-319; and Darien Pun, 'Rethinking Espionage in the Modern Era', (2017) *Chicago Journal of International Law* 353, 359-368. For an alternative perspective, elaborating a new and innovative legal framework for articulating the law and practice of interstate peacetime espionage operations, see Lubin, *supra* n.41; and Asaf Lubin, 'Cyber Law and Espionage Law as Communicating Vessels', in Tomáš Minárik et al. (eds), *CyCon X: Maximising Effects* (NATO CCD COE Publications, 2018) 203, 219-224.

[44] Ido Kilovaty, 'World Wide Web of Exploitations – The Case of Peacetime Cyber Espionage Operations Under International Law: Towards a Contextual Approach', (2016) *Columbia Science & Technology Law Review* 42, 66-69.

[45] Ashley S. Deeks, 'Confronting and Adapting: Intelligence Agencies and International Law', (2016) *Virginia Law Review* 599, 621-623.

[46] Ibid., 615-621.

[47] Ibid., 631-650. The present analysis of peacetime cyber espionage operations is not intended to be exhaustive – omitting, for example, consideration of the relationship between cyber espionage and diplomatic and consular law, as well as the relationship between economic cyber espionage and the World Trade Organisation. For a comprehensive overview of cyber espionage and international law, see generally, Buchan, *supra* n.41.

programmes to constitute violations of State-focused international legal rules.[48] Most prominently, the Brazilian President at the time, Dilma Rousseff, characterised the NSA surveillance programme as a situation of "disrespect to […] national sovereignty […and] a breach of international law".[49] The Foreign Ministry of Mexico issued a press release condemning US surveillance practices with respect to the Mexican government and president as "unacceptable, unlawful, and contrary to Mexican law as well as international law".[50] Indonesia also claimed that extraterritorial surveillance practices violate international law and the UN Charter,[51] while the Bahamas argued that the NSA's secret interception of virtually every cell phone conversation in the country had led its citizens to question "what these high ideals of territorial integrity, sovereignty and respect for the rule of law actually mean in practice".[52] The Chinese government also declared that the NSA's surveillance practices had "flagrantly breached international laws […and] deserve to be rejected and condemned by the whole world".[53] Yet, not only were these statements small in number compared to the extensive reach of the surveillance programmes revealed by the Snowden leaks, they were also variable and ambiguous in their specificity.[54] In addition, the sincerity of some of these statements is questionable in light of media reports that reveal similar intelligence practices conducted by some of the States that raised these allegations.[55]

In general, therefore, silence concerning the compatibility of peacetime cyber espionage operations with State-focused international legal rules continues to be the prevailing policy of States.[56] To take a prominent example, the US response to the massive data theft from the Office of Personnel Management between 2014 and 2015 has to date been muted. Despite being dubbed "one of the most potentially damaging

[48] For additional analysis of these and other statements submitted by States in response to the Snowden disclosures, see Navarrete and Buchan, *supra* n.43. Beyond State reactions to the Snowden leaks, other practices in support of State-focused intentional legal regulation of espionage operations include the International Court of Justice's provisional measures order in the case between Timor-Leste and Australia, which was based on the plausibility that Australia's interception of information belonging to East Timor located on Australian territory violated East Timor's sovereignty, as well as the German Foreign Ministry's indication to the UK that "tapping communications from a diplomatic mission would be a violation of international law". See generally, Deeks, *supra* n.45, 641-645.

[49] 'Remarks by Dilma Rousseff at the 68th UN General Assembly', *Voltaire Network*, 24 September 2013.

[50] 'Mexico Slams US Spying on President', *Der Spiegel*, 21 October 2013.

[51] Deeks, *supra* n.45, 644.

[52] 'Bahamas Raises NSA Spy Scandal at OAS Summit', *Curaçao Chronicle*, 5 June 2014.

[53] 'China demands halt to 'unscrupulous' US cyber-spying', *The Guardian*, 27 May 2014.

[54] See, in this regard, *Tallinn Manual 2.0, supra* n.3, 169 (noting that there remains insufficient State practice and *opinio juris* to conclude that customary international law prohibits espionage *per se*).

[55] See, for example, 'Brazil Says It Spied on U.S. and Others Inside Its Borders', *The New York Times*, 4 November 2013.

[56] In a notable exception, however, Brian Egan, US State Department Legal Adviser, recently confirmed the US legal position that "there is no *per se* prohibition on such activities under customary international law". Egan, *supra* n.4, 174. This type of statement does not, however, offer insight into how the US views the compatibility of the constituent acts of cyber espionage with general rules of international law. See, in this regard, *Tallinn Manual 2.0, supra* n.3, 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 of the methods employed to conduct cyber espionage are unlawful").

cyber heists in U.S. government history",[57] the only notable public response by a US official has been one of seeming admiration – James Clapper, then-head of the Office of the Director of National Intelligence, remarking that "you have to kind of salute the Chinese for what they did".[58]

By contrast, the compatibility of the acts underlying peacetime cyber espionage operations with *individual-focused* international legal rules has achieved a prominent position on the agenda of the international community. In particular, the question of the compatibility of peacetime cyber espionage practices with international human rights law has been visible in at least three respects.[59]

First, a number of States have responded to disclosures about the espionage practices of other States by alleging violations of international human rights law. The then-President of Brazil, Dilma Rousseff, for example, characterised the NSA's surveillance programme as a "situation of grave violations of human rights and of civil liberties", adding that "[t]he right to safety of citizens of one country can never be guaranteed by violating fundamental human rights of citizens of another country".[60]

Second, States have expressly recognised the dangers posed by surveillance programmes to individual human rights in a series of resolutions adopted by the UN General Assembly and the Human Rights Council concerning the right to privacy in the digital age.[61] In Resolution 68/167 of 2013, for example, the UN General Assembly expressly recognised "the negative impact that surveillance and/or interception of communications, *including extraterritorial surveillance and/or interception of communications*, as well as the collection of personal data, in particular when carried out on a mass scale, may have on the exercise and enjoyment of human rights".[62] In the same resolution, the General Assembly called upon all States to review their

[57] 'Hacks of OPM databases compromised 22.1 million people, federal authorities say', *The Washington Post*, 9 July 2015.

[58] 'U.S. Intelligence Chief James Clapper Suggests China Behind OPM Breach', *The Wall Street Journal*, 25 June 2015.

[59] Deeks, *supra* n.45, 635-641 (also discussing a fourth context, namely constraints placed on State intelligence agencies by each other). In addition to the examples discussed here, another important area of individual-focused norms is data protection law, in particular the EU's General Data Protection Regulation, which indirectly affects espionage activities by regulating the personal data processing practices of private actors like social media companies, which intelligence agencies sometimes compel to release data.

[60] 'Remarks by Dilma Rousseff at the 68th UN General Assembly', *Voltaire Network*, 24 September 2013. See also, 'China demands halt to 'unscrupulous' US cyber-spying', *The Guardian*, 27 May 2014 (noting how the Chinese government also concluded that the NSA programme "seriously infringed upon […] human rights").

[61] UN General Assembly Resolution 68/167, 18 December 2013, U.N. Doc. A/RES/68/167; UN General Assembly Resolution 69/166, 18 December 2014, U.N. Doc. A/RES/69/166; UN General Assembly Resolution 71/199, 19 December 2016, U.N. Doc. A/RES/71/199; Human Rights Council Resolution 28/16, 26 March 2015, U.N. Doc. A/HRC/RES/28/16; and Human Rights Council Resolution 34/7, 23 March 2017, U.N. Doc. A/HRC/RES/34/7. See generally, Carly Nyst and Tomaso Falchetta, 'The Right to Privacy in the Digital Age', (2017) 9 *Journal of Human Rights Practice* 104.

[62] UN General Assembly Resolution 68/167, 18 December 2013, U.N. Doc. A/RES/68/167, Preamble.

procedures, practices, and legislation regarding their surveillance programmes to ensure their compatibility with international human rights law.[63]

Finally, States have begun to be held accountable for their cyber espionage practices through the findings of human rights treaty bodies and litigation. Determining the compatibility of State espionage programmes with international human rights law generally entails answering two questions: first, whether human rights obligations are applicable to extraterritorial surveillance practices; and second, whether human rights obligations have been violated by such practices.

As regards the first question, apart from notable exceptions such as the US and Israel, there is widespread support amongst States that in certain circumstances human rights obligations apply extraterritorially.[64] In this regard, the prevailing view is that the extraterritorial application of human rights obligations requires power or effective control by the State concerned over territory (the spatial model of jurisdiction) or the person affected (the personal model of jurisdiction).[65] Traditionally, this test has been understood to require *physical* control, a condition which is ill-suited to the cyber domain where control over infrastructure and individuals tends to be virtual in nature. Nonetheless, recent indications from human rights experts, treaty bodies, and courts suggest that the "power or effective control" test may be sufficiently malleable to encompass the extraterritorial cyber surveillance practices of States.[66] In this regard, it is notable that the UN Human Rights Committee has concluded that "measures should be taken to ensure that any interference with the right to privacy complies with the principle of legality, proportionality and necessity, *regardless of the nationality or location of the individuals whose communications are under direct surveillance"*.[67] More recently, in the landmark surveillance case, *Big Brother Watch & Others v. the UK*, the European Court of Human Rights (ECtHR) was able to side-step the question of the applicability of the European Convention on Human Rights to extraterritorial surveillance because the UK government decided not to raise a jurisdictional objection on this point. The case offers an example of how the silence of a State can, in certain contexts, enable scrutiny of its practices; the ECtHR was able to proceed "on the

[63] Ibid., para. 4(c).

[64] Monika Heupel, 'How do States Perceive Extraterritorial Human Rights Obligations? Insights from the Universal Periodic Review', (2018) 40, *Human Rights Quarterly* 52.

[65] See, for example, UN Human Rights Committee, 'General Comment 31 – The Nature of the General Legal Obligations Imposed on States Parties to the Covenant', U.N. Doc. CCPR/C/21/Rev1/Add.13, 29 March 2004, para. 10; and *Al-Skeini v. The United Kingdom*, Application No. 55721/07, ECtHR, Judgment, 7 July 2011, paras 133-140.

[66] See, in particular, Barrie Sander, 'Democracy Under The Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections', *Chinese Journal of International Law* (2019, *forthcoming*); Vivian Ng and Daragh Murray, Extraterritorial Human Rights Obligations in the Context of State Surveillance Activities?, *HRC Essex Blog*, 2 August 2016; and Report of the Office of the UN High Commissioner for Human Rights: The Right to Privacy in the Digital Age, 30 June 2014, U.N. Doc. A/HRC/27/37, para. 34.

[67] UN Human Rights Committee, 'Concluding observations on the fourth periodic report of the United States of America', U.N. Doc. CCPR/C/USA/CO/4, 23 April 2014, para. 22 (emphasis added).

assumption that the matters complained of fall within the jurisdictional competence of the United Kingdom".[68]

On the second question, a significant body of case law has developed concerning the compatibility of State surveillance practices with international human rights law, with an emphasis on the right to privacy in particular. The seminal judgment concerning cyber surveillance is the aforementioned case of *Big Brother Watch & Others v. the UK*, in which the ECtHR scrutinised the UK's bulk interception of content and certain metadata relating to so-called "external communications" (i.e. foreign-to-foreign, foreign-to-domestic, and domestic-to-foreign communications), its receipt of US signals intelligence collection, and its compulsion of communication service providers to provide certain metadata on a targeted basis. While space does not permit a thorough examination of the judgment, two aspects were particularly notable.[69] First, the ECtHR effectively normalised the practice of mass surveillance by concluding that "the decision to operate a bulk interception regime in order to identify hitherto unknown threats to national security is one which continues to fall within States' margin of appreciation" and characterising bulk interception as "a valuable means to achieve the legitimate aims pursued, particularly given the current threat level from both global terrorism and serious crime".[70] The judgment's legitimation of mass surveillance programmes – confining its role to determining whether sufficient safeguards have been adopted in their implementation – stands in contrast to sentiments expressed by the Court of Justice of the European Union (CJEU) in the *Schrems* decision of 2015, in which the CJEU stated that "legislation permitting public authorities to have access on a generalised basis to the content of electronic communications must be regarded as compromising the essence of the fundamental right to respect for private life".[71] Second, the ECtHR adjusted in various ways the application of the safeguards it had developed in the context of scrutinising *targeted* surveillance regimes – for example, by dispensing with the requirement for objective evidence of reasonable suspicion in relation to the persons on whom data is being sought – thereby introducing a differentiated approach to the regulation of surveillance that distinguishes between bulk and targeted surveillance practices.[72]

As this analysis indicates, States have generally been far more reticent to discuss the applicability of *State-focused* compared to *individual-focused* norms of international

[68] *Big Brother Watch & Others v. The United Kingdom*, Application Nos 58170/13, 62322/14 and 24960/15, ECtHR, Judgment, 13 September 2018, para. 271. The issue was also not addressed in *Centrum För Rättvisa v. Sweden*, Application No. 35252/08, ECtHR, Judgment, 19 June 2018 (a case in which the ECtHR upheld Swedish legislation that authorised the gathering of covert bulk signals intelligence).

[69] See generally, Theodore Christakis, 'A Fragmentation of EU/ECHR Law on Mass Surveillance: Initial Thoughts on the Big Brother Watch Judgment', *European Law Blog*, 20 September 2018. For commentary on the similar case of *Centrum För Rättvisa v. Sweden*, see Asaf Lubin, 'Legitimizing Foreign Mass Surveillance in the European Court of Human Rights', *Just Security*, 2 August 2018.

[70] *Big Brother Watch & Others, supra* n.68, paras 314 and 386.

[71] *Maximilian Schrems v. Data Protection Commissioner*, C-362/14 ECLI:EU:C:2015:650, Court of Justice of the EU, Judgment, 6 October 2015, para. 94.

[72] *Big Brother Watch & Others, supra* n.68, paras 303 and 316-320.

law in the context of peacetime cyber espionage operations. A number of reasons likely explain this divergence, including the heightened external pressures exerted by human rights groups and courts on States to conform their espionage practices to individual-focused norms, the unpalatability of States arguing that international human rights law does not apply to espionage practices, and the fact that State agencies surrender less flexibility of action in conceding that individual-focused norms apply to their practices compared to State-focused norms.[73]

4. PEACETIME CYBER INFORMATION OPERATIONS

Peacetime cyber information operations are content-based cyber operations, encompassing any act undertaken clandestinely or under false pretences by a State – or actors whose conduct is attributable to a State under international law – that harnesses information in the cyber domain to influence political sentiment in a foreign State, which falls below the threshold required to constitute a prohibited use of force and occurs outside the context of an armed conflict.[74] Examples of cyber information operations include:[75] *dis*information operations, which involve the spread of "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public";[76] and *mal*information operations, which involve threatening, abusive, discriminatory, harassing or disruptive behaviour that aims to cause harm to a person, organisation or State.[77] With the rise of social media, cyber information operations have become increasingly prevalent in recent years, the most high profile being Russia's cyber information operation on the 2016 US presidential election.[78] Importantly, the targets of information operations are the perceptions of an adversary which reside in the cognitive dimension of the information ecosystem.[79] Since the regulation of cyber information operations embroils States in defining the boundaries of content control

[73] Deeks, *supra* n.45, 665-667 (noting that interpreting State-focused norms as strictly applying to espionage activities "would bring to a halt most spying and covert action, as so many of those activities violate other states' territorial integrity and sovereignty, broadly interpreted"). It should be emphasised that the extent to which States have been willing to engage in discussions concerning the application of individual-focused norms to espionage practices has been variable. While States in Europe have proven particularly vocal, other States – such as China, for example – have been relatively silent.

[74] This definition draws on Jen Weedon et al., 'Information Operations and Facebook', *Facebook*, 27 April 2017, 4.

[75] Claire Wardle and Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking* (2017), 20 (also distinguishing the further category of "mis-information", namely information that is false but not created with the intention of causing harm).

[76] 'Communication – Tackling Online Disinformation: A European Approach', European Commission, COM(2018) 236 final, 26 April 2018, 3-4.

[77] Chris Tenove et al., *Digital Threats to Democratic Elections: How Foreign Actors Use Digital Techniques to Undermine Democracy* (2018), 22-25.

[78] Office of the Director of National Intelligence, *Assessing Russian Activities and Intentions in Recent US Elections*, ICA 2017-01D, 6 January 2017. See also, Freedom House, *Freedom on the Net 2017: Manipulating Social Media to Undermine Democracy*, November 2017 (noting that disinformation tactics "played an important role in elections in at least 17 other countries over the past year").

[79] Herbert Lin and Jaclyn Kerr, 'On Cyber-Enabled Information/Influence Warfare and Manipulation', *SSRN* (2017), 6.

and freedom of expression, it is perhaps unsurprising that approaches adopted at the international level to date have been highly divergent.

According to what may be termed the digital authoritarian perspective – whose adherents include China, Russia, and other members of the Shanghai Cooperation Organisation (SCO) – cyber information operations encompass a broad category of "information security" threats, including internal dissent and anti-government information disseminated through cyberspace.[80] As Roger Hurwitz explains, adherents to this perspective tend to be motivated by a desire "to control the ideational space that cyber networks afford their populations", based on a characterisation of cyberspace as "a vector for dissident political information and organizing – one not easily suppressed, but easily exploited by external rivals, in particular the United States".[81] In line with this stance, in 2009 the SCO adopted an agreement which defined "information war" in broad terms as "dissemination of information harmful to political, social and economic systems, as well as spiritual, moral and cultural spheres of other States".[82] Towards the end of 2011, a number of SCO members, including China and Russia, submitted to the UN General Assembly a draft International Code of Conduct for Information Security,[83] which they updated in early 2015.[84] The 2011 draft advocated "curbing the dissemination of information that incites terrorism, secessionism or extremism or that undermines other countries' political, economic and social stability, as well as their spiritual and cultural environment".[85] As Tim Stevens notes, this provision "has been widely interpreted as a defence of internet censorship and states' rights to prohibit access to materials deemed inimical to their ideologies".[86]

[80] Adam Segal, 'Chinese Cyber Diplomacy in a New Era of Uncertainty', *Aegis Paper Series No. 1703* (Hoover Institution, 2017), 3; and Tim Maurer, *Cyber Mercenaries: The State, Hackers, and Power* (CUP, 2018), 54-55.

[81] Roger Hurwitz, 'A New Normal? The Cultivation of Global Norms as Part of a Cybersecurity Strategy', in P.A. Yannakogeorgos and A.B. Lowther (eds.), *Conflict and Cooperation in Cyberspace: The Challenge to National Security* (Taylor & Francis, 2014) 233, 238.

[82] Agreement between the Governments of the Member States of the Shanghai Cooperation Organization on Cooperation in the Field of International Information Security (16 June 2009), unofficial translation available online at: https://ccdcoe.org/sites/default/files/documents/SCO-090616-IISAgreement.pdf (last accessed 5 January 2019).

[83] International Code of Conduct for Information Security (2011), Annex to the Letter dated 12 September 2011 from the Permanent Representatives of China, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General, U.N.Doc. A/66/359 (14 September 2011).

[84] International Code of Conduct for Information Security (2015), Annex to the 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, U.N.Doc. A/69/723 (13 January 2015).

[85] International Code of Conduct for Information Security (2011), *supra* n.83, para (c). See similarly, Draft International Code of Conduct for Information Security (2015), *supra* n.84, para. 2(3) (advocating for States not to use ICTs "to interfere in the internal affairs of other States or with the aim of undermining their political, economic and social stability"); and Astana Declaration of the Heads of State of the Shanghai Cooperation Organization, 9 June 2017 ("member states will continue to strengthen practical interaction in countering propaganda and justifications of terrorism, separatism and extremism in the media").

[86] Tim Stevens, 'A Cyberwar of Ideas? Deterrence and Norms in Cyberspace', (2012) 33 *Contemporary Security Policy* 148, 162.

By contrast, the US and members of the EU have generally refrained from discussing cyber information operations at the multilateral level, rejecting the language of "information security" in favour of a narrower discussion of technical security risks under the banner of "cyber security".[87] The silence of those States at the multilateral level should not, however, be mistaken for a lack of concern for the regulation of cyber information operations at the regional or domestic levels. All States regulate the dissemination of content in their territories, the difference between them being essentially one of degree.[88]

In the EU, for example, illegal content includes incitement to terrorism, xenophobic and racist speech that publicly incites hatred and violence, as well as child sexual abuse.[89] Even the US, which is host to one of the most permissive free speech environments in the world, has federal criminal laws that restrict, for example, child pornography and knowingly providing material support to designated foreign terrorist organizations.[90] These types of *content restriction* laws are often paired with *intermediary liability* laws, which establish the conditions under which intermediaries – including social media platforms – may be held liable for illegal content generated by their users.[91] Germany, for example, recently enacted the Network Enforcement Act (*NetzDG*), which requires major social media platforms with at least two million registered German users to set up an effective and transparent complaints management infrastructure that can ensure illegal content is deleted or blocked within specified timeframes, or risk facing the prospect of penalties of up to €50 million.[92]

Viewed in this light, the reticence of certain States such as the US and members of the EU to discuss the regulation of cyber information operations at the multilateral level is not driven by a disdain for content regulation *per se*, but a fear that an international treaty would serve to legitimise the highly intrusive online censorship practices implemented by digitally authoritarian governments such as China and Russia.[93] Indeed, it is possible that one of the reasons why the Obama administration decided to characterise Russia's information operation on the 2016 US presidential election as merely a "violation of established international *norms* of behavior" was a concern

[87] Segal, *supra* n.80, 3; and Maurer, *supra* n.80, 54-55.

[88] See similarly, Zhixiong Huang and Kubo Mačák, 'Toward the International Rule of Law in Cyberspace: Contrasting Chinese and Western Approaches', (2017) 16 *Chinese Journal of International Law* 271, 294.

[89] 'Communication – Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms', European Commission, COM(2017) 555 final, 28 September 2017, 2.

[90] Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money', *Aegis Series Paper No. 1807* (Hoover Institution, 2018), 12.

[91] Rebecca MacKinnon et al., *Fostering Freedom Online: The Role of Internet Intermediaries* (UNESCO, 2014), 40-43.

[92] See generally, William Echikson and Olivia Knodt, 'Germany's NetzDG: A Key Test for Combatting Online Hate', *Counter-Extremism Project Research Paper No. 2018/09*, November 2018.

[93] See, for example, 'Statement by the Delegation of the United States', *Other Disarmament Issues and International Security Segment of Thematic Debate in the First Committee of the Sixty-Seventh Session of the United Nations General Assembly*, 2 November 2012, available online here: https://perma.cc/3C7Q-DUDP ("we cannot support approaches proposed in the draft Code of Conduct for Information Security that would only legitimize repressive state practices").

that alleging a violation of international law might serve to lend legitimacy to Russia's efforts to significantly restrict freedom of expression, including, for example, the practices of human rights NGOs and other civil society groups.[94]

5. CONCLUSION

This paper has sought to demonstrate the explanatory value of distinguishing between cyber attacks, cyber espionage, and cyber information operations when examining the silences of States concerning the relationship between international law and peacetime cyber operations. Three insights emerge from the analysis.

First, this paper has illuminated the different *targets* of State silences. States may be silent as to the attribution of a cyber operation to another State or the measures taken in response to a particular operation. State silences may also pertain to the *existential* questions of whether or not particular rules fall within the corpus of international law or whether or not specific norms of international law are applicable to particular cyber operations – for example, determining the applicability of international human rights obligations to extraterritorial espionage practices. And finally, State silences may also concern the *expository* question of *the meaning* to be assigned to applicable norms of international law in the cyber context – whether provisions of a treaty or norms of customary international law.[95]

Second, this paper has revealed how the *scope* of State silences can vary depending on the security threat under examination. For some peacetime cyber operations, States have been silent about the applicability of a specific subset of international legal norms and more vocal about others. In the context of cyber espionage operations, for example, States have generally been silent about the applicability of *State-focused* norms of international law compared to their greater openness to discuss *individual-focused* norms of international law such as international human rights law. For other peacetime cyber operations, the spread of silence across different States concerning the applicability of international law in the cyber domain has been uneven. In the context of cyber information operations, for example, digitally authoritarian States have actively sought to legitimize their intrusive censorship practices through the

[94] White House, 'Statement by the President on Actions in Response to Russian Malicious Cyber Activity and Harassment', *Press Release*, 29 December 2016. See, in this regard, Beatrice Walton, 'Duties Owed: Low-Intensity Cyber Attacks and Liability for Transboundary Torts in International Law', (2017) 126 *Yale Law Journal* 1460, 1513 ("[H]olding states responsible for too many cyber [operations] might encourage states to impose draconian restrictions on internet use. […] And broadening the concept of intervention or sovereignty could result in severe problems for NGOs and other supporters of human rights who engage in what might be called low-level coercive activity").

[95] On *existential* and *expository* functions of interpretation, see generally, Duncan B. Hollis, 'The Existential Function of Interpretation in International Law', in Andrea Bianchi et al. (eds), *Interpretation in International Law* (OUP, 2015) 78, 79 ("international law's interpretative process can thus be likened to an iceberg – a rule's meaning arrived at by an interpreter is not simply a function of the method and technique employed (the visible tip) but rests on an array of earlier choices about whether the rule 'exists' to be interpreted in the first place (the iceberg's hidden, critical mass)").

adoption of new multilateral treaties, whereas members of the EU and the US have tended to confine their governance of online content to the regional or domestic levels through the adoption of a mixture of legal and non-legal regulatory measures.

Finally, this paper has revealed some of the possible *rationales* that may underpin the silences of States concerning the applicability and meaning of international law in the cyber domain. These include technical difficulties and geopolitical sensitivities regarding the attribution of peacetime cyber operations to other States, preferences regarding the desired degree of international legal permissibility within cyberspace, a desire to uphold particular values such as freedom of expression rather than risk legitimising intrusive censorship practices, and inclinations towards a "wait and see" approach to the applicability and contours of international law in the context of a fast-changing technological landscape.

Bearing in mind these insights, the significance of the typology outlined in this paper is threefold.

First, by revealing the variable targets, scope, and rationales behind State silences concerning the international law applicable to peacetime cyber operations, the typology reveals an important dimension of the politics that is "part and parcel of international law's structural DNA".[96] As Nicholas Tsagourias explains: "whether states will claim that a violation of international law occurred and take countermeasures depends on many factors, primarily political ones. There is no automaticity as far as the application and enforcement of international law is concerned because states are at the same time law creators, interpreters, and enforcers".[97]

Second, the typology is also salient to the extent that it cautions against the tendency to refer to State silences in uniform terms and to automatically cast such silences in a negative light. Amongst international lawyers, there is often a propensity to fetishise the value of international law, underpinned by an unspoken faith in the transformative potential of law to create order and stability.[98] Yet, as Umut Özsu points out, "the legal form has often underwritten and legitimated precisely the substantive injustice and inequality it is nominally designed to counter".[99] By identifying the distinct targets, scope, and rationales of State silences, this paper has sought to demonstrate that, in certain contexts, a policy of silence may be constructive – for example, to enable the scrutiny of extraterritorial surveillance practices or to prioritise the value of freedom of expression over the potential legitimation of invasive censorship practices. In practice, whether a policy of State silence is deemed appropriate will always be contingent on

[96] Tsagourias, *supra* n.4, 74. See also, Sander, supra n.1.

[97] Ibid.

[98] Jean d'Aspremont, 'Cyber Operations and International Law: An Interventionist Legal Thought', (2016) 21 *Journal of Conflict & Security Law* 575.

[99] Umut Özsu, 'Against Legal Fetishism (Part Two)', *Legal Form*, 3 November 2017.

the type of security threat to which the policy relates, the international legal norms in question, and the observational viewpoint from which the policy is evaluated.

Finally, the typology also sets the foundations for future research examining the legal significance of State silences for the development of international law applicable to different types of peacetime cyber operations. Avenues for future exploration in this context include explaining how State silences may be relied upon to make inferences about the scope and content of international legal obligations,[100] as well as examining how State actions, reactions, accusations, initiatives, and the like – which are silent as to their international legal implications – may over time inform the scope and content of international legal rules applicable to peacetime cyber operations.[101]

[100] See, for example, International Law Commission, 'Identification of Customary International Law: Text of the Draft Conclusions as Adopted by the Drafting Committee on Second Reading', U.N.Doc. A/CN/4/L.908, 17 March 2018, Conclusion 10(3) ("Failure to react over time to a practice may serve as evidence of acceptance as law (*opinio juris*), provided that States were in a position to react and the circumstances called for some reaction"). On different approaches to silence in international law in general, see Helen Quane, 'Silence in International Law', (2014) 84 *British Yearbook of International Law* 240; and Roland Tricot and Barrie Sander, 'Recent Developments: The Broader Consequences of the International Court of Justice's Advisory Opinion on the Unilateral Declaration of Independence in Respect of Kosovo', (2011) *Columbia Journal of Transnational Law* 321, 330-336.

[101] See, for example, Mačák, *supra* n.4, 894 (arguing that the articulation of non-binding voluntary norms in cyberspace may be viewed "as an intermediate stage on the way towards the generation of cyber 'hard law'"); and Hollis and Finnemore, *supra* n.12, 11-12 (arguing that cyber accusations by States concerning State or State-sponsored hostile cyber operations which are silent as to their international legal implications, may serve as early evidence of State practice from which *opinio juris* may emerge over time).

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Addressing Adversarial Attacks Against Security Systems Based on Machine Learning

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

Mirco Marchetti
Department of Engineering
"Enzo Ferrari"
University of Modena and
Reggio Emilia
Modena, Italy
mirco.marchetti@unimore.it

Abstract: Machine-learning solutions are successfully adopted in multiple contexts but the application of these techniques to the cyber security domain is complex and still immature. Among the many open issues that affect security systems based on machine learning, we concentrate on adversarial attacks that aim to affect the detection and prediction capabilities of machine-learning models. We consider realistic types of poisoning and evasion attacks targeting security solutions devoted to malware, spam and network intrusion detection. We explore the possible damages that an attacker can cause to a cyber detector and present some existing and original defensive techniques in the context of intrusion detection systems. This paper contains several performance evaluations that are based on extensive experiments using large traffic datasets. The results highlight that modern adversarial attacks are highly effective against machine-learning classifiers for cyber detection, and that existing solutions require improvements in several directions. The paper paves the way for more robust machine-learning-based techniques that can be integrated into cyber security platforms.

Keywords: *adversarial attacks, machine learning, deep learning, poisoning attacks, evasion attacks, intrusion detection*

1. INTRODUCTION

Solutions based on machine- and deep-learning algorithms are becoming pervasive in multiple fields [1], with documented successes for computer vision, speech processing, social media analysis and healthcare [2]. However, the application of these techniques to cyber security is still affected by several shortcomings that limit their effectiveness in real scenarios. Recent results evidence that utmost care and due diligence should be adopted when considering defensive methods based on machine learning to protect current organizations [3, 4, 5]. There are several motivations for these problems: attacks are relatively infrequent compared to the massive number of events generated by modern enterprises; they evolve rapidly, with consequences for possible ground truth for validation; and attackers are not constrained by rules as in artificial intelligence gaming. In this paper, we consider the additional problem presented by the inherent vulnerability of machine-learning methods to adversarial attacks, through which opponents can thwart the system by inducing the generation of incorrect or undesirable results [6]. This issue is aggravated by the multiple variations of malicious actions that can be performed during the training- or test-time of the machine-learning algorithms [7, 8].

Adversarial attacks against machine learning have been explored in image processing [9], but lack adequate analyses in the cyber security domain. The papers that evaluate the performance of cyber detectors in adversarial settings (e.g., [7, 10]) consider a limited number of cyber security problems, few machine-learning classifiers, and a restricted subset of adversarial attacks. The main focus is on spam and malware analysis [11, 12], while we consider this issue from a network intrusion detection perspective [13], where experimental evaluations and novel solutions are lacking [5]. We provide a comprehensive overview of adversarial attacks against cyber security applications of machine learning and propose a taxonomy of these threats in three areas: network intrusion detection, malware analysis, and spam and phishing detection. We present existing solutions to counter this menace, and propose an original method for mitigating attacks based on data poisoning. We have executed a large set of experiments to evaluate and compare the performance of cyber detectors under normal and adversarial settings. In addition, we have measured the effectiveness of some countermeasures, including the strategy proposed in this paper.

The remainder of this paper is structured as follows. Section 2 provides a thorough description of adversarial attacks in the cyber security sphere. Section 3 explores existing strategies for countering these threats and proposes our original methodology against poisoning attacks. Section 4 presents the experimental results and evaluations. Section 5 concludes the paper with final remarks and future work.

2. ADVERSARIAL ATTACKS AND CYBER SECURITY

To defend against cyber threats, security operators rely on techniques borrowed from the machine-learning domain [14, 15] because of their *anomaly detection* capabilities, which may identify novel attacks and which are not recognizable through *signature-based* approaches [16, 17]. Machine-learning algorithms can be divided into *supervised* and *unsupervised* techniques, depending on the requirement of the training phase, with a set of labelled data [18]. Both groups can solve cyber security problems [3], but supervised methods are appreciated due to their ability to provide actionable results, such as detecting an attack [4]. On the other hand, unsupervised techniques are employed for ancillary tasks such as data clustering [19]. All these methods present several open issues that must be considered when integrating them into security systems [18]. Here, we focus on the topic of adversarial attacks.

Adversarial attacks against machine-learning solutions represent a major limitation to the adoption of a fully autonomous cyber defence platform. These threats are based on the generation of specific samples that induce the model to produce an output that is favourable to the attacker, and leverage the intrinsic sensitivity of machine-learning models to their internal configuration settings [14, 20, 21]. Although adversarial perturbations affect all applications of machine learning, the cyber security field presents several characteristics that further aggravate this menace: there is a constantly evolving arms race between attackers and defenders; the system and network behaviour of an organization can be subject to continuous modifications. These unavoidable and unpredictable changes are denoted as the *concept drift* [22] problem, which decreases the performance of any model based on anomaly detection. Mitigations involve periodic retraining and adjustment processes that can identify behavioural modifications and recent related threats. While performing such operations is a challenging task in itself [18], it also facilitates the execution of adversarial attacks [23].

Many research results (e.g., [6, 24, 8]) show that machine-learning algorithms are unsuitable to face adversarial settings. The first examples of adversarial attacks date back to 2004 [25], but the advent of deep learning drew the attention of the research community to this issue [26]. Possible countermeasures have appeared in the computer vision literature [9], with several papers proposing solutions for improving the robustness of deep neural networks for image classification in adversarial environments [27]. However, the performance of machine-learning algorithms depends on their application contexts, hence it is of paramount importance to understand the effects of adversarial threats against cyber security detectors. We consider different classes of attacks by proposing a taxonomy inspired by the work of Huang et al. [6], where threats are classified on the basis of two properties: the *influence* determines whether an attack is performed at training-time or test-time; the *violation* denotes the type of security violation that may affect availability or integrity of the system.

- **Influence**
  - *Training-time*: these attacks include the manipulation of the training set used by the machine-learning model through the insertion or removal of specific samples that alter the decision boundaries of the algorithm. They are also known as *poisoning attacks*.
  - *Test-time*: These attacks assume that the detector has been deployed and aim to subvert its behaviour through the submission of specific samples during its operational phase.
- **Violation**
  - *Integrity*: often referred to as *evasion attacks*, these attacks aim to increase the false negative rate of the model by introducing malicious samples that are classified as benign. Hence, when successful, these stealthy threats do not cause any defensive action to be taken by the targeted organization.
  - *Availability*: these attacks make the targeted model useless, for example by causing overwhelming spikes of false alarms. For this reason, attacks of this type usually induce some sort of response action by the defending side, such as temporary shut-down and recalibration of the model.

A comprehensive classification of adversarial attacks requires a definition of the attacker model. According to Biggio et al. [24], we should consider the following main features.

- The **goal** is related to the security violation purpose of the adversarial attack.
- The **knowledge** denotes the information possessed by the attacker on the machine-learning system that may include the adopted algorithm, its parameters, and its training data set. Depending on the type of information,

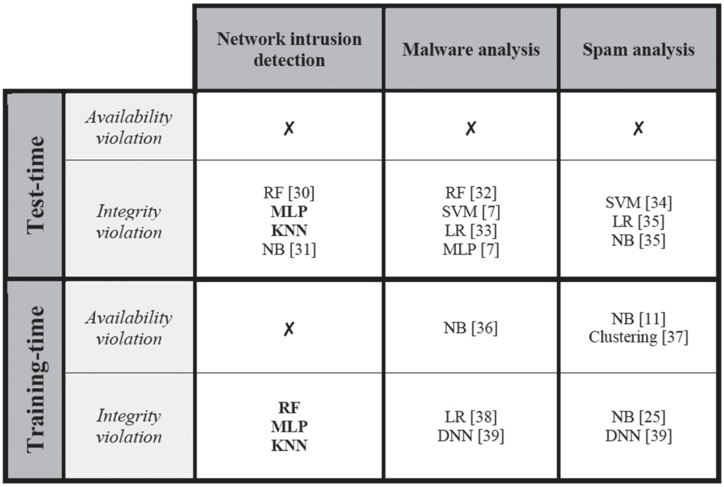
we can distinguish between *black box* attacks (zero knowledge), *grey box* attacks (partial knowledge), and *white box* attacks (complete knowledge).

- The **capability** determines the type of actions that an attacker can perform against the targeted environment that includes, but is not limited to, the machine-learning system. As a strict requirement, it is important to specify which kind of access the attacker has to the cyber detector: he can have full access (that is, reading its output and modifying its internals), limited access (can only read its output) or no access at all.
- The **strategy** denotes the workflow pursued by the attacker to achieve his goal by leveraging previous knowledge and capabilities.

The attacker model distinguishes the adversarial attacks against cyber security systems from offences against other domains of application of machine learning. For example, most papers on image recognition [9, 28] assume that the attacker has complete knowledge and capability. These assumptions are unrealistic in cyber security applications for two reasons: cyber detectors are protected by multiple defence layers; if an attacker overcomes these barriers and can modify the detector, he can achieve his goals without relying on adversarial attack strategies. Thus, in the remainder of this paper, we consider attacks in which the attacker has limited or no access to the machine-learning system.

In Table 1, we classify the most important examples of adversarial attacks against three cyber security areas (Network intrusion detection, Malware analysis, Spam detection) representing scenarios where machine-learning methods are achieving appreciable results (e.g., [14, 15, 29]). In this table, columns indicate the cyber security problem while rows denote the adversarial attack class. Each cell reports the machine-learning algorithms that are tested against the related class of attacks. We remark that algorithms written in bold are evaluated for the first time in this paper. The existing literature focuses mainly on integrity attacks, with several algorithms evaluated for Malware analysis and Spam analysis. Few solutions exist and are tested in the Network intrusion detection context, and this observation motivates this paper. There are few documented attacks targeting the system availability, and there are no specific studies at test-time.

TABLE 1. MAPPING OF THE CATEGORIES OF ADVERSARIAL ATTACKS TO CYBER SECURITY PROBLEMS. LEGEND: RF=RANDOM FOREST; MLP=MULTI-LAYER PERCEPTRON; KNN=K-NEAREST NEIGHBOUR; NB=NAÏVE BAYES; SVM=SUPPORT VECTOR MACHINE; LR=LOGISTIC/LINEAR REGRESSION; DNN=DEEP NEURAL NETWORK.



		Network intrusion detection	Malware analysis	Spam analysis
Test-time	*Availability violation*	✗	✗	✗
	Integrity violation	RF [30] **MLP** **KNN** NB [31]	RF [32] SVM [7] LR [33] MLP [7]	SVM [34] LR [35] NB [35]
Training-time	*Availability violation*	✗	NB [36]	NB [11] Clustering [37]
	Integrity violation	**RF** **MLP** **KNN**	LR [38] DNN [39]	NB [25] DNN [39]

3. DEFENCES AGAINST ADVERSARIAL ATTACKS

Devising effective solutions against adversarial attacks is a challenging task. We present existing methods proposed in the literature that aim to mitigate these critical threats. Countermeasures can be divided into two groups: those conforming to the *security-by-design* paradigm that are effective against perfect-knowledge attacks; and methods that are only effective against partial- or zero-knowledge attacks. One of the main limitations of most solutions against adversarial attacks is that they may worsen the performance of the cyber detector in the absence of adversarial attacks, typically causing higher false positive rates (e.g., [40, 27, 41, 42]).

A. Defences Against Attacks at Test-time

Since there are no known examples of availability attacks at test-time, we focus on defences against attacks targeting the integrity of the system. These threats involve the creation of specific samples that evade the detection mechanism. For example, an opponent can alter a malicious sample to induce its classification as a benign sample. The security-by-design countermeasures aim to improve the machine-learning system capabilities to detect even adversarially manipulated samples.

- **Adversarial training.** These solutions train the model on datasets that include samples of possible adversarial attacks [43]. A recent proposal [42] suggests the adoption of a generative adversarial network (GAN) to automatically generate a similar dataset, achieving promising results. However, these approaches are not a "catch-all" solution, because it is simply unfeasible to obtain a dataset that contains all possible variations of realistic adversarial samples.
- **Robust optimisation.** The authors in [44] and [45] propose techniques aimed at smoothing the decision boundaries of the machine-learning algorithm, thus reducing the effects of adversarial samples. Similar solutions can help to mitigate some attacks, but expert opponents are still able to craft malicious samples that look like licit activities.
- **Feature selection.** Other proposals (e.g., [40, 5]) suggest training the detection model by considering only the subset of features that cannot be manipulated by an attacker. While this method can prevent certain types of evasion attacks, feature removal reduces the detection rates in non-adversarial scenarios [40].
- **Game theory.** These approaches represent the problem of adversarial attacks as a zero-sum game between the attacker and the defender, and work under several assumptions. They require a model of the attacker knowledge and capabilities that must be integrated into the machine-learning algorithm. The optimal defence course against the modelled attacker is found when the system reaches an equilibrium. An example of application to spam detection is described in [46]. The main limitation of these strategies is that they are only able to counter attacks that strictly conform to the considered attacker's model, because even small deviations nullify their effectiveness. Since the cyber security world is intrinsically unpredictable and fuzzy, most of these solutions are not applicable to real contexts.
- **Ensemble methods.** The paper by Biggio et al. [47] shows that it is possible to counter evasion attacks at test-time by devising systems composed by multiple classifiers. However, each classifier represents a weak link in the security chain because the misconfiguration of even one component can lead to poor results, as shown in [48].

Most black- and grey-box evasion attacks involve a *probing* step, in which the adversary aims to gather information on the detector by submitting specific inputs to the system and observing the subsequent response. Thus, existing defences address these malicious exploratory activities by providing misleading information to the attacker. For example, the authors in [47] suggest classifiers that are difficult to reverse-engineer or propose a randomization of the detector output. The problem of these solutions is that they tend to work against attackers with limited time or skill that adopt automated tools. Expert opponents can detect such deception activities and bypass them.

B. Defending Against Attacks at Training-time

Attacks performed at training-time alter the decision process of the machine-learning algorithm by modifying the configuration of the model before the training phases, that is, by manipulating the training dataset(s). Existing solutions focus on protecting the training dataset with the objective of minimizing the effects of adversarial perturbations. We identify the following two groups of security-by-design defences.

- **Data sanitization.** Poisoning attacks are countered through a data sanitization process that aims to detect and remove poisoned samples introduced in the training data [49]. The problem is that some assumptions of these approaches are not always applicable to the cyber security field. For example, the work in [50] assumes that each poisoning sample significantly affects the training process. This assumption is not valid in many situations in which an attacker introduces few samples just to avoid some specific detections of his interest. Other solutions [51] leverage the *machine unlearning* concept that allows the effects of poisoned data to be cancelled without the need to retrain the machine-learning model. The main limitation of this approach is that it needs to know which (poisoned) data to unlearn, that is, it requires the knowledge of which poisoned data samples have been introduced by the attacker. This is an unrealistic assumption in real cyber security contexts.
- **Ensemble methods.** The adoption of multiple-classifier systems can also be effective against attacks at training-time [50]. These solutions present the same advantages and problems characterizing their test-time version, that is, a misconfiguration of even one component can damage the results of the entire detection mechanism.

Defences against partial- or zero-knowledge attacks include the collection of training data from randomized sources [52] with the goal of making it harder for the attacker to devise effective adversarial samples; and the application of strategies to prevent the attacker from controlling the actual training dataset [52]. As an example of this latter group, we propose an original methodology based on the idea of generating the actual training set only at training-time. The approach introduces data transformation procedures on the training dataset. In this way, even if an adversary manages to poison the stored dataset by injecting malicious samples that are labelled as benign, the data transformation step ensures that the model is not trained on those exact poisoned samples. The expected result is that these samples will have a significantly smaller impact on the detector. The complete description of this solution is as follows.

We assume an organization that adopts a cyber detector relying on a supervised algorithm, which is periodically retrained. The training is based on a dataset X' that is stored on a dedicated database server. Let T be an invertible function with domain K so that:

$$T^{-1}(T(k)) = T^{-1}(k') = k, \quad \forall\, k \in K$$

The organization employs the transformation defined by T. More specifically, each time a new piece of data x is added to the dataset X', it is transformed as $T(x) = x'$. When it is necessary to retrain the detector, the dataset X' is retrieved and is inversely transformed through T^{-1}, providing the original training dataset, X.

Now, let us assume that an attacker obtains full access to the database server containing X'. The attacker attempts a poisoning attack by introducing some samples $\overline{x}$ in X' that are labelled as benign, and that represent malicious actions. (For example, the underlying code or network behaviour of a piece of malware, or a spam email). As the attacker is unaware of the data transformation, he does not try to infer the existence of a similar function by analysing the dataset and does not apply the data transformation T to the $\overline{x}$ samples. When the detector is retrained, these samples $\overline{x}$ will undergo the transformation T^{-1}, resulting in samples $\overline{x}^{-1}$ with different characteristics than those of the malicious actions that the attacker wanted to evade detection. This results in poisoning samples whose effect on the detector will be different from that desired by the attacker. We report the entire workflow of the proposed approach in Figure 1 and Figure 2.

FIGURE 1. WORKFLOW OF THE PROPOSED POISONING COUNTERMEASURE: OPERATIONS PERFORMED BEFORE THE (RE)TRAINING.

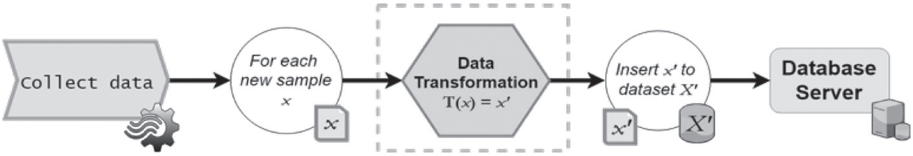
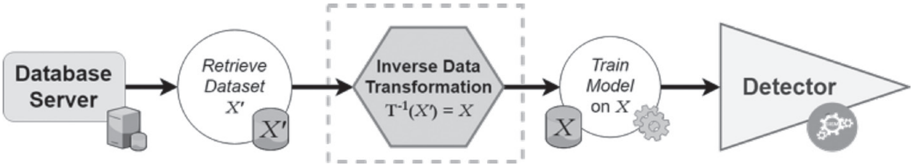


FIGURE 2. WORKFLOW OF THE PROPOSED POISONING COUNTERMEASURE: OPERATIONS PERFORMED AT (RE)TRAINING-TIME.



To provide an improved understanding of the proposed method, we present the following example. Consider an organization adopting a classifier C that analyses network flows [53] to distinguish between malicious and benign traffic; let $\hat{X}$ be the dataset of network flows used to train the classifier, and let $\hat{\mathrm{T}}$ be a transformation that modifies a flow sample by multiplying the *flow_duration* by $d \in \mathbb{R}$, and dividing the *flow_exchanged_bytes* by $b \in \mathbb{R}$; conversely, $\hat{\mathrm{T}}^{-1}$ modifies a flow sample by dividing its *flow_duration* by d and multiplying its *flow_exchanged_bytes* by b. With these assumptions, the dataset $\hat{X}$ is stored in the organization database as $\hat{X}'$. That is, every flow sample $\hat{x} \in \hat{X}$ is modified into $\hat{x}'$ by having the values of its *flow_duration* multiplied by d, and the values of its *flow_exchanged_bytes* divided by b. Therefore, every time the dataset $\hat{X}'$ is updated with a new set of flows, the flows are subject to the transformation denoted by $\hat{\mathrm{T}}$. Consequently, whenever the classifier C undergoes a retraining process, each flow $\hat{x}' \in \hat{X}'$ will be inversely transformed by $\hat{\mathrm{T}}^{-1}$ into its original version, $\hat{x}$.

Now, if an unaware attacker attempts to poison the stored dataset $\hat{X}'$ by inserting some adversarial samples $\hat{\hat{x}}$ that are wrongly labelled, he will not perform the transformation defined by $\hat{\mathrm{T}}$, that is, the adversarial flows will not have their *flow_duration* and *flow_exchanged_bytes* modified. Hence, when the classifier, C is retrained, the adversarial samples $\hat{\hat{x}}$ will be transformed by $\hat{\mathrm{T}}^{-1}$ into $\hat{\hat{x}}^{-1}$. As a practical example, if b=10 and d=2, and if an attacker introduces in $\hat{X}'$ the adversarial sample, $\hat{\hat{x}}$ having *flow_duration*=2 and *flow_exchanged_bytes*=240, then $\hat{\mathrm{T}}^{-1}$ will modify it into $\hat{\hat{x}}^{-1}$ having *flow_duration* =1 and *flow_exchanged_bytes*=2400. Thus, this sample will have different effects on the retraining process of classifier C than the ones intended by the attacker.

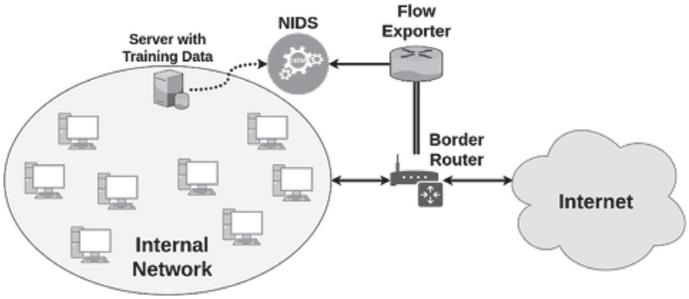
4. EXPERIMENTAL RESULTS

We present an original evaluation of integrity attacks performed at training- and test-time against network intrusion detection systems based on three supervised machine-learning algorithms that achieve appreciable detection performance [14]: Random Forest, Multi-layer Perceptron, K-Nearest Neighbour. We initially present the application scenario, the experimental testbed, and the baseline performance of the considered detectors. Then, we evaluate them in adversarial scenarios and assess the effectiveness of possible countermeasures.

A. Experimental Environment

We consider a typical context, shown in Figure 3, where the network of a large enterprise is monitored by a NIDS based on a machine-learning classifier that inspects the network flows of the border router [53]. The NIDS is periodically retrained with updated data stored on a dedicated database server.

FIGURE 3. SCENARIO ADOPTED FOR THE EXPERIMENTS.



The testbed is based on a publicly available collection of multiple datasets of network flows captured in a monitored environment with dozens of hosts, where some machines are infected with malware belonging to seven botnet families [54]. Overall, these datasets contain over 20 million network samples that are labelled as either legitimate or illegitimate. In the evaluation, we split each detector into several instances, each devoted to one botnet family. Each instance is trained on a training set containing 80% of the malicious samples of the related botnet family, while the remaining 20% is used in the test-set. We use a fixed 85:15 ratio of legitimate-to-illegitimate samples for each training- and test-set. The quality of each detector is measured through the traditional performance indicators *Precision, Recall* (or *Detection Rate*), *F1-score* and *Accuracy*:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. A positive refers to a malicious sample. The values presented in Table 2 represent the average of the results for each detector. These detectors obtain an appreciable performance that is comparable to the state-of-the-art [14, 55].

TABLE 2. BASELINE PERFORMANCE OF THE CLASSIFIERS.

	Precision	Recall	F1-score	Accuracy
Random Forest	0.9774	0.9684	0.9729	0.9978
Multi-layer Perceptron	0.9616	0.9438	0.9526	0.9912
K-Nearest Neighbour	0.9558	0.9375	0.9466	0.9909

To measure the effectiveness of adversarial integrity attacks and their countermeasures, we introduce the *attack severity (AS)* metric, where attacks with higher (respectively, lower) magnitude will obtain *AS* scores that are closer to 1 (respectively, 0):

$$AS = 1 - \frac{Recall\ \text{(after the attack)}}{Recall\ \text{(before the attack)}}$$

B. Evaluation of Adversarial Attacks at Test-time

The first experiments involve integrity violations performed at test-time. We consider an attacker that has already established a foothold within the enterprise's internal network by compromising one or more machines with botnet malware; these bots communicate with an external Command and Control server. The attacker model is based on the following three assumptions: his goal is to evade detection in order to expand his control of the internal network [56]; he knows that the organization adopts a botnet detector based on machine learning, which is trained on malware samples that are similar to the variant used by the bots; he can interact with the controlled bots, but he cannot access the botnet detector. To achieve his goal, the attacker plans to slightly modify the network communications performed by the bots (e.g., small increments in the amount of exchanged data and in the communications duration) so that these small perturbations can induce misclassifications of botnet flows. We simulate this realistic attack scenario by altering the following flow-based features: *exchanged_bytes, duration, total_packets*. This process is repeated for all the samples of each botnet variants. Table 3 reports the average results for the three detectors considered, when tested against these adversarial samples. All these algorithms are severely affected by the adversarial attacks: the detection rate in the second column is about one-third of the original rate.

TABLE 3. EFFECTS OF THE EVASION ATTACK ON EACH CLASSIFIER.

	Recall (before the attack)	Recall (after the attack)	Attack Severity
Random Forest	0.9684	0.3429	**0.6459**
Multi-layer Perceptron	0.9438	0.3012	**0.6809**
K-Nearest Neighbour	0.9375	0.3121	**0.6671**

To defend against similar threats, we explore two of the countermeasures proposed in the literature: *adversarial retraining* and *feature removal*.

For the former case, we harden the detectors by inserting some of the adversarial samples that we manually crafted into their training sets (with the appropriate malicious label), and we repeat the training process. Then, we test the classifiers again on the respective adversarial datasets. The results of this evaluation are reported in Table 4, which compares the severity of the attacks before and after retraining. The decreased severity of the attack after retraining shows the validity of *adversarial retraining*. However, it should be observed that this technique does not guarantee detection against other types of adversarial perturbations.

TABLE 4. EVALUATION OF THE COUNTERMEASURE BASED ON ADVERSARIAL RETRAINING.

	Attack Severity	Attack Severity (after Retraining)
Random Forest	0.6459	**0.3842**
Multi-layer Perceptron	0.6809	**0.4089**
K-Nearest Neighbour	0.6671	**0.4772**

The defences based on *feature removal* aim to nullify the effects of evasion attacks by adopting a set that does not include features related to *duration, exchanged_bytes* and *total_packets*. By training each detector without these features, the results are optimal because the attack severity measure drops to 0. The problem with this approach is that it typically affects the detector performance in scenarios that are not subject to adversarial attacks. By comparing the performance in non-adversarial settings for each detector before and after retraining with the modified feature set, we obtain the results presented in Table 5. All the performance metrics considered fall well below acceptable values for any NIDS. It is possible to attenuate the performance drop by excluding only those features that have a small impact in the decision process of the detector, but this approach will not prevent all evasion attacks.

TABLE 5. EVALUATION OF THE COUNTERMEASURE BASED ON FEATURE REMOVAL IN NON-ADVERSARIAL SETTINGS, AND COMPARISON WITH THE BASELINE PERFORMANCE.

	Precision		*Recall*		*F1-score*		*Accuracy*	
	Original	New	Original	New	Original	New	Original	New
Random Forest	0.9774	**0.8561**	0.9684	**0.8885**	0.9729	**0.8719**	0.9978	**0.9711**
Multi-layer Perceptron	0.9616	**0.7934**	0.9438	**0.7561**	0.9526	**0.7743**	0.9912	**0.9816**
K-Nearest Neighbour	0.9558	**0.8298**	0.9375	**0.8091**	0.9466	**0.8193**	0.9909	**0.9838**

C. Adversarial Attacks at Training-time

We analyse the effects of poisoning attacks that focus on integrity violations. The attacker model considers an opponent who has compromised the targeted network and plans to infect other hosts with novel malware. He is aware that the network is monitored by a NIDS based on some supervised machine-learning algorithms, and he also knows that this detector is periodically retrained. His goal is to ensure that the deployed new malware variants evade detection mechanisms. The attacker has full access to the server that contains the training dataset, but he cannot interact with the detector. To reach his goal, the attacker plans to poison the training dataset through malicious samples representing the behaviour of the deployed malware variant, but that is classified with the benign label.

To simulate this attack scenario, we craft sets of malicious flows that slightly differ (to account for the novel malware variant) from those contained in the testbed, and we label them as benign. This procedure is performed by selecting the existing malicious samples and increasing their *duration* by [1-5] seconds, their *exchanged_bytes* by [1-1024], and their *total_packets* by [1-10]. Then, we inject some of these samples into each training dataset. We measure the effectiveness of a similar attack by comparing the performance of the detectors on the poisoned samples before and after the poisoned retraining phase. The results shown in Table 6 highlight that, before the poisoning attempt, the classifiers were able to identify the novel attack samples with detection rates comparable to other proposals against zero-day malware [17]. The performance of the same algorithms suffered a significant drop after a retraining phase with the poisoned data. The high attack severity score gives a clear idea of the impact of the effect.

TABLE 6. EFFECTS OF THE POISONING ATTACK ON EACH CYBER DETECTOR.

	Recall (before the attack)	Recall (after the attack)	Attack Severity
Random Forest	0.8834	0.2636	**0.7016**
Multi-layer Perceptron	0.8674	0.2777	**0.6798**
K-Nearest Neighbour	0.8611	0.2391	**0.7223**

We now evaluate the original methodology presented in Section 3.B by introducing a custom data transformation procedure on the training set, and then replicating the poisoning attack. For the sake of clarity, we consider a simple function $\widehat{\mathrm{T}}$ that multiplies the *duration* by $d\in\mathbb{R}$, and divides the exchanged_bytes by $b\in\mathbb{R}$. In this way, the poisoned samples introduced by the attacker are (inversely) transformed into samples that are different from the flows generated by the malware variant, because they have durations of $+[\frac{1}{d},\frac{5}{d}]$ seconds (instead of +[1,5]) in which the hosts exchange +[1**b*,1024**b*] bytes (instead of +[1,1024]).

In Table 7, we compare the attack severity of the poisoning attempt before and after the application of the countermeasure, from which we can deduce that the proposed approach can significantly mitigate the effects of a poisoning attack.

TABLE 7. EVALUATION OF THE PROPOSED DEFENSIVE METHOD.
THESE RESULTS ARE OBTAINED BY SETTING d=2 AND m=5.

	Attack Severity	Attack Severity (with Data Transformation)
Random Forest	0.7016	**0.1587**
Multi-layer Perceptron	0.6798	**0.1741**
K-Nearest Neighbour	0.7223	**0.2830**

5. CONCLUSIONS

Machine- and deep-learning algorithms are adopted in many application domains, but in the cyber security field, they are affected by several open issues. In this paper, we consider adversarial attacks where the machine-learning model is compromised to induce an output favourable to the attacker. Literature on this subject is still immature, and most documented examples of adversarial attacks against security systems consider only few algorithms and few application areas. We present a taxonomy of adversarial attacks that evidences which cyber security areas and which machine-learning algorithms have been evaluated against what type of threat. This analysis evidences that there is space for novel research in the context of adversarial attacks against network intrusion detection systems based on machine learning. We are confident that the large set of original experiments and the novel way to address issues related to adversarial attacks presented here can pave the way for cyber detection platforms that are based on more robust machine-learning algorithms.

REFERENCES

[1] Y. LeCun, Y. Bengio and G. Hinton, “Deep learning,” *Nature*, 2015.

[2] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, 2015.

[3] S. Dua and X. Du, Data mining and machine learning in cybersecurity, Auerbach Publications, 2016.

[4] M. Stevanovic and J. M. Pedersen, “On the use of machine learning for identifying botnet network traffic,” *Journal of Cyber Security and Mobility*, 2016.

[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. Rubinstein and J. Tygar, “Adversarial machine learning,” in *Proc. 4th ACM Workshop Security and Artificial Intelligence*, 2011.

[7] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto and F. Roli, “Evasion attacks against machine learning at test time,” in *Joint Eur. Conf. Machine Learning and Knowledge Discovery in Databases*, 2013.

[8] N. Papernot, P. McDaniel, A. Sinha and M. Wellman, “Towards the science of security and privacy in machine learning,” *arXiv preprint arXiv:1611.03814*, 2016.

[9] Please change: N. Papernot, P. McDaniel, A. Sinha and M. Wellman, “SoK: Security and privacy in machine learning,” 2018 IEEE Europ. Symp. on Security and Privacy.

[10] B. Biggio, B. Nelson and P. Laskov, “Poisoning attacks against support vector machines,” Proc. 29th Int. Conf. Machine Learning, 2012.

[11] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar and K. Xia, “Exploiting Machine Learning to Subvert Your Spam Filter,” *Proceedings of the LEET, USENIX Association*, 2008.

[12] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto and F. Roli, “Yes, machine learning can be more secure! a case study on android malware detection,” *IEEE Trans. Depend. Sec. Comput*., 2017.

[13] F. Pierazzi, G. Apruzzese, M. Colajanni, A. Guido and M. Marchetti, “Scalable architecture for online prioritization of cyber threats,” in *2017 NATO 9th Int. Conf. Cyber Conflict*.

[14] A. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Communications Surveys & Tutorials*, 2015.

[15] E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” *Artificial Intelligence Review*, 2008.

[16] R. P. V. Sommer, “Outside the closed world: On using machine learning for network intrusion detection,” in *2010 IEEE Symp. Security and Privacy*.

[17] M. Alazab, S. Venkatraman, P. Watters and M. Alazab, “Zero-day malware detection based on supervised learning algorithms of API call signatures,” in *Proc. 9th Australasian Data Mining Conf*., 2011.

[18] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido and M. Marchetti, “On the effectiveness of machine and deep learning for cyber security,” in *2018 10th Int. Conf. Cyber Conflict*.

[19] G. Apruzzese, M. Marchetti, M. Colajanni, G. Gambigliani Zoccoli and A. Guido, “Identifying malicious hosts involved in periodic communications,” in *2017 16th IEEE Int. Symp. Network Computing and Applications*.

[20] M. Mannino, Y. Yang and Y. Ryu, “Classification algorithm sensitivity to training data with non representative attribute noise,” in *Decision Support Systems*, 2009.

[21] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, Data Mining: Practical machine learning tools and techniques, 2016.

[22] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy and A. Bouchachia, “A survey on concept drift adaptation,” *ACM Computing Surveys*, 2014.

[23] A. Kantchelian, S. Afroz, L. Huang, A. C. Islam, B. Miller, M. C. Tschantz, R. Greenstadt, A. D. Joseph and J. Tygar, "Approaches to adversarial drift," in *Proc. 2013 ACM Workshop Artificial Intelligence and Security*, 2013.

[24] B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto and F. Roli, "Security evaluation of support vector machines in adversarial environments," in *Support Vector Machines Applications*, 2014.

[25] N. Dalvi, P. Domingos, S. Sanghai and D. Verma, "Adversarial classification," in *Proc,. 10th ACM Int. Conf. Knowledge Discovery and Data Mining*, 2004.

[26] M. Fredrikson, S. Jha and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM Conf. Computer and Communications Security*, 2015.

[27] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symp. Security and Privacy*.

[28] S. Huang, N. Papernot, I. Goodfellow, Y. Duan and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[29] T. Chakraborty, F. Pierazzi and V. Subrahmanian, "Ec2: Ensemble clustering and classification for predicting android malware families," *IEEE Trans. Depend. Sec. Comput*, 2017.

[30] G. Apruzzese and M. Colajanni, "Evading Botnet Detectors Based on Flows and Random Forest with Adversarial Samples," in *2018 IEEE 17th Int. Symp. Network Computing and Applications*.

[31] C. V. Wright, S. E. Coull and F. Monrose, "Traffic Morphing: An Efficient Defense Against Statistical Traffic Analysis.," in *Network and Distributed System Security Symp*., 2006.

[32] P. Laskov, "Practical evasion of a learning-based classifier: A case study," in *2014 IEEE Symp. Security and Privacy*.

[33] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert and F. Roli, "Is feature selection secure against training data poisoning?," in *Int. Conf. Machine Learning*, 2015.

[34] D. Lowd and C. Meek, "Adversarial learning," in *Proc. 11th ACM Int. Conf. Knowledge Discovery in Data Mining*, 2005.

[35] D. Lowd and C. Meek, "Good Word Attacks on Statistical Spam Filters," in *CEAS*, 2005.

[36] J. Newsome, B. Karp and D. Song, "Paragraph: Thwarting signature learning by training maliciously," in *Int. Workshop Recent Advances in Intrusion Detection*, 2006.

[37] B. Biggio, I. Pillai, S. Rota Bulo, D. Ariu, M. Pelillo and F. Roli, "Is data clustering in adversarial settings secure?," in *Proc. 2013 ACM Workshop Artificial Intelligence and Security*.

[38] C. Liu, B. Li, Y. Vorobeychik and A. Oprea, "Robust linear regression against training data poisoning," in Proc. *10th ACM Workshop Artificial Intelligence and Security*, 2017.

[39] L. Munoz-Gonzalez, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. *10th ACM Workshop Artificial Intelligence and Security*, 2017.

[40] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE Trans. Cybern*., 2016.

[41] B. Biggio, B. Nelson and P. Laskov, "Support vector machines under adversarial label noise," in *Asian Conf. Machine Learning*, 2011.

[42] H. S. Anderson, J. Woodbridge and B. Filar, "DeepDGA: Adversarially-Tuned Domain Generation and Detection," in *Proc. 2016 ACM Workshop Artificial Intelligence and Security*.

[43] A. Kantchelian, J. Tygar and A. Joseph, "Evasion and hardening of tree ensemble classifiers," in *Int. Conf. Machine Learning*, 2016.

[44] H. Xu, C. Caramanis and S. Mannor, "Robustness and regularization of support vector machines," *Journal of Machine Learning Research*, 2009.

[45] P. Russu, A. Demontis, B. Biggio, G. Fumera and F. Roli, "Secure kernel machines against evasion attacks," in *Proc. 2016 ACM Workshop Artificial Intelligence and Security*.

[46] M. Bruckner and T. Scheffer, "Stackelberg games for adversarial prediction problems," in *Proc. 17th ACM Int. Conf. Knowledge Discovery and Data Mining*, 2011.

[47] B. Biggio, G. Fumera and F. Roli, "Adversarial pattern classification using multiple classifiers and randomisation," in *Joint IAPR Int. Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*, 2008.

[48] I. Corona, B. Biggio, M. Contini, L. Piras, R. Corda, M. Mereu, G. Mureddu, D. Ariu and F. Roli, "Deltaphish: Detecting phishing webpages in compromised websites," in *Eur. Symp. Research in Computer Security*, 2017.

[49] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo and A. D. Keromytis, "Casting out demons: Sanitizing training data for anomaly sensors," in *Proc. 2008 IEEE Symp. Security and Privacy*, 2008.

[50] B. Biggio, I. Corona, G. Fumera, G. Giacinto and F. Roli, "Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," in *Int. Work Multiple Classifier Systems*, 2011.

[51] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symp. Security and Privacy*.

[52] A. D. Joseph, P. Laskov, F. Roli, J. D. Tygar and B. Nelson, "Machine learning methods for computer security," in *Dagstuhl Manifestos*, 2013.

[53] "Netflow," [Online]. Available: https://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netflow/index.html.

[54] S. Garcia, M. Grill, J. Stiborek and A. Zunino, "An empirical comparison of botnet detection methods," *Computers & Security*, 2014.

[55] M. Stevanovic and J. M. Pedersen, "An analysis of network traffic classification for botnet detection," in *2015 Int. Conf. Cyber Situational Awareness, Data Analytics and Assessment*.

[56] G. Apruzzese, F. Pierazzi, M. Colajanni and M. Marchetti, "Detection and threat prioritization of pivoting attacks in large networks," *IEEE Trans. Emerg. Topics. Comput*., 2017.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise

Nicolas Känzig
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
kaenzign@student.ethz.ch

Roland Meier
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Luca Gambazzi
Science and Technology
armasuisse
Thun, Switzerland
luca.gambazzi@armasuisse.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: The diversity of applications and devices in enterprise networks combined with large traffic volumes make it inherently challenging to quickly identify malicious traffic. When incidents occur, emergency response teams often lose precious time in reverse-engineering the network topology and configuration before they can focus on malicious activities and digital forensics.

In this paper, we present a system that quickly and reliably identifies Command and Control (C&C) channels without prior network knowledge. The key idea is to train a classifier using network traffic from attacks that happened in the past and use it to identify C&C connections in the current traffic of other networks. Specifically, we leverage the fact that – while benign traffic differs – malicious traffic bears similarities across networks (e.g., devices participating in a botnet act in a similar manner irrespective of their location).

To ensure performance and scalability, we use a random forest classifier based on a set of computationally-efficient features tailored to the detection of C&C traffic. In order to prevent attackers from outwitting our classifier, we tune the model parameters to maximize robustness. We measure high resilience against possible attacks – e.g., attempts to camouflaging C&C flows as benign traffic – and packet loss during the inference.

We have implemented our approach and we show its practicality on a real use case: Locked Shields, the world's largest cyber defense exercise. In Locked Shields, defenders have limited resources to protect a large, heterogeneous network against unknown attacks. Using recorded datasets (from 2017 and 2018) from a participating team, we show that our classifier is able to identify C&C channels with 99% precision and over 90% recall in near real time and with realistic resource requirements. If the team had used our system in 2018, it would have discovered 10 out of 12 C&C servers in the first hours of the exercise.

Keywords: *malware, botnets, machine learning, digital forensics, Locked Shields, network defense*

1. INTRODUCTION

Large enterprise or campus networks handle data from a vast set of different applications, protocols, and devices. Identifying malicious traffic in such networks is similar to the figurative problem of finding a needle in a haystack, raising the need for effective tools to automate this process and to support defenders such as computer emergency response teams (CERTs) in their operation. As network traffic is not only voluminous but also very diverse, these tools need to adapt to different contexts.

Recent alarming examples of malicious software exploiting a remote infrastructure in order to issue directives to steal or modify data or performing distributed denial-of-service attacks include CryptoLocker [1] or the Mirai botnet [2].

Machine learning-based models have repeatedly been proven to outperform humans in tasks involving large data volumes and high-dimensional feature spaces. However, training these models to detect malicious activity in networks is a particularly challenging task, because the methods used by modern threat actors are continuously evolving. Moreover, the profiles of legitimate background traffic can vary strongly among different networks and their users. Consequently, such solutions might perform well in the environment they have been trained in, while failing in new deployments.

In this paper, we focus on one particular type of malicious traffic: communication between compromised hosts and their Command and Control (C&C) servers. C&C traffic only depends on the botnet (i.e. the communication scheme between the C&C server and the bots) and is invariant to the networks to which the bots are connected. This makes the development of machine learning-based models that perform reliably in different contexts more feasible. We argue that identifying this type of traffic is fruitful because it means that compromised hosts can be identified (and eventually blocked, isolated or patched) before an actual attack is launched.

The work that we present in this paper is based on data from Locked Shields [3], the world's largest cyber defense exercise. While Locked Shields is only an exercise, it reproduces critical infrastructure under the intense pressure of severe cyberattacks. Moreover, it provides a setting that closely matches the real world: in practice, defenders have limited resources to protect a large, heterogeneous network against unknown attacks. And because it is an exercise, we obtained a ground-truth of logs from the attackers describing when and where they were active, something which is hardly possible for real incidents.

Problem statement: Given the constraints (e.g. in terms of computational resources and lack of familiarity with the network) that defending teams face during the Locked Shields exercise, we aim to design a system that can identify C&C traffic and compromised hosts.

Challenges: Solving this problem is challenging for the following reasons:

- Benign and malicious traffic profiles can vary considerably between different Locked Shields exercises.
  This requires a solution with high generalization and robustness.

- Defenders have a very limited budget for computational resources.
  This requires an efficient classification technique.
- Defenders have a small amount of storage capacity.
  This prevents them from storing large amounts of network traffic.
- Defenders have a small bandwidth to access the attacked network.
  This makes it impossible to send large amounts of data to an external system.

Our approach: Our key idea is to use data from past iterations of Locked Shields to efficiently identify similar-looking C&C traffic in future exercises. We do this by creating a labeled dataset containing flow-based features extracted from raw Locked Shields traffic captures, which we then use to train a supervised classifier (random forest) to flag C&C traffic. Our approach is efficient enough to be deployed during future Locked Shields exercises.

Novelty and related work: Detecting C&C traffic has been the focus of many research papers in recent years (cf. surveys in [4] [5]), many of which also pursue classifier-based approaches using machine learning algorithms. [6] proposes a two-stage system for identifying P2P C&C traffic using a decision tree and a random forest classifier. To train a random forest classifier, [7] leverages the fact that malware-related domains are likely to have an inconsistent pool of requesting hosts. [8] develops a system for classifying malicious C&C servers using NetFlow data, extracting features related to flow sizes, client access patterns and temporal behavior.

In contrast to these approaches, we use a new set of flow-based features and evaluate our models on two new and completely labeled datasets (Locked Shields 2017 and 2018). While most studies train and evaluate their models on different parts of the same dataset, we use train- and test-sets that have been acquired independently in different setups. This provides strong evidence for the ability of our system to perform in new environments. Moreover, a minority of the solutions proposed in past investigations claim to run in real time [4]. In our approach, we combine quickly computable features (e.g. number of packets per flow) with an efficient random forest algorithm, which makes real-time calculation feasible.

Contributions: The main contributions of this paper are:

- A selection of features that allow identifying C&C channels while being fast and efficient to compute.
- An efficient random forest model that classifies between C&C traffic and normal traffic with high accuracy.
- An implementation of the system that is suitable for deployment in future Locked Shields exercises.

- An evaluation based on real data from Locked Shields 2017 and 2018, which shows that our system allows defenders to identify C&C traffic, C&C servers and compromised hosts.

Organization: The remainder of this paper is organized as follows. In Section 2, we provide background information on the Locked Shields exercise and define the attacker model. In Section 3, we present our system to identify C&C traffic before we evaluate it in Section 4. In Section 5, we discuss the outcome and finally, we conclude in Section 6.

2. BACKGROUND ON LOCKED SHIELDS

In this section, we explain how Locked Shields is organized and give details about the roles of defenders and attackers.

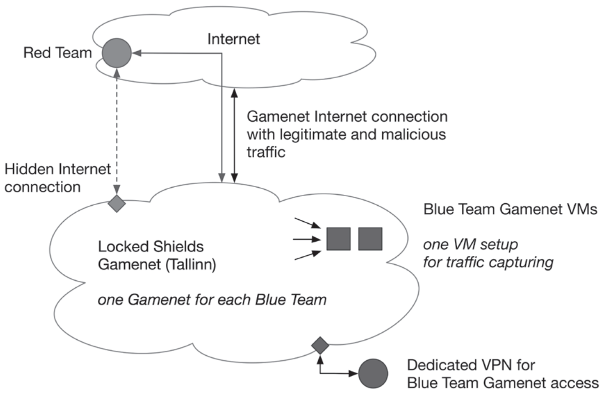
A. Exercise Organization

Locked Shields is the largest and most complex live-fire global cyber defense exercise, with more than 1000 participating cyber experts from 30 nations [9]. It takes place every year and is organized by the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) in Tallinn (Estonia) [3].

For the exercise, participating countries send *Blue Teams*, which represent response teams whose main task is to secure and protect the network infrastructure. Whereas each Blue Team operates in an isolated instance of the network (*Gamenet*), a *Red Team* runs attacks against all these networks in order to compromise or degrade the performance of the connected systems.

In Figure 1, we illustrate the environment during an exercise.

FIGURE 1. LOCKED SHIELDS ENVIRONMENT OVERVIEW.


Red Team
Internet
Gamenet Internet connection with legitimate and malicious traffic
Hidden Internet connection
Blue Team Gamenet VMs
Locked Shields Gamenet (Tallinn)
one VM setup for traffic capturing
one Gamenet for each Blue Team
Dedicated VPN for Blue Team Gamenet access

The environment simulated during the exercise changes every year. In this paper, we focus on the last two occurrences of Locked Shields (2017 and 2018). In 2017, the Blue Teams had to maintain the services and networks of a military air base; in 2018, a major civilian Internet service provider, a military base and other critical infrastructures of a fictional country were targeted in cyber attacks.

B. Environment and Constraints for the Defenders

Prior to the exercise, the defenders (Blue Teams) receive an architecture scheme of the original Gamenet that shows the topology and connected devices. However, the scheme does not show changes put in place by the Red Team (e.g. additional connections between the Gamenet and the Internet to bypass the main gateways).

In addition, each Blue Team obtains two virtual machines (VMs) inside the Gamenet, which it can use during the exercise to install its own tools (e.g. to perform forensics or deploy patches). Moreover, the traffic exchanged in the Gamenet is forwarded to one VM in order to allow the Blue Team to perform on-site analysis and detection. However, the performance of this VM is limited and access to it is only possible via a low-bandwidth VPN tunnel. In order to rapidly counter Red Team activity, the Blue Team has to deploy efficient analysis tools (given the constraints on computation and bandwidth), intrusion detection systems, and to avoid sending voluminous data to an external infrastructure. The system that we present in this paper is designed to work in such a restricted environment.

After the exercise, the Red Team delivers reports to the Blue Teams summarizing their malicious activities.

C. Attacker Model

The attackers (Red Team) perform their activities according to a tight schedule of missions and goals. Waves of attacks hit the Blue Team for the entire duration of the exercise. Some attacks are limited to a specific phase of the exercise while others are repeated during the entire exercise.

Prior to the exercise, the Red Team knows the configuration of the entire Gamenet and can use this knowledge to prepare suitable attacks (e.g. leveraging outdated systems).

In order to systematically orchestrate the large number of attacks on all Gamenets, the Red Team uses Cobalt Strike as a C&C framework. This allows automatizing injections, deployment of malicious code and C&C datalink management.

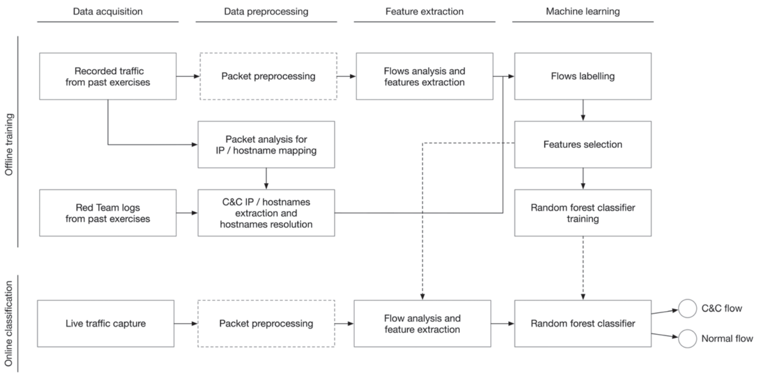
3. SUPERVISED MACHINE LEARNING FOR DETECTING C&C CHANNELS

In this section, we explain how we use supervised machine learning to identify C&C channels in the Locked Shields exercise. First, we provide an overview of our approach. Afterwards, we describe the data and labeling that we used. Finally, we explain how we selected the features and the machine learning model for this task.

A. Overview

Our system consists of two basic phases (Figure 2): offline training and online classification. In the offline training phase (which was done prior to the exercise), we used data from past Locked Shields exercises and processed them in order to obtain a labeled dataset to train a supervised classifier that could be used for live classification of C&C flows during the exercise.

FIGURE 2. SYSTEM OVERVIEW.


Data acquisition
Data preprocessing
Feature extraction
Machine learning
Offline training
Recorded traffic from past exercises
Packet preprocessing
Flows analysis and features extraction
Flows labelling
Packet analysis for IP / hostname mapping
Features selection
Red Team logs from past exercises
C&C IP / hostnames extraction and hostnames resolution
Random forest classifier training
Online classification
Live traffic capture
Packet preprocessing
Flow analysis and feature extraction
Random forest classifier
C&C flow
Normal flow

B. Data Analysis and Enrichment

In this section, we describe the data sources we used for labeling and training and the preprocessing steps we applied.

1) Available Data Sources

We built our labeled dataset from two sources: raw traffic captures and Red Team logs.

a) Raw Traffic Captures

We obtained pcap traffic traces containing the Gamenet activity recorded during Locked Shields 2017 and 2018 (LS17, LS18) from a participating country (Switzerland). The packets are not sampled, anonymized, or truncated. We extracted the features used to train our models from this data.

b) Red Team Logs

The activities of the Red Team are logged in different documents generated by the Cobalt Strike framework [10]. Among others, these documents contain *indicators of compromise* (e.g. IP addresses and domain names of C&C servers) and an *activity report*, which contains a timeline of all Red Team activities (e.g. commands that were executed on compromised machines). We used these log files to label the C&C flows.

2) Data Preprocessing

Before extracting features, we preprocessed the dataset in three ways: *(i)* we truncated packets to reduce the size of the dataset; *(ii)* we aggregated packets to flows; and *(iii)* we mapped domain names and IP addresses in the traffic capture.

a) Truncating Packets

Since capturing and analyzing full packets in real time is difficult for the Blue Team during the exercise, our approach does not require packet payloads. We used only the first 96 bytes (enough to capture everything up to the header of the transport layer) of each packet, which reduced the size of our dataset by approximately 75 percent. The performance of our final models did not decrease due to the truncation.

b) Flow Extraction

We aggregated the packets from the packet trace into flows, since our model operates at the flow level. A flow is defined by its 5-tuple (source IP, destination IP, source port, destination port, transport layer protocol). It starts with a TCP SYN packet and ends when the first TCP FIN packet is sent or after a timeout of 15s. We used CICFlowMeter [11] to extract flow-based features from the raw traffic traces.

c) Domain Name Resolution

The Red Team logs list some devices only by their domain name, thus we needed a mapping from these domain names to the associated IP addresses. We used Bro [12], a network analysis framework, to resolve the domain names to IP addresses from the packet traces, using information contained in the HTTP (host header), TLS (with server name indication), or DNS.

C. Data Labeling

After extracting a list of IP addresses and domain names of C&C servers from the Red Team logs, the labeling process was straightforward: we labeled all flows where at least one endpoint was a C&C server (i.e. listed in the Red Team logs) as malicious and all other flows as benign. The intuition behind this approach was that there was no benign reason for any device to contact a C&C server. It is safe to assume that any device communicating with a C&C server is compromised.

D. Feature Selection and Extraction

In this section, we explain how we selected and extracted the features that our classifier would use to identify C&C flows.

1) Feature Extraction

For computing the features, we used CICFlowMeter [13] (version 3.0), an open source tool for extracting flows from packet traces and computing large sets of features. CICFlowMeter focuses on time-related features such as the inter-arrival time of packets, active and idle times separately for packets in each direction, while including minimum, maximum, mean and standard deviation [11]. These features are suitable for our purposes because they can be extracted with little computational effort.

To capture the fact that C&C servers are typically located outside the internal network, we added an additional feature (Int/Ext Dst IP), indicating whether the destination IP address of a flow is within the internal address space.

Table I lists all features that we considered in our selection process.

TABLE I: COMPLETE LIST OF FEATURES CONSIDERED IN THE FEATURE SELECTION.
ONE ROW CAN DESCRIBE MULTIPLE FEATURES (E.G. THE MINIMUM, MAXIMUM, MEAN AND STANDARD DEVIATION OF A PROPERTY)

Nr	Feature	Description
1	Flow Duration	Duration of the flow in microseconds
2-3	Tot Fwd/Bwd Pkts	Total packets in the fwd/bwd direction
4-5	TotLen Fwd/Bwd Pkts	Total size of packets in fwd/bwd direction
6-13	Fwd/Bwd Pkt Len	Min, Max, Mean, Std size of packet in fwd/bwd direction
14-23	Fwd/Bwd IAT	Total, Min, Max, Mean, Std time between two packets sent in the fwd/bwd direction
24-35	Flag Counts	Flag Counts PSH, URG, SYN, FIN, RST, ACK, URG, CWE, ECE in Fwd/Bwd and both directions. (0 for UDP)
36-37	Fwd/Bwd Header Len	Total bytes used for headers in the fwd/bwd direction
38-40	Fwd/Bwd/Tot Pkts/s	Number of fwd/bwd/tot packets per second
41	Flow Byts/s	Number of flow bytes per second
42-45	Pkt Len	Min, Max, Mean, Std packet length of a flow
46-49	Flow IAT	Min, Max, Mean, Std packet inter-arrival time in fwd/bwd direction
50	Down/Up Ratio	Download and upload ratio
51	Pkt Size Avg	Average size of packet
52-53	Fwd/Bwd Seg Size Avg	Average size observed in the fwd/bwd direction
54-55	Fwd/Bwd Byts/b Avg	Average number of bytes bulk rate in the fwd/bwd direction
56-57	Fwd/Bwd Pkts/b Avg	Average number of packets bulk rate in the fwd/bwd direction
58-59	Fwd/Bwd Blk Rate Avg	Average number of bulk rate in the forward direction
60-61	Subflow Fwd/Bwd Pkts	Average number of packets in a subflow in the fwd/bwd direction
62-63	Subflow Fwd/Bwd Byts	Average number of bytes in a subflow in the fwd direction
64-67	Active Time	Min, Max, Mean, Std time a flow was active before becoming idle
68-71	Idle Time	Min, Max, Mean, Std time a flow was idle before becoming active
72-73	Init Fwd/Bwd Win Byts	TCP window size in the fwd/bwd direction
74	Fwd Act Data Pkts	Count of fwd packets with at least 1 byte of TCP payload
75	Fwd Seg Size Min	Minimum segment size in the forward direction
76	Int/Ext Dst IP	0 if Dst-IP of a flow belongs to Blue Teams subnet, 1 if external
77	L3/L4 Protocol	0 for TCP, 1 for UDP, 2 for ICMP

2) Feature Selection

To identify the best set of features, we removed correlating and irrelevant features by applying a recursive feature elimination scheme based on random forest Gini importance scores [14].

In each iteration, we trained a random forest classifier with the dataset from LS17 and all the considered features. Afterwards, we removed the feature with the lowest score from the set of considered features. Thus, we obtained a feature ranking, where the one that is first removed has the lowest rank. Eliminating features one by one is crucial, as importance scores can spread over multiple features with redundant information (i.e. if multiple important features are strongly correlated, their scores can all be low in a particular iteration).

The 20 most important features according to our feature selection are listed below in descending order of importance (except for the last two features, which we included in the feature set based on preliminary evaluations).

Tot Fwd Pkts, Flow IAT Mean, Fwd IAT Max, Flow Pkts/s, Bwd Pkt Len Min, FIN Flag Cnt, Init Fwd Win Byts, Active Mean, Bwd IAT Mean, Bwd Pkt Len Std, Fwd Seg Size Min, Fwd Pkt Len Std, Tot Bwd Pkts, Bwd Header Len, Subflow Fwd Byts, Subflow Bwd Pkts, Fwd IAT Tot, Flow IAT Max, Int/Ext Dst IP, L3/L4 Protocol

E. Model Selection

We tested a variety of different supervised models on our data: Artificial Neural Network, Support Vector Machine, Logistic Regression, Naive Bayes, K-Nearest Neighbors and Random Forest (RF). The main difficulty in our task was that the distribution of the background traffic was different in the LS17 and LS18 data, as benign and attack traffic profiles change every year. However, the distribution of the C&C session features hardly varies, due to the fact that the same tool (Cobalt Strike) is used to maintain these sessions. We found that RF performed best under these circumstances. Furthermore, RF models are highly efficient and require low training and inference times, which is decisive for real-time deployments.

1) Model Configuration

As a baseline model, we used an RF classifier with default configurations from scikit-learn [15] (i.e. an ensemble of 10 fully expanded trees). However, this resulted in large trees (30,000 nodes for the model trained on LS17, 70,000 for LS18) and we found that constraining the maximal tree-depth significantly increased the robustness of our model. We empirically found that a maximum tree-depth of 10 drastically reduced the node count (to 700 for LS17 and 900 for LS18). However, reducing the depth further had a negative impact on the performance. Moreover, we found that increasing the number of trees to 128 further improved the robustness and prediction quality with negligible impact on computational cost. In the following, we refer to configurations with a maximum depth of 10 and 128 trees as “tuned” configurations.

2) Robustness Against Camouflage

In the following, we analyze possible attack vectors against our model, assuming a white-box scenario where the attacker has full knowledge of the model and the features we deploy. We focus on two strategies that the attacker can follow: modifying Cobalt Strike’s C&C configuration, and altering the C&C flows by other means (e.g. by changing the network stack on the infected machines).

a) Changing the appearance of the C&C sessions using Cobalt Strike

As our model detects C&C sessions maintained using Cobalt Strike, we first analyze the options this framework provides to alter their appearance. The two main parameters the Red Team can use during the exercise are the sleep-period and jitter of a C&C session. The sleep-period defines the time interval used to periodically contact the C&C server. The jitter configures the deviation from this periodicity. Our features are invariant to both of these parameters, as they focus on timing statistics within single connections and do not depend on the time elapsed between the periodic connections of a C&C session.

Cobalt Strike's Malleable C2 tool [16] allows the custom design of the HTTP headers of the packets exchanged within C&C sessions to avoid detection. However, our model does not rely on features extracted from HTTP headers.

We conclude that bypassing detection of our model by altering Cobalt Strike's C&C configurations is infeasible as our features are invariant to the options the framework provides.

b) Identifying attack vectors for manipulating feature values

Our classifier identifies flows that look like Cobalt Strike C&C channels. To avoid this, an attacker might attempt to camouflage these C&C flows as normal traffic for the given network.

We observe that most of the feature values can be altered either by injecting additional packets (to manipulate statistics such as inter-arrival time or packet counts) or by altering the packet sizes (which affects features such as the download size). Many of these tampering attempts could be prevented by additional checks in the feature extraction phase (e.g. sequence number checking for packet injections). However, since this is computationally expensive, we assume that the defenders cannot do this.

To simulate the robustness of our model in such scenarios, we conducted experiments involving tampering with the feature values, as described in Section 4.D.

4. EVALUATION

In this section, we evaluate our classifiers based on data recorded by the Swiss Blue Team from Locked Shields 2017 and 2018. After providing more details about the methodology (Subsection A), we evaluate precision and recall (Subsection B), runtime (Subsection C), robustness against camouflaging (Subsection D) and incomplete traffic captures (Subsection E).

A. Methodology

In this section, we summarize the datasets that we used for the evaluation, the environment in which we conducted the experiments and the parameters that we used.

1) Datasets

To evaluate the performance of our models, we used the complete LS17 dataset for training and the LS18 dataset for testing and vice versa. Therefore, our evaluation corresponds to a case where our classifier is used for classifying previously unseen data in a different network. In the following, we will refer to models trained on the full LS17 or LS18 datasets as LS17-models and LS18-models, respectively (cf. Table III).

In Table II, we summarize the baseline information about the datasets that we used for the evaluation.

TABLE II: BASELINE INFORMATION ABOUT THE DATASETS USED.

Dataset	Size	Packets	Flows	C&C Flows
LS17	114 GB	288'940'662	9'070'828	1'239'041 (13.7%)
LS18	216 GB	557'783'930	16'379'346	1'818'006 (11.1%)

2) Environment

We conducted all experiments and calculations on a virtual machine running Ubuntu 16.04 (64 bit), with 10 Intel Xeon E5-2699 cores and 16 GB RAM. The implementation was based on Python 3.6 and scikit-learn (0.19.2) [17].

3) Parameters and Models

We evaluated two configurations of our classifier: one with the default scikit-learn parameters [15], and the other with the tuned parameters described in Section 3.E. We refer to these configurations as "baseline" and "tuned" and summarize them in Table III. We trained all models using the 20 features obtained from the recursive feature elimination scheme described in Section 3.D.

TABLE III: CHARACTERIZATION OF MODELS USED IN OUR EVALUATION.

Model	Training data	Testing data	RF size	RF depth
LS17-baseline	LS17	LS18	10 trees	unconstrained
LS17-tuned	LS17	LS18	128 trees	10
LS18-baseline	LS18	LS17	10 trees	unconstrained
LS18-tuned	LS18	LS17	128 trees	10

B. Precision/Recall

We used widespread metrics precision (i.e., the percentage of reported C&C flows that are actual C&C flows) and recall (i.e., the ratio between the correctly identified C&C flows and all the C&C flows present in the dataset) to measure the prediction quality of our models. High precision is particularly important in the given task because a high number of false positives would mislead the defenders during their operation.

Table IV lists the precision and recall scores for all models. We repeated the evaluation ten times with different random seeds to train the models, and we report the medians of the results. The results show that all models achieve high precision and recall while tuned configuration clearly outperforms the baseline configuration.

TABLE IV: THE TUNED MODELS ACHIEVE HIGH PRECISION AND RECALL (MEDIANS)

Model	Precision	Recall
LS17-baseline	0.94	0.98
LS17-tuned	0.99	0.98
LS18-baseline	0.98	0.86
LS18-tuned	0.99	0.90

C. Runtime

In this experiment, we evaluate the runtime of three phases:

1. Extracting features from the training dataset
2. Training the model
3. Applying the model on the testing dataset

In Table V, we report the time it takes to extract features from both datasets (using CICFlowMeter). We note that the feature extraction tool extracts all 77 features from Table I. The runtime could be significantly improved by calculating only the 20 selected features and by using a more efficient implementation.

TABLE V: FEATURE EXTRACTION TAKES LESS THAN 45 MIN (LS17) AND LESS THAN 90 MIN (LS18) FOR DATASETS CONTAINING ABOUT 38 HOURS OF NETWORK TRAFFIC.

Dataset	Runtime	Extracted Flows
LS17	42 min	9'070'828
LS18	85 min	16'379'346

In Table VI, we report the time it takes to train and test the model on both datasets. As above, we point out that the training phase is not time-critical as it is done prior to the exercise. As the results show, running predictions on the whole dataset takes less than one minute. In a practical deployment, the inference would be performed on much smaller sets of samples, which makes real-time detection feasible.

TABLE VI: TRAINING AND INFERENCE TIMES OF THE BASELINE AND TUNED MODEL CONFIGURATIONS

Model	Training time	Inference time
LS17-baseline	120 s	6 s
LS17-tuned	1117 s	50 s
LS18-baseline	390 s	4 s
LS18-tuned	2828 s	30 s

D. Robustness Against Camouflaging

In this experiment, we simulate an attacker attempting to camouflage C&C flows as normal traffic. To model an attack against a particular feature, we replace the feature values in the malicious samples (i.e. the C&C flows) with values randomly subsampled from benign samples. As a result, this feature no longer helps in distinguishing C&C flows from normal flows.

In Figure 3 (LS17) and Figure 4 (LS18), we plot the precision and recall of the respective models depending on the number of tampered features. The results hold under the assumption that an attacker that attacks n features would target the n most relevant features according to Section 3.D. (which is a promising strategy). We evaluate the impact of tampering with 5 to 14 features on each model with 10 different random seeds and plot the median as well as the 95% confidence interval.

The results show that the tuned model reacts much less sensitive to camouflaging attempts and achieves high performance even if many features are tampered with (precision falls below 90% when manipulating >12 features). Recall of the LS18 model drops sharply when attacking more than 5 features, however, its precision remains high, meaning that the predictions the model makes are still reliable. Further, we observe that the variance among the tuned models is much lower than that of the baseline models.

FIGURE 3. ACHIEVED PRECISION AND RECALL FOR LS17 IF AN ATTACKER TRIES TO CAMOUFLAGE C&C FLOWS. OUR TUNED MODEL IS ROBUST AGAINST TAMPERING, FOR UP TO 10 FEATURES.

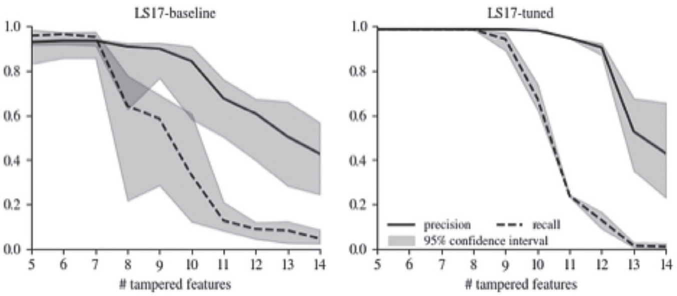
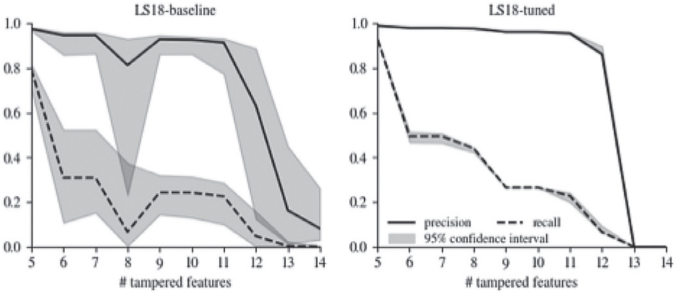


FIGURE 4. ACHIEVED PRECISION AND RECALL FOR LS18 IF AN ATTACKER TRIES TO CAMOUFLAGE C&C FLOWS. OUR TUNED MODEL ACHIEVES A HIGH PRECISION EVEN IF 12 FEATURES ARE ATTACKED BUT THE RECALL DROPS.

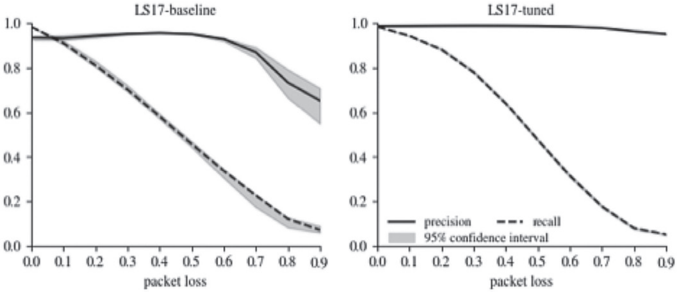


E. Robustness Against Packet Loss

In this experiment, we evaluate the impact of packet loss, which could occur due to the limited resources of the defenders to capture packets in real time during the exercise. We simulate this by randomly dropping between 10 and 90 percent of the packets.

The results in Figure 5 show that the tuned models achieve high precision (> 95%) even for 90% packet loss. This means that even for high losses, the raised alerts stay accurate. However, the recall decreases approximately linearly with the packet loss. Presumably, this is because C&C flows with too many dropped packets are no longer recognized as such, while the model still detects less affected flows.

FIGURE 5. IMPACT OF PACKET LOSS ON THE LS17-MODEL. THE CURVES SHOW THE MEAN VALUES OVER 10 MEASUREMENTS. IN OUR TUNED MODEL, PACKET LOSS HARDLY IMPACTS PRECISION.



5. DISCUSSION

In this section, we discuss the outcomes of the experiments conducted in this paper as well as details of possible real-world deployments and potential extensions.

A. Identifying C&C Servers

The ability to detect individual C&C flows can obviously be used to identify C&C servers (the destinations of such flows) and compromised hosts (the sources of the flows). In an additional experiment, we observed that running our system for a short time period of 30 minutes at the beginning of the exercise (11am-12 pm in Locked Shields 2018) was enough to identify most of the C&C servers (10 out of 12 listed in the Cobalt Strike reports). We further observed 5 different source IP addresses from the Blue Team's network communicating with these servers, suggesting that these hosts had been compromised at this point in time.

B. Running Multiple Models in Parallel

In this paper, we used datasets from two occurrences of Locked Shields: one to train the model, and the other to test it. In the future, when more datasets are available, we suggest training multiple models and conducting live classification during the exercise on all of them. This would make it even harder for the Red Team to camouflage C&C traffic as benign flows, because it needs to match the features of benign flows in multiple different models (while the features of C&C flows are similar in each model). Performing the inference only slightly increases the computational cost and is thus feasible during the exercise. Since we have data from only two iterations of Locked Shields, we could not evaluate this approach.

C. Practical Deployment for Future Locked Shields Exercises

In order to use our system in the next Locked Shields exercise, a Blue Team needs to perform three steps:

1. Train one or multiple models with labeled data from past exercises
2. Prepare the VM to record network traffic and compute the features
3. Run the trained models with the recorded features during the exercise

Step 1 is not time-critical and can be done at any time prior to the exercise. To counteract camouflaging attempts by the Red Team, we suggest using data from different years and training multiple models (cf. Section 5.B).

For Step 2, the Blue Team can use any tool to capture the traffic (no payloads required) and calculate the flow features. In our experiments, we used CICFlowMeter; however, more efficient implementations are possible.

Step 3 consists of feeding the extracted features to one or more models. Information about detected C&C flows can be passed to an intrusion alert system used by the defenders to coordinate security responses.

As our evaluation shows, our classifier is able to predict C&C flows with 99% precision and over 90% recall. By evaluating the system on two datasets originating from two different occurrences of Locked Shields (2017 and 2018), we provided strong evidence for the success of a deployment in future exercises on previously unseen data.

In an additional experiment, we simulated a real-case deployment, where we applied our system for a short 30 minutes time interval in the first phase of the LS18 exercise. There, our system unveiled almost the complete C&C infrastructure used by the Red Team (10 out of 12 C&C Servers).

D. Challenges and Deployment in Other Environments

In this paper, we have focused on a very specific use case for C&C detection (Locked Shields, Cobalt Strike). One of the main limitations of supervised-learning-based systems is that while they are highly effective in detecting anomalies that were labeled in the training set, they fail to detect new and unknown attacks. A further challenge is that the distribution of the legitimate background traffic may strongly vary among different networks.

By expanding the training data with more C&C traffic types and including a wider range of legitimate traffic profiles, our approach could be adapted for deployment in other environments. Moreover, data augmentation techniques such as domain randomization – currently applied with great success in the deep learning domain – are other promising paths towards broader generalization. For instance, OpenAI recently developed a human-like robotic hand to manipulate physical objects with unprecedented dexterity [18]. The training was performed solely in a simulated environment, but by randomizing the physical properties in the simulation, the final model generalized well enough to be deployed on a real physical hand. Although our application is very different, the same concepts could be applied to network traffic data to obtain richer training sets leading to more robust detection systems.

6. CONCLUSION

In this paper, we present a system for identifying C&C channels using supervised machine learning. As a typical use case for such a system, we focus on Locked Shields, the world's largest cyber defense exercise. Our evaluation shows that the system could be deployed by defenders in this exercise and that it identifies C&C traffic with high precision and recall. We use real data from one participating Blue Team and show that if this team had trained the classifier with the data from 2017, it would have identified C&C channels in Locked Shields 2018 with 99% precision and 98% recall. Further, running the system during a time interval of just 30 minutes in LS18 would have been enough to identify 10 out of 12 C&C servers used by the Red Team.

Acknowledgments

We thank the Swiss Blue Team for sharing their data and expertise with us and for their constant support throughout this project.

REFERENCES

[1] "How a British SMB survived a nightmarish cryptolocker ransom attack | Security | Computerworld UK," [Online]. Available: https://www.computerworlduk.com/security/how-british-smb-survived-nightmarish-cryptolocker-ransom-attack-3677593/.

[2] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis and others, "Understanding the mirai botnet," in *USENIX Security Symposium*, 2017.

[3] "Locked Shields 2017," [Online]. Available: https://ccdcoe.org/ locked-shields-2017.html.

[4] A. H. Lashkari, G. D. Gil, J. E. Keenan, K. Mbah and A. A. Ghorbani, "A Survey Leading to a New Evaluation Framework for Network-based Botnet Detection," in *Proceedings of the 2017 the 7th International Conference on Communication and Network Security*, 2017.

[5] M. Feily, A. Shahrestani and S. Ramadass, "A survey of botnet and botnet detection," in *Third International Conference on Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09*.

[6] 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.

[7] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou and D. Dagon, "Detecting Malware Domains at the Upper DNS Hierarchy," in *USENIX security symposium*, 2011.

[8] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda and C. Kruegel, "Disclosure: detecting botnet command and control servers through large-scale netflow analysis," in *Proceedings of the 28th Annual Computer Security Applications Conference*, 2012.

[9] "CCDCOE News (26 April 2018)," [Online]. Available: https://ccdcoe.org/more-1000-cyber-experts-30-nations-took-part-locked-shields.html.

[10] "Cobalt Strike Reporting," [Online]. Available: https://www.cobaltstrike.com/help-reporting.

[11] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun and A. A. Ghorbani, "Characterization of Encrypted and VPN Traffic using Time-related," in *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, 2016.

[12] "The Bro Network Security Monitor," [Online]. Available: https://www.bro.org/.

[13] "CICFLOWMETER A network traffic Biflow generator and analyzer," [Online]. Available: http://www.netflowmeter.ca/.

[14] G. Louppe, L. Wehenkel, A. Sutera and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, 2013.

[15] "sklearn RandomForestClassifier," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[16] "Cobalt Strike 3.11 Manual," [Online]. Available: https://www.cobaltstrike.com/downloads/csmanual311.pdf.

[17] "scikit-learn Machine Learning in Python," [Online]. Available: https://scikit-learn.org.

[18] OpenAI, "Learning dexterous in-hand manipulation," 2018.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Neural Network-Based Technique for Android Smartphone Applications Classification

Roman Graf
Austrian Institute of Technology GmbH
Vienna, Austria
roman.graf@ait.ac.at

L. Aaron Kaplan
CERT.AT
Vienna, Austria
kaplan@cert.at

Ross King
Austrian Institute of Technology GmbH
Vienna, Austria
ross.king@ait.ac.at

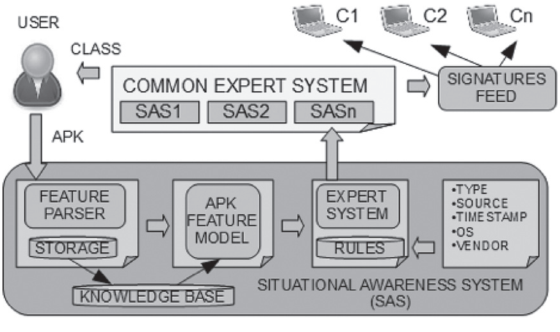
Abstract: With the booming development of smartphone capabilities, these devices are increasingly frequent victims of targeted attacks in the 'silent battle' of cyberspace. Protecting Android smartphones against the increasing number of malware applications has become as crucial as it is complex. To be effective in identifying and defeating malware applications, cyber analysts require novel distributed detection and reaction methodologies based on information security techniques that can automatically analyse new applications and share analysis results between smartphone users. Our goal is to provide a real-time solution that can extract application features and find related correlations within an aggregated knowledge base in a fast and scalable way, and to automate the classification of Android smartphone applications. Our effective and fast application analysis method is based on artificial intelligence and can support smartphone users in malware detection and allow them to quickly adopt suitable countermeasures following malware detection. In this paper, we evaluate a deep neural network supported by word-embedding technology as a system for malware application classification and assess its accuracy and performance. This approach should reduce the number of infected smartphones and increase smartphone security. We demonstrate how the presented techniques can be applied to support smartphone application classification tasks performed by smartphone users.

Keywords: *Cyber security, neural network, AI, smartphone, malware*

1. INTRODUCTION

The Android operating system for tablets, phones and smart devices is by far the most widespread mobile operating system in the world, with millions of active devices. Millions of new malware programs have been released for this platform in recent years. The market share of exploits that target the Android platform makes it the second most targeted platform for running exploit attacks. The goal of this paper is to train a neural network to evaluate the discoverability and explainability of upcoming attack patterns. Classification capabilities of neural networks are heavily reliant on the quality of the underlying datasets, and subsequently even more dependent on the granularity of extracted features. The presented technique (see Figure 1) will apply deep neural networks and supervised learning to evaluate the capabilities of detecting smartphone malware applications in Android. Currently there is a lack of technology supporting an integrated solution of large-scale feature extraction and neural network training. The goal of this approach is to release an open source framework that provides integrated functionality along the required workflow. This workflow comprises application source code extraction, feature composition, neural network training and analysis of results. The components of this system are executed at scale within Hadoop and GPU clusters. The platform supports publishing of the harvested ground-truth dataset, the extracted features and the trained neural network on an open data platform. To visualize the projects results and to raise awareness for malware applications prevention in the general public, a demonstrator was developed that allows live inspection of the trustworthiness of Android applications.

FIGURE 1. THE OVERVIEW OF ESTABLISHING THE CYBER SITUATIONAL AWARENESS USING NEURAL NETWORK FOR APK CLASSIFICATION.



The neural network approach is widely used for different analytical tasks. A machine learning framework based on word-embedding techniques can be used for the classification of text files. Standard machine learning algorithms are incapable of processing strings or plain text in their raw form; rather, they require numbers as inputs to perform any type of calculations. In the word-embedding approach, words are mapped to numerical vectors. The difference from other language processing methods is that the embedding vector also keeps the context of the word in a sentence or file. This improves the overall accuracy of the prediction model, compared to simple counting of words in a file. Our approach provides a numerical representation of contextual similarities between Android Package (APK) features extracted in text format. Each feature is represented by a real-value vector with tens or hundreds of dimensions. In contrast, other methods, such as a one-hot encoding, employ thousands or millions of dimensions required for sparse word representations.

This paper is structured as follows. Section 2 gives an overview of related work and concepts. Section 3 explains the APK classification workflow including the feature extraction method and neural network training for APK classification. The expert system issues and related rule engine are covered in Section 4. Section 5 presents the experimental setup, applied methods end evaluation. Section 6 presents our conclusions.

2. RELATED WORK

The design of the presented framework is inspired by the DREBIN project (Rieck, 2004; Hoffmann, 2013), which combines a broad static analysis of gathered smartphone application features and applies machine learning for identifying patterns that are indicative of malware. The manifest and decompiled dex (Dalvik Executable) codes are scraped to extract feature sets and DREBIN utilizes a linear SVM algorithm (Shawe-Taylor, 2000), which assumes real-value inputs. The manifest file provides features such as requested hardware components and system granted permissions, declared components such as services or broadcast receivers and filtered intents which are used for inter-process communication. By analysing the disassembled bytecode, additional "hidden" features are gathered, such as restricted API calls, actually used permissions, calls to sensitive resources (e.g. frequently used for obfuscation) and a list of all network addresses. This demonstrated approach provides both effective detection rates and explainable results and was able to outperform related approaches as well as 9 out of 10 popular anti-virus scanners with a detection rate of 94% and a false positive rate of 1%, and reliably detect all malware families except Gappusin. DREBIN showed the importance of the different features sets and that their proper composition can lead to reliable and explainable detection results using neural networks and machine learning. While the methodology is well-documented, and the collected corpus of 120 thousand apps (including 22% malware samples) is published for academic re-use, the corpus itself is outdated (SDK level 12) and the DREBIN framework and neural network itself are closed-source. Furthermore, DREBIN is highly restricted in its learning-based detection capabilities as the project targeted the smartphone as runtime and detection environment where such a dataset must be heavily maintained and updated. The SVM approach is limited by the choice of the kernel, which is a general weak point of SVM applications. Alternative algorithms employing categorical features and labels are Naive Bayes (Schütze, 2008), Logistic Regression (Cox, 1958) and Random Forests (Ho, 1995). Approaches based on decision trees such as Random Forests are very fast to train, but quite slow to create predictions once trained. A higher degree of accuracy requires additional trees, which means losing performance. Naive Bayes often serves as a robust method for data classification, but the vectors representing incident in Naive Bayes are larger than in word-embedding methods and also Naive Bayes classifiers make a very strong assumption on the shape of the data distribution. Further problems may result due to data scarcity, which can result in probabilities going towards 0 or 1, leading to numerical instabilities and worse detection results. Logistic regression like a Naive Bayes method requires that each feature in an incident is independent from all other features. Logistic regression models are also vulnerable to overconfidence as a result of sampling bias.

A brief overview of related approaches for the detection of Android malware lists some comparable methods for this task. Kirin (McDaniel, 2009) checks application permissions, Stowaway (Wagner, 2011) analyses API calls to detect overprivileged applications and RiskRanker (Jiang, 2012) identifies applications with different security risks. However, none of these approaches includes multiple features sets or features received from reverse-engineering the applications' source code, elements that were proven crucial for the detection results in DEBRIN. Open source tools such as Smali2 and Androguard3 enable dissecting the application's content for subsequent feature extraction and are evaluated for their use within framework's extraction pipeline. The dedicated analysis system DroidScope (Droidscope, 2012), which enables introspection at different layers of the Android platform, allows users to dynamically monitor applications in a protected environment at runtime. Methods of sandboxing try to mimic a real-world environment and aim to discover malicious behaviour but are limited due to sophisticated obfuscation methods used in modern malware. ParanoidAndroid (Bos, 2010) creates a virtual clone of the smartphone that runs in parallel on a dedicated server and synchronizes with the activities of the device. This configuration allows for monitoring the behaviour of applications on the clone without disrupting the functionality of the real device, but the resources required for a large number of devices are often not technically feasible. Dynamic analysis tools, such as DroidRanger (Jiang Y. Z., 2012) are suitable for filtering malicious applications from Android markets.

Dedicated open source frameworks in the domain of malware detection are rare. A prominent but outdated system is MobileSandbox (Hoffmann, 2013), which is designed to automatically analyse Android applications by combining a static and dynamic approach, which for example allowed the analyst to log system calls to native APIs. MobileSandbox provides a highly complex and immature system due to the enormous integration effort and customizations required for the Davlik virtual machine and emulators.

The advantage of the embedding method is that it not only takes into account features such as count and word context, but also learns automatically from examples. The autoencoder (Cheng-hua, 2008) makes use of neural networks, which are already in use by latent semantic analysis for text categorization to reduce dimensionality and to improve performance; but this method has the disadvantage of not using the context of the feature. Another application (Lee, 1999) employs an artificial neural network to improve text classifier scalability.

Classification methods implemented in these threat intelligence tools suffer from large vector sizes and are less effective as the number of features rises. The main drawback of existing text classification methods such as SVM or the Gensim tool is that they require a huge database for training to provide meaningful results. Another common disadvantage of these techniques is the lack of result 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 words that were not previously included in the training vocabulary.

Multiple researchers are developing an automated technology that will support an information classification system. An attempt to classify the relationships between documents and concepts employs principles of ontology. Currently, APKs can be classified based on the features included in the package and in source code. Contrary to this approach, we classify not only by data extracted from APK that can differ from dataset to dataset, but we also employ additional rules implemented in an expert system and take in account APK source, type, timestamp, dataset and other parameters. This technique provides more accurate prediction.

Neural networks with word-embeddings in general also require large training datasets, but for APK classification, taking into account the fact that we have multiple different datasets, we will train multiple models for each dataset and additionally employ a rule engine to produce accurate results compared to the case if we would just train one huge model ignoring intrinsic differences in the datasets. Consequently, for the particular use case of APK classification task, we suggest using the word-embedding neural network solution that scales well because of the split-models concept and supportive rule engine, while maintaining a high level of accuracy. Our goal is to make a more accurate prediction for a specific dataset, employing the whole aggregated expert knowledge and applying expert rules.

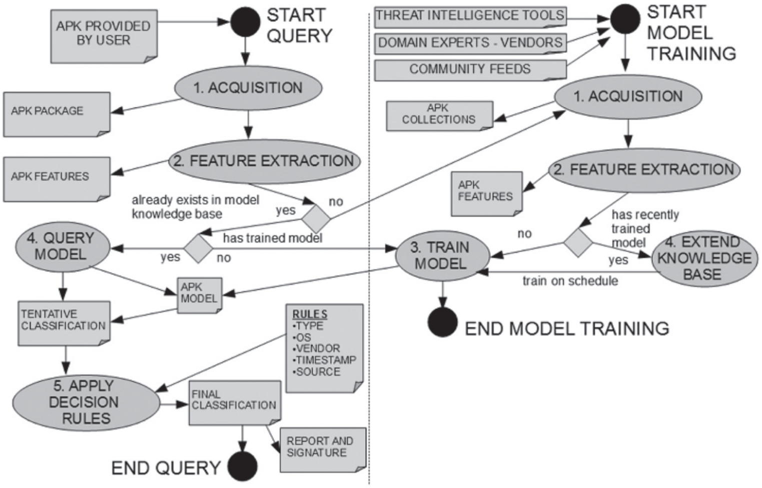
3. APPLICATION CLASSIFICATION METHOD

APK classification employs application features extraction and training of neural network to produce a model for the queries. Deep Learning is employed for learning in neural networks and describes a subset of machine learning algorithms that deal with accurately assigning weights across many neural network layers. Three main types of machine learning can be distinguished: Supervised, Unsupervised and Reinforcement Learning. Supervised learning can solve classification problems. Classification predicts previously defined categories for a given sample. In the case of Android malware these categories are binary: “benign” or “malware”. Supervised learning employs labelled training data to learn mapping functions from a given input (embedding vector in our case) to a desired output value. A supervised learning algorithm analyses the data through weights and activation functions that activate neurons and produces an inferred function, which is then used for mapping new samples or correctly determining classification labels for unseen instances.

A. Application Classification Using Neural Networks

Figure 1 provides an overview of establishing the cyber situational awareness using neural networks for APK classification. This approach is based on a knowledge base containing large number of labelled smartphone applications. This data can be provided by different vendors, collected at different times for particular operating systems, and may be separated by type of application. Therefore, for each use case (Situational Awareness System – SAS) we propose to have a separate expert system and associated decision rules. All such SAS systems are then aggregated in a common expert system, which performs final classification. A user uploads their APK package to the SAS. The system extracts features from this package, stores them for further analysis and queries an APK model that was trained based on knowledge base. The final classification result in the form of a report and signatures is disseminated by means of a signatures feed for subscribed clients C1-Cn.

FIGURE 2. THE WORKFLOW FOR FEATURE EXTRACTION AND CLASSIFICATION OF APK USING A NEURAL NETWORK APPROACH.



To employ the embedding method, features aggregated in text form must be converted into numerical values, since machine learning algorithms and deep learning architectures cannot process plain text. Therefore, each uploaded APK (see Figure 2) is converted into an array of strings, where each string represents a particular feature. Then strings are encoded by indices, and each feature string has a unique index. If this feature repeats in the APKs, we re-use its index. Finally, arrays of indexes are converted in one-hot encoded vectors, meaning that the position of each feature in the original feature set is encoded using “1” if a feature exists in the given place or “0” if not. After defining the number of latent factors expressed in the length of the embedding vector, we convert produced on-hot vectors into embedding vectors, giving an array of float numbers. Therefore, we create a list of embedding sequences for each APK with embedding vector representation of each feature. Embedding vectors are an input to the neural network.

The neural network is composed of an input embedding layer, a flattening layer and two hidden layers, where the model will be trained to classify APKs as either “benign” or “malware.” The flattening layer is required to enable a connection between the dense and embedding layers. We flatten the two-dimensional output matrix of the embedding layer (with one embedding for each feature in the input sequence of features) to a one-dimensional vector used by the dense layer.

B. Application Features Extraction

The workflow process is composed of two parts. One process is a neural network model training, where workflow acquires APK data from different sources such as community feeds, threat intelligence tools and domain experts, which are vendors or anti-malware producers. The model is trained and regularly updated by extended knowledge from new APK collections. The query workflow execution begins with reading a smartphone application package (see step 1 in Figure 2) provided by a user and parsing the extracted content for features extraction in step 2. For the acquisition computation we employ a parsing method developed by researchers who reimplemented the DREBIN parsing method described in the DREBIN paper (Rieck, 2004). By means of extracted features, we obtain an APK vector. If the given APK is not in the model, we additionally extend the model knowledge base for subsequent training. In the next step we train the APK model using a neural network (step 3) or a query trained model in step 4, applying the created feature vector. The model responds with a tentative classification. Finally, we calculate the APK classification employing an expert system and the decision rules in step 5. These rules comprise decision logic and expert profile settings that are specific for an organisation. Factors such as APK type, operating system, vendor, creation time and origin have an impact on the resulting decision. At the end we provide a report accompanied by an APK signature.

4. EXPERT SYSTEM

Table 1 lists the layers that are employed in the neural network, including their type, activation function, size and parameter number.

TABLE 1: DEPENDENCY CHART WITH INTERACTIONS AMONG THE RULES AND ASSOCIATED IMPACT FACTORS.

Rules/Actions	Install	Remove	Ignore	Alarm	Quarantine	Clean	Log
Neural network model classification (benign/malware)	+	+	+	+	+		+
Metadata (has conclusive feature)	+	+	+	+	+	+	+
File size (large/small)	+	+	+				+
File name (known/unknown)	+	+	+	+			+
Malware signature (known/unknown)	+	+	+	+	+	+	+
Operating system (Android/Mac OS/Win)	+	+	+				
Vendor (trusted/not trusted)	+	+	+	+	+		+
Type (game/office...)	+	+	+				
Source (trusted/not trusted)	+	+	+		+		+
Creation time (old/new)	+	+	+				+

To organize the knowledge base (see Figure 1), we must structure the information that has been obtained from the domain experts of APK domain and from conducted experiments.

We aim to achieve the following objectives:

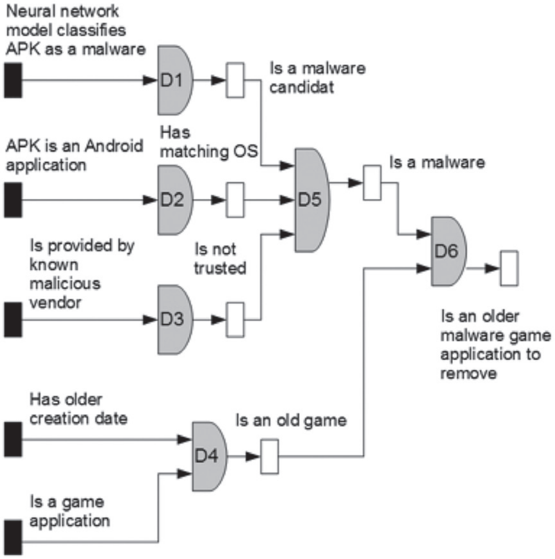
1) Define typical scenarios for smartphone application handling;
2) Identify the parameters used by cyber experts for APK handling;
3) Define the linguistic labels that are used by the experts to classify measured values of each parameter and identify the range of each label when possible; and
4) Define typical scenarios for determining the conditional rules that relate these linguistic labels to specific control actions.

Knowledge acquisition for the knowledge base occurs through the domain expert. In our case, these are cyber analysts and SOC operators who provide the knowledge with typical application use cases, metrics and parameters that characterize the APK analysis processes. Information retrieved from the APK packages is processed by the customized domain model. This model enables structured and maintainable handling of analysed data and its storage in a database for further treatment. Inferred data is processed in an inference engine by rules application in order to provide the rationale for a particular analysis action. A user communicates with the expert system using GUI by sending a request query and receiving an advice in response.

The development of a knowledge base is an iterative process. Knowledge can be encoded, tested, added, updated and removed. Potential problems with rule definition and coverage are redundant rules, conflicting rules, rules that are subsumed by other rules, unreachable rules, inconsistent rules and circular rules chains. In order to avoid the rule-based systems faults described by Arman (2007), we generated a dependency chart that shows the interactions among the rules (Nguyen, 1985). The dependency chart presented in Table 1 gives an overview of the identified rules and associated impact on the knowledge base. The dependency chart helps to find potential rules problems and to keep an overview of the rules.

Among the most important rules (see Figure 3) are those regarding APK issues, like “neural network model classification”, “metadata”, “file size”, “file name” and “malware signature”. According to the requirements and circumstances for a particular APK, an expert could leverage these rules; for example, if a file name has a semantic meaning or if file size is of interest for analysis. Sometimes metadata contains important and useful information. The “malware signature” rule becomes significant in the case of known malware signature in an application source code. The issue of “type” means that APK has significant risk if it belongs to particular type of application e.g. gambling. The issues “vendor” and “source” could have higher severity if the actors are known for producing malicious APKs.

FIGURE 3. AN EXAMPLE SELECTION OF FORWARD RULE CHAINING FOR SMARTPHONE APPLICATIONS CLASSIFICATION.


Neural network model classifies APK as a malware
D1
Is a malware candidat
APK is an Android application
Has matching OS
D2
D5
Is a malware
Is provided by known malicious vendor
Is not trusted
D3
D6
Is an older malware game application to remove
Has older creation date
Is an old game
D4
Is a game application

Rules are associated with related actions. Table 1 gives an overview of these relations. Upon the provided inputs, the rule engine can trigger actions, such as “install”, “remove”, “ignore”, “alarm”, “quarantine”, “clean” or “log” the given APK. The “clean” action is the most challenging and supposes an attempt to remove malware from the APK, which is applicable only by a high value of APK. Other actions are self-descriptive.

The previously defined rules should be organized in order to process input statements (assertions) and to infer appropriate action and conclusions. A process of the forward rule chaining for APK collection is presented in Figure 3. It is a process of moving from the “if” patterns (antecedents) to the “then” patterns (consequents) in a rule-based reaction system. We consider the antecedent as satisfied when “if” pattern matches the assertion. Assertions are depicted in the figure as the black rectangles on the input side and as the white rectangles on the output side. The rules are presented in the form of blue semi-circles (*R1-Rn*). The rule is triggered if all the antecedents are satisfied. A triggered rule is considered as fired if it produces a new assertion or performs an action as output (white rectangle in the figure). Since our expert system is presently focused on APK collections, we do not need a conflict-resolution procedure to resolve possible rules conflicts. Managing dependencies, as depicted in Table 1, reduces the risk of conflicting rules and lists rules required to distinguish malware from other applications. A variable x acquires value as antecedent pattern is matched to assertion. For example, using information from well-established and reliable "FDroid tool", rule *R3* determines that an application stems from known malicious source:

R3: If ?x is provided by known malicious source
Then ?x is not trusted

The rule-based system starts APK classification with the rule *R1*. Suppose that particular APK was classified by the neural network model as malware. Then if the antecedent pattern "*?x is a malware*" matches that assertion, the value x becomes "*is a malware candidate*" and rule *R1* fires. Because application is an Android APK, rule *R2* fires, establishing that the document "*has matching OS*". Rule *R3* fires with the value "*is not trusted*". If two input assertions match an antecedent pattern, rule *R4* fires. The output assertions of the first three rules become the input assertion for the rule *R5* and if there is a match to the antecedent pattern the rule fires with the value "*is a malware*". Finally, if the input assertions of rule *R6* match, the rule fires with resulting action "*is an older malware game application to remove*".

The output of the rule-based system is a conclusion for a malware classification. The classification of the given APK is calculated based on the features of the associated APK. The inference engine performs conditional rules and classification analysis, infers appropriate action and formulates advice using relation of linguistic labels to specific control actions.

5. EXPERIMENTAL EVALUATION

A. Evaluation Data Set

The experimental dataset with ground-truth labels was provided by firms I and C and processed on an ABC server, which comprises Hadoop and GPU clusters. We split samples into test (5,640), validation (5,076) and training sets (45,676). For feature extraction we employ APK feature extractor described on a research site[1] and reimplemented on GitHub.[2]

B. Experimental Results and Interpretation

Classification of APK samples into benign and malware was evaluated employing techniques described in the previous sections. Features were extracted from APK

[1] https://www.sec.cs.tu-bs.de/~danarp/drebin/

[2] https://github.com/MLDroid/drebin

packages and converted in embedding vectors. Here is an example selection of loaded features:

- UsedPermissionsList_android.permission.VIBRATE
- UsedPermissionsList_android.permission.ACCESS_NETWORK_STATE
- UsedPermissionsList_android.permission.INTERNET
- BroadcastReceiverList_com.google.android.apps.analytics.AnalyticsReceiver
- SuspiciousApiList_Landroid/content/Context.getSystemService
- SuspiciousApiList_Landroid/app/Activity.getSystemService

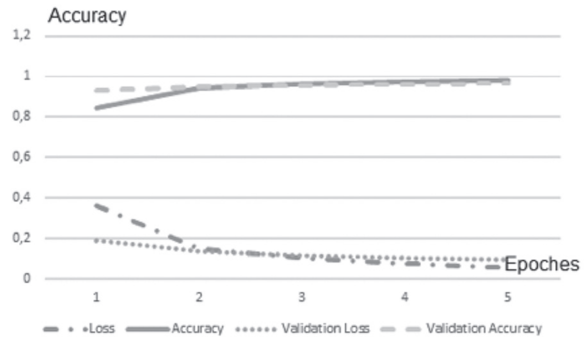
Embedding vectors describing loaded features were used as an input to a neural network. Table 2 lists the layers employed in the neural network including their type, activation function, size and parameter number. The total number of parameters used in the input and hidden layers during the training was 19,546,001. We employed an embedding approach for the input layer and sigmoid activation function for the dense layer. The total training time was 29,578 seconds. The model parameter settings for this particular training is presented in the fifth row in Table 3.

TABLE 2: SUMMARY OF THE NEURAL NETWORK TRAINING PROCESS.

Layer	Type	Activation Function	Size	Parameters #
Input layer	Embedding		200x30	19,245,900
Hidden layer 1	Flatten		6,000	0
Hidden layer 2	Dense	Sigmoid	50	300,050
Hidden layer 3	Dense	Sigmoid	1	51

Figure 4 visualizes the training results. We can see that, in general, the model training accuracy improves with every iteration (epoch) from 0.845 at the beginning to 0.982 at the end, which is sufficiently good; whereas training loss (error) of original information decreases from 0.362 to 0.056. This means that the outputs will be degraded compared to the original inputs, but it is an acceptable rate. Similarly, validation accuracy is in the range between 0.931 and 0.966. Validation loss decreases from 0.191 to 0.096.

FIGURE 4. ACCURACY AND LOSS CHARACTERISTICS BY NEURAL NETWORK TRAINING.



C. Evaluation Effectiveness

Table 3 shows the impact of parameter tuning on the neural network output and accuracy.

TABLE 3: IMPACT OF PARAMETER CHANGING ON NEURAL NETWORK OUTPUT AND ACCURACY.

LR	MVL	EVL	Time	TL	TA	VL	VA	NNA	TP	FP	FN	TN	TPR	FPR
0.001	200	30	1,287	0.0050	0.9989	0.1168	0.9663	99.936	2,880	82	99	2,659	96	2
0.01	200	30	3,231	0.0082	0.9980	0.1398	0.9639	99.875	2,719	163	44	2,714	98	5
0.0001	200	30	31,769	0.0432	0.9866	0.0919	0.9697	98.885	2,761	121	64	2,694	97	4
0.0001	100	30	31,335	0.0497	0.9840	0.0902	0.9675	98.791	2,761	121	64	2,694	97	4
0.0001	50	30	29,578	0.0563	0.9819	0.0961	0.9657	98.640	2,773	109	85	2,673	97	3
0.001	50	30	28,852	0.0047	0.9989	0.1446	0.9547	99.927	2,824	58	147	2,611	95	2
0.01	50	30	5,843	0.0107	0.9973	0.1381	0.9618	99.877	2,783	99	97	2,661	96	3
0.01	100	30	3,836	0.0094	0.9974	0.1465	0.9675	99.840	2,709	173	40	2,718	98	5
0.001	100	30	3,957	0.0047	0.9988	0.1256	0.9667	99.796	2,812	70	114	2,644	96	2
0.001	200	20	16,673	0.0045	0.9988	0.1395	0.9665	99.873	2,764	118	54	2,704	98	4
0.001	200	10	496	0.0055	0.9986	0.1373	0.9565	99.811	2,823	59	156	2,602	94	2
0.01	50	10	559	0.0049	0.9988	0.1638	0.9636	99.859	2,754	128	79	2,679	97	4
0.2	5	5	951	0.1784	0.9533	0.2922	0.9033	95.262	2,316	566	92	2,666	96	17
0.5	5	5	1217	0.2898	0.9156	0.3325	0.8936	91.925	2,138	744	99	2,659	95	21

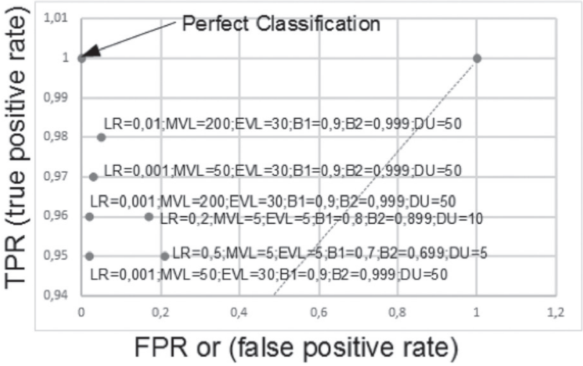
Multiple factors can impact characteristics of the neural network model; some of them are depicted in Table 3. These factors are optimization algorithm, maximal length of the embedding vector, dense units number, activation functions, number of training epochs, learning rate (LR), maximal vector length (MVL) and embedding vector length (EVL). The characteristics of the model are training loss (TL), training accuracy (TA), validation loss (VL), validation accuracy (VA), total training accuracy (NNA), time in seconds, number of hidden layer parameters and classification accuracy expressed in receiver operating characteristic (ROC) points. Some well-known and established default settings for tested neural network problems were applied in our evaluation. As an optimization algorithm for the learning model we selected "Adam", which is an extension to stochastic gradient descent that is widely adopted for deep learning applications in natural language processing. This method differs from standard stochastic gradient descent by changing the learning rate during training. This algorithm can be tuned using parameters such as: "alpha", learning rate (0.001); "beta1", the exponential decay rate for the first moment estimates (0.9); "beta2", the exponential decay rate for the second-moment estimates (0.999); "epsilon", a very small number to prevent any division by zero in the implementation (1E-8); and "decay", the learning rate decay over each update (0.0).

During the APK's classification calculation using the neural network, there was a minor fluctuation of accuracy value (between 95.65 and 99.36). This is because the model employs a random weights initialization. Therefore, it is possible that the highest level of accuracy can be achieved with different parameter configurations. In the test scenario, we investigated the provided test APK collection to classify those applications by threat level (malware or benign) without involvement of a human analyst. Due to the large number of possible configurations in Table 3, we describe only the selected configurations, which demonstrate typical cases. LR is presented in the first column. The second column shows the MVL of the extracted features. In the third column, we show the length of embedding vector. Column "time" depicts the time required to train a model with the given parameter settings. The next five columns are related to the model training process and show training and validation accuracy and error. The final six columns show ROC values to assess evaluation accuracy based on labelled training dataset.

The figure shows that the most productive settings for highest accuracy (up to 99.93) are LR=0.001, MVL=200, EVL=30, whereas "LR" and "MVL" are dominating. For a given training collection, the most accurate classification (TPR=97, FPR=3) was achieved by LR=0.0001, MVL=50, EVL=30. The smallest duration for model training was 496 seconds (LR=0.0001, MVL=200, EVL=10) and the longest operation time was 31,769 seconds with settings (LR=0.0001, MVL=200, EVL=30). This difference can be explained by the different embedding vector sizes. The larger the vector, the longer it takes to calculate the model. This evaluation also gives a simple overview of the detected impact of a particular setting, such as "EVL" for calculation speed, "LR" for learning accuracy and "maximal input vector length" for classification accuracy.

Having evaluated the model for different parameter configurations, we can conclude that a smaller LR provides higher accuracy, while employing more time for calculation. The MVL size has limited impact on the presented model, since APKs comprise a relatively small number of significant features, although the longer the EVL the more accurate the result. To prove that the remaining parameters were selected optimally and that their change would reduce the overall quality and accuracy of the model, in last two measurements presented in Table 3 we additionally reduced the “beta1” and “beta2” parameters (0.8, 0.899 and 0.7, 0.699) and the number of dense units of the activation function to 10 and 5 respectively. The reduced accuracy of the last two results confirms our hypothesis that the noted settings provide the best possible result, thus making model optimization easier. Higher accuracy is also related to the number of training parameters in the dense hidden layers of the model, which ranges between 130 and 1,200,200. The number of these parameters is dependent on all the other aforementioned settings.

FIGURE 5. ROC PLOT OF NEURAL NETWORK TRAINING.



The classification effectiveness can be determined in terms of a Relative Operating Characteristic (ROC) using the labelled ground-truth query dataset. The SA analysis makes use of the separation of the provided APK samples into the two groups “benign” and “malware” provided by domain experts. For example, in the first sample in Table 3, the provided algorithm detected 2,880 TP (True Positive), 82 FP (False Positive), 2,659 TN (True Negative) and 99 FN (False Negative) APKs. The primary statistical performance metrics for ROC evaluation are sensitivity (highest is 0.98) or true positive rate and false positive rate (lowest is 0.02). For the first sample, the associated ROC value is represented by the point (0.02, 0.96). 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 distribution of collection points above the red diagonal demonstrates quite good classification results that could be improved by refining the model settings. Two ROC points with deliberately roughly selected parameters are still situated above the red line, but as expected shifted lower, away from the perfect classification point. The calculation results demonstrate that the calculated classification values for the query APKs are located very close to the labelled classification. These results demonstrate that an automatic approach for APK classification of the method described is very effective and is a significant improvement on manual analysis. Therefore, an analysis method based on neural network technique can be suggested as an effective method for APK 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 APKs, thus making analysis of a large number of APKs 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 Android smartphone applications (APKs) for establishing cyber situational awareness using neural networks. We have combined expertise gathered during the development of methods for application features extraction with the power of the neural network approach and expert system for decision support.

The main contribution of this work is a real-time automatic solution that can classify smartphone applications as either “malware” or “benign” in a fast and effective manner based on a large number of labelled applications, in order to detect malware applications and to secure user devices. The presented method employs a knowledge base collected from domain experts to detect situational awareness risks. Ultimately, our research will lead to the creation of automated security assessment tools with more effective handling of smartphone applications.

REFERENCES

Arman, N. (2007). Fault detection in dynamic rule bases using spanning trees and disjoint sets. *The International Arab Journal of Information Technology, Vol. 4, No. 1*, pp. 67-72. Palestine.

Auria, L. (2008). Support Vector Machines (SVM) as a Technique for Solvency Analysis. *DIW Berlin*.

Bos, H. (2010). Paranoid android: Versatile protection for smartphones. In *Proc. of Annual Computer Security Applications Conference (ACSAC)*.

Cheng-hua, Y. B.-b. (2008). Latent semantic analysis for text categorization using neural. *in Knowledge-Based Systems, 21*, pp. 900-904.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 215-242). Royal Statistical Society, Wiley.

Droidscope, L.-K. Y. (2012). Seamlessly reconstructing os and dalvik semantic views for dynamic android malware analysis. In *Proc. of USENIX Security Symposium*, (pp. 393–407).

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, (pp. 278-282).

Hoffmann, M. S. (2013). MobileSandbox: Looking Deeper into Android Applications. In *Proc. 28th International ACM Symposium on Applied Computing (SAC)*.

Jiang, M. G. (2012). Riskranker: scalable and accurate zero-day android malware detection. In *Proc. of International Conference on Mobile Systems, Applications, and Services (MOBISYS)*, (pp. 281–294).

Jiang, Y. Z. (2012). Hey, you, get off of my market: Detecting malicious apps in official and alternative android markets. In *Proc. of Network and Distributed System Security Symposium (NDSS)*.

Lee, S. L. (1999). Feature reduction for neural network based text categorization. In *Proceedings 6th International Conference on Advanced Systems for Advanced Applications*, (pp. 195-202). Hsinchu.

McDaniel, W. E. (2009). On lightweight mobile phone application certification. In *13 Proc. of ACM Conference on Computer and Communications Security (CCS)*, (pp. 235-245).

Molloy, H. P.-R. (2012). Using probabilistic generative models for ranking risks of android apps. In *Proc. of ACM Conference on Computer and Communications Security (CCS)*, (pp. 241–252).

Rieck, D. A. (2004). *Drebin: Efficient and Explainable Detection of Android Malware in Your Pocket*. 21th Annual Network and Distributed System Security Symposium (NDSS).

Schütze, C. D. (2008). Introduction to Information Retrieval. *Cambridge University Press*. New York, USA.

Shawe-Taylor, N. C. (2000). *An introduction to support vector machines and others*. Cambridge University Press.

T. A. Nguyen, W. A. (1985). Checking an Expert System Knowledge Base for Consistency and Completeness. In *Proc of IJCAI-85*, (pp. 375-378).

Wagner, A. P. (2011). Android permissions demystified. In *Proc. of ACM Conference on Computer and Communications Security (CCS)*, (pp. 627–638).

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Cyber-Physical Battlefield Platform for Large-Scale Cybersecurity Exercises

Joonsoo Kim
Senior Researcher
National Security Research Institute
Daejeon, South Korea
joonsoo@nsr.re.kr

Kyeongho Kim
Senior Researcher
National Security Research Institute
Daejeon, South Korea
lovekgh@nsr.re.kr

Moonsu Jang
Senior Researcher
National Security Research Institute
Daejeon, South Korea
moonsujang@nsr.re.kr

Abstract: In this study, we propose a platform upon which a cyber security exercise environment can be built efficiently for national critical infrastructure protection, i.e. a cyber-physical battlefield (CPB), to simulate actual ICS/SCADA systems in operation. Among various design considerations, this paper mainly discusses scalability, mobility, reality, extensibility, consideration of the domain or vendor specificities, and the visualization of physical facilities and their damage as caused by cyber attacks. The main purpose of the study was to develop a platform that can maximize the coverage that encompasses such design considerations. We discuss the construction of the platform through the final design choices.

The features of the platform that we attempt to achieve are closely related to the target cyber exercise format. Design choices were made considering the construction of a realistic ICS/SCADA exercise environment that meets the goals and matches the characteristics of the Cyber Conflict Exercise (CCE), an annual national exercise organized by the National Security Research Institute (NSR) of South Korea. CCE is a real-time attack-defense battlefield drill between 10 red teams who try to penetrate a multi-level organization network and 16 blue teams who try to defend the network. The exercise platform provides scalability and a significant degree of freedom in the design of a very large-scale CCE environment. It also allowed us to fuse techniques such as 3D-printing and augmented reality (AR) to achieve the exercise goals.

This CPB platform can also be utilized in various ways for different types of cybersecurity exercise. The successful application of this platform in Locked Shields 2018 (LS18) is strong evidence of this; it showed the great potential of this platform to integrate high-level strategic or operational exercises effectively with low-level technical exercises. This paper also discusses several possible improvements of the platform which could be made for better integration, as well as various exercise environments that can be constructed given the scalability and extensibility of the platform.

Keywords: *cyber exercise, cyber conflict, cyber-physical systems, ICS/SCADA testbed*

1. NEED FOR A CYBER-PHYSICAL BATTLEFIELD (CPB) IN LARGE-SCALE CYBER EXERCISES

The purpose of a national cyber security exercise is to assess the national readiness with regard to cyber threats and to enhance the cyber defense capability of national cyber warriors. In cyberspace, there is no clear boundary to determine who will fight together for national security. Recent national cyber exercises currently attempt to invite as many entities as possible to participate regardless of whether they are private companies, public institutions, national critical infrastructure operators, or from the military or academia. To handle a national cyber crisis effectively, it is critical to prepare all potential players within the country so that they can become involved and effectively perform their expected roles whenever necessary. International cooperation with allied countries or international organizations also becomes more important. The capacity of national cyber security involves readiness for well-ordered cooperation or coordination between all possible cyber stakeholders. This is one of the reasons why increasing numbers of large-scale cyber exercises to cover national and international cooperation have tended to be introduced recently.

Recent cyber exercises have also attempted to integrate their technical hands-on exercises with high-level operational or strategic table-top exercises. The omnipotence of the advanced ICT technologies also defines the unlimited power of malicious cyber attacks. However, to ground the exercise scenario in reality and to keep the exercise participants immersed without questioning the authenticity of the scenario, scenario injects[1] for operational or strategic table-top exercises should be connected to technical scenarios. This also provides an opportunity to test one of the most important cyber crisis management capabilities: cyber crisis communication to support rapid and accurate decision-making. To assess the current situation of cyber security exercises accurately, reporting to high-level decision-makers with the correct, often non-technical, terms and with succinct but sufficient information is crucial. Therefore, providing interesting and realistic scenarios to trigger the need for technical players to report to high-level table-top players is constantly being emphasized during efforts to prepare national and international cybersecurity exercises.

Not all cyber attacks should be reported to the high-level officials' table, requesting their timely decisions. What determines the need to report is the damage that the cyber attacks cause or will soon cause to organizations, a nation or to the international community. Therefore, another trend in current cyber security exercises is that they expend much more effort on exercising scenarios in which critical infrastructure must be protected. Damage to critical infrastructure through cyber-based attacks can have a significant impact on national security and on the economy and citizens' livelihoods and safety [2]–[4]. It is, therefore, important to develop a comprehensive national strategy to deal with cyber security issues. This effort should be followed by constantly testing and improving the strategy in national exercises on CPBs simulated around national critical infrastructure installations.

When developing national cyber crisis exercise scenarios, many different factors are considered, such as the objectives, participants, and target capabilities of the exercise, among others. Moreover, one of the most interesting questions when preparing national exercises at present centers on what national critical infrastructure sectors should be chosen to be simulated as a CPB for the exercise. One determining factor is how significant the physical harm to individuals or properties may be if and when the sector is compromised by cyber attacks. Efforts to answer this question can create a sense of alertness within the national cyber community and an incentive to develop true national response capabilities against future cyber threats. A system of cooperation will be established.

Therefore, constructing a realistic CPB for large-scale national or international exercises has become a critical goal. This provides a magnifying glass for exercise participants to focus on certain sectors of national critical infrastructure and to assess our preparedness, as a nation or along with our international allies. The exercise should be able to visualize the most devastating effect of cyber threats on our critical infrastructure based on realistic, but somewhat worst-case, scenarios. It can provide an opportunity to examine how well we are prepared to battle the future threats of

[1] *Injects* are defined as events, typically planned through entries on the Master Scenario Events List, that controllers must simulate, including directives, instructions, and decisions [1]. Exercise controllers provide injects to exercise players to drive exercise play towards the achievement of objectives. Injects can be written, oral, televised, and/or transmitted via any means (e.g., fax, phone, email, voice, radio, or sign).

cyber attacks on our critical infrastructure. We claim that these tools are fundamental to prepare for such battles in cyberspace.

2. TOWARDS A UNIVERSAL EXERCISE PLATFORM FOR CONSTRUCTING A CPB

Good exercise scenarios should provide decision challenges based on a wide spectrum of scope, duration, and the intensity of the cyber operation consequences. Therefore, exercise preparation groups can leverage a widely known tool developed to make use-of-force assessments [5]–[9]. Known as a Schmitt analysis, it introduces different factors that can be used in the assessment of whether cyber operations violate the prohibition of the use of force; such as severity, immediacy, directness, invasiveness, the measurability of the effects and military characteristics, among others.[2] The most important scenario to cover is when national critical infrastructure is targeted by cyber operations in a manner that may have a severe impact on a State's security, economy, public health, or environment [5].

To experiment with the various criteria of a Schmitt analysis, a versatile CPB exercise environment should be developed. For example, it should be able to visualize the *severity* of physical consequences that can cause great harm to the nation and society. Different types of consequences should be representable. Regarding *immediacy*, given that the timeline of the exercise scenarios may not precisely match the actual exercise time, the time for which to visualize consequences should be controllable based on the exercise progress or the exercise scenario. If technical exercises are integrated with operational or strategic exercises, the process can be *directed* to visualize the consequences and to issue high-level table-top scenarios only when a red team (RT) successfully compromises the blue team (BT)'s network. By designing cyber systems as isolated and highly secured, or military-related, we would also like to consider the *invasiveness* or *military character* factors.

Hence, the technical means of constructing a CPB should be established. Doing so is challenging, because many requirements to support the developed technical and table-top exercise scenarios must be met. One way to tackle this problem is to run the design from scratch. This is the usual means of developing ICS/SCADA testbeds for academic research, for security validation, or for training and exercises [10]–[19]. One main problem with this approach is reusability. For every new critical infrastructure sector introduced, the entire cycle of the CPB development should be iterated with

[2] It is desirable to develop exercise scenarios in which each criterion of the Schmitt analysis can be configured as an adjustable parameter and its variation can be maximized, considering the unsettled nature of the "use-of-force" or "armed attack" threshold. In an ideal situation, this tool can work as a framework in determining the next critical infrastructure target to build as a CPB. In many cases in reality, however, after a CPB is constructed based on its technical or practical availability, exercise scenarios will be developed accordingly.

nearly the same amount of resources used in the previous cycle. Another problem is that conventional testbeds are mostly developed for academic research or for the training of field experts. In many cases, they are not suitable for general cybersecurity exercises and can be leveraged only for very limited purposes, due to their lack of scalability or flexibility.

There are increasing cases of critical infrastructure simulations specifically designed for hands-on or live-fire exercises [20]–[28]. They have many different features, but they also seem to share the common philosophy of realistically emulating the actual field environment and emphasizing the visualization of the physical world as controlled by digital systems in cyberspace and the damaging effects of cyber attacks.

However, it appears that their focus has been on constructing offline cyber ranges. Even when they were intended to provide online exercises, their exercise environments were developed while assuming that the participants would connect to the cyber range network remotely, usually through a virtual private network (VPN) [21], [25]. In such a case, the scalability issue of providing the same environment to each participant is resolved by time-dividing the online access to the system and sharing the environment within the same participant group. A miniaturized diorama city composed of different cyber-physical elements, such as power stations, traffic lights, a water treatment system, military sites, and other elements is developed. The city is controlled and supervised by a realistic ICS/SCADA system and the developers incorporate interesting ideas and technologies to visualize CPBs more realistically.

These cyber ranges, however, are not designed with a view to the reproduction of the same environment to provide a separate environment simultaneously to different BT participants. The scalability issue remains in this sense when targeting large-scale exercises to provide each BT with a separate defense mission on their cyber-physical battleground.

Moreover, cyber ranges are not mobile. When cyber ranges may not be able to accommodate all exercise participants, they can only be experienced through remote video cameras.

Another important issue is extensibility. Exercise coverage is becoming more widespread, and the affiliations to which the training participants belong are becoming more diverse. There is also a growing demand for an exercise environment to cover various areas. Custom designs have limitations. It is necessary to develop a general cyber exercise platform that fosters continuous innovation with integrated knowledge and with accumulated CPB design management experience.

The exercise platform used in this paper refers to a set of technical means for establishing an exercise environment for technical hands-on and table-top exercises to simulate various types of critical infrastructure. The platform should be developed to visualize provocations and responses on the CPBs. It is designed to utilize various technical elements to express the physical properties of cyber-physical systems and the damage that may be caused by a cyber attack on them.

The platform should be capable of extensibility to represent different elements of critical infrastructure on the same platform and to enable the inter-domain integration of different infrastructure sectors seamlessly. In other words, we aim to establish a virtuous process cycle to perform system development on new areas and integrate them while reusing or improving existing systems. Thus, we sought to develop a 'platform' that could gradually encompass all areas of the infrastructure that should be considered to assess and strengthen national cybersecurity capabilities on an ongoing basis.

3. PLATFORM DESIGN CONSIDERATIONS

In this chapter, we describe the design considerations when developing a CPB for the large-scale national or international cybersecurity exercises.

A. Target Exercises

We had two main target exercises, the Cyber Conflict Exercise (CCE)[29] and Locked Shields (LS)[30], [31]. The main development phase lasted approximately one year, from March of 2017 (CCE 2017 planning phase) to March of 2018 (before the LS18 test-run).

1) CCE 2017

Since November of 2017, CCE has been held as an annual national live-fire attack-defense exercise, organized by the National Security Research Institute (NSR) of South Korea. CCE is a real-time battlefield drill between 10 RTs who try to penetrate a multi-level organization network on a virtualized platform and 16 BTs who try to defend the network. The maximum number of people per team is limited to five, and all participants gather at an offline venue for this event. CCE can attract the interest of young national cyber security talents or experienced pen-testers to join the RT and to practice their knowledge and skills. BT participants have included many cyber security specialists working in different public sector areas, including those in the military, government, or who work with critical infrastructure, as well as those from the major private industries, including ISPs, banks, major game companies, and other sectors. Online preliminary competition rounds for RTs and BTs are held one month before the final exercise execution to select finalists from all the applicants.

Each BT is presented with a realistic virtualized network composed of four different zones, in this case a DMZ, an internet-connectable work zone, an intranet zone, and an ICS/SCADA zone. As usual, vulnerabilities and misconfigurations have been pre-built into this game network. Each RT should engage in step-by-step intrusion activities to access this hierarchically constructed network, pivoting through compromised machines. The ICS/SCADA Zone has served as the core element of the exercise network. It is the main cyber-physical battlefield and the final destination of the RTs. There will be significant damage if RTs succeed in penetrating this layer and committing a successful cyber attack.

One of the main exercise objectives was to provide interesting challenges which demonstrate realistic cyber incident challenges in the realistically complex full-network environment for each BT. Therefore, the highest priority is to build a realistic ICS/SCADA zone with a realistic implementation of all of the core elements included.

We also wanted to provide the participants with a dramatic visualization of their defense target, our CPB or our society, and the consequences of failures to defend these targets. Though CCE is still a highly technical live-fire attack-defense exercise, some ‘soft’ skills are also tested through some injects to request accurate, succinct and prompt situational reports to be sent to decision-makers and to provide sensitive and time-critical media interview questions. Here, visualization will serve a critical function by providing situational awareness on the progress of the exercise overall. This will also help high-level decision-makers who are observing the exercise to raise their cyber security awareness.

2) LS18

Locked Shields is the world’s largest and most advanced international technical live-fire cyber defense exercise, as described by the NATO-affiliated Cooperative Cyber Defence Centre of Excellence (CCDCOE), which has run it since 2010 in Tallinn, Estonia. During the design of this platform, cooperation between NSR and CCDCOE for LS18 was underway.

Because the goal of LS is to offer a full-stack exercise that integrates LS technical hands-on exercises with operational or strategic/policy/media table-top exercises, there has been a long-standing desire to experiment with various scenarios covering more critical infrastructure sectors. However, another important principle of the LS team is that the technical game and the table-top exercise must constantly be integrated. Therefore, exercise scenarios could be introduced only when the technical implementation of a new CPB is possible and the scalability issues are resolved. Therefore, our platform should have very flexible options to adapt based on changing LS demands, and it must have distinct features for specializing in large-scale cyber exercises.

B. Scalability

Originally, CCE 2017 targeted 20 BTs. For LS, the number of participating BTs has been growing rapidly, such that LS18 was expected to host more than 20 BTs. Our objective was to provide an identical and complete ICS/SCADA zone for each participating team. This means that we need to develop up to 30 sets, considering the backups and demos for the observers.

However, the costs of the specialized hardware elements, such as PLC (programmable logic controllers), actuators, and other electronic and physical devices, as well as software elements, such as an HMI (human-machine interface), historian DB, PMS (patch management system), are very high and open-source alternatives may not be available. Building scalable systems for large-scale national or international exercises was the most important goal of the project, and this goal needed to be considered in all of the design considerations listed below.

C. Mobility and Ease of Deployment

In most cases, mobility and ease of deployment are essential when considering a situation in which work cannot always be done because a venue is rented and a remote exercise site must be constructed within a limited time immediately before the event. The goals are to design and construct an environment that minimizes unnecessary annoyances which arise when moving the platform, to ensure ease of moving the platform, to establish a remote exercise site, and to establish a connection with the main server hosting the virtualized exercise network.

D. Reality or Similarity to the Field Environment

It is fairly odd to emphasize this because it is the most important consideration when building a critical infrastructure simulation system and must always be considered. Ironically, in reality, most of the ICS/SCADA simulation systems tend to be criticized for not being realistic, for many different reasons. This may be unavoidable unless the original systems and network are identically copied. For security reasons, it is often not possible or even desirable to have a complete copy of an actual operating network. Performing cyber attack-defense exercises on actual networks has many risk factors.

The basic principles for developing this platform are as follows. First, we conduct on-site visits to understand the actual network, security threats, and actual working environment of each field and design the exercise environment after consistent and in-depth discussions with operational experts and cyber security experts in each field. Second, the key to reality concerning the exercise goals is whether the cyber crisis scenarios offered during the exercise are based on real-life cases or highly probable future threats. To maximize the exercise effect given to the BT and to ensure the immersive participation of all on this team, we strive not to compromise practicality, completeness or complexity with the technical implementation of the essential elements of the scenario.

E. Extensibility, Flexibility, and Reusability

When selecting the target critical infrastructure sector to represent the damage situation of major national infrastructure elements caused by a cyber attack, it is necessary to consider the following factors comprehensively: the exercise objective; the exercise participants; the accessibility of the technical information of the sector; the extent of the effect of damage; recent actual cyber accident cases; the cost of system and software development; LS strategy game scenario concerns; interdependency between critical infrastructure sectors; and other related factors. The coverage of the target sectors should be gradually expandable based on these criteria.

One of the most important effects of the platform is the accumulation of knowledge. Providing a shared framework of thinking that facilitates continuous innovation and improvement should be a key function of the platform. When we develop one critical infrastructure simulation system from scratch, the result will be very different from another, depending on the design choices, i.e. the system size, the implementation scope or level, the visualization concept of the simulated physical world, among other considerations. This heterogeneous collection of knowledge cannot be combined naturally. It is not cumulative. Therefore, it is necessary to develop a universal cyber exercise platform that will foster continuous innovation by providing a well-established framework when developing exercise scenarios, creating technical measures when developing new CPB designs that will enable a rich set of challenging and interesting exercise scenarios and integrating them with the existing exercise environment seamlessly.

F. Domain-Independency and Vendor-Independency

There are various types of ICS communication protocols [32], [33]. Depending on the practices or main suppliers in each sector, organization or site, the operating communication protocols differ. The characteristics of the communication subjects, organizations, sites, and the construction completion year can all make a difference as well. It is not uncommon for decades-old legacy systems to continue to operate with multiple security vulnerabilities and without major software or security updates. Depending on the vendor or contractor, the system architecture can also differ greatly.

Many major vendors often use their proprietary communication protocols instead of standard open-source protocols.

The point is that supporting all possible implementation scenarios is not possible. The platform was designed to support as many protocols as possible under the given setup. PLC models were chosen considering the ease of recreating various cybersecurity threat scenarios. For example, to reproduce many types of cyber attack, PLC models that support multiple protocols which are compatible with Internet protocols are considered, such as Modbus TCP, CIP Ethernet/IP, Profinet, OPC, or ICCP. The platform is designed to support two or more PLC models so as not to be dependent on specific vendors. If a new protocol requirement arises, certain elements such as PLCs, SWs or APIs should be replaceable with existing ones to support them. The platform should enable a modular design in this sense.

G. Visualization

The goal was to develop a platform with a visualization layer that represents physical facilities and the damage caused by cyber attacks. As noted above, this is one of the main differences between the exercise platform and typical ICS/SCADA testbeds. Considering scalability, extensibility, and reality, it was determined early on that 3D-printing technology would be used to design and produce the diorama city in a more cost-effective and modular approach. This platform can best utilize the advantages of small-volume production of various designs of 3D printing.

The established design principles are as follows. In the center of the visualization layer, symbolic structures that represent each critical infrastructure sector are located. The surrounding area, which includes the residential, commercial and/or industrial districts, represents the physical world we live in and will show the spreading damage when needed. City districts should be designed to connect and expand with adjacent districts.

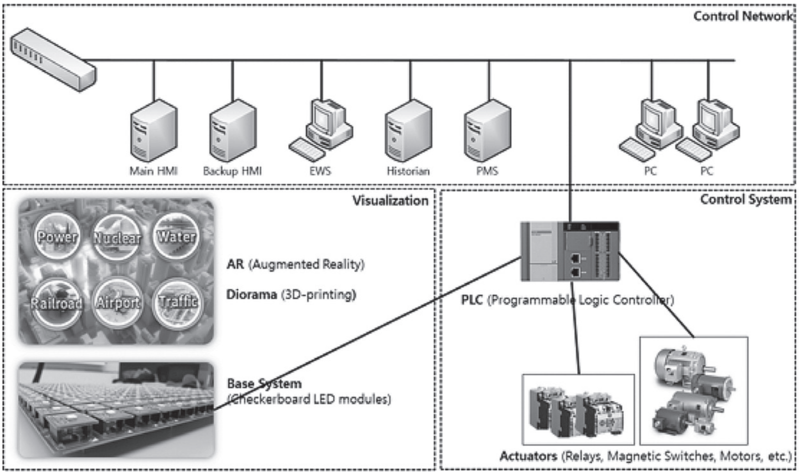
There should also be a way to provide situational awareness on top of the created cyber-physical world. At the very least, there should be a technical means of representing the normal state and the level of the damage caused by a massive cyber attack. Though there is a vast range of options from which to choose, scalability and extensibility are the top priorities. In relation to this, a basic system that uses different colors of RGB (tri-color) LED (light-emitting diode) lights is introduced first, while more dramatic and physical representation techniques could be used. It is simple but effective, with little risk of physical failure. We also devised a method to utilize AR (augmented reality) visualization technology in the 3D-printed diorama city to maximize this effect.

4. ARCHITECTURE OF THE EXERCISE PLATFORM

Based on the design considerations discussed in the previous chapter, we created the basic architecture of the exercise platform, developed a prototype, validated it, produced a modified version by fixing its faults, and used it for two major target exercises, CCE 2017 and LS18. It satisfied most of the considerations in the original design phase and contributed greatly to the success of the exercises. The platform is a system with scalability and extensibility, which were most important, and has thus far shown remarkably different concepts and possibilities compared to those of existing systems. Its visualization showed great potential and it received numerous favorable reviews, along with some criticism, as might be expected.

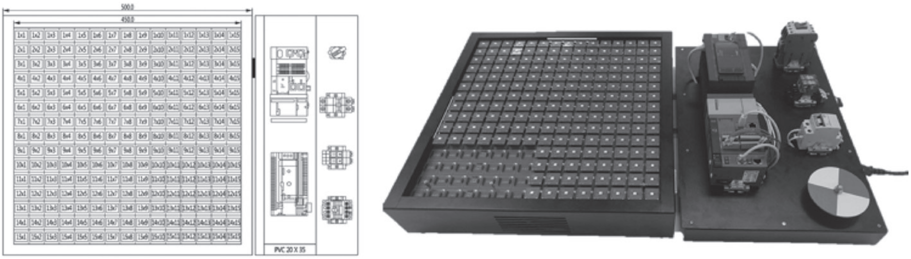
The platform consists of three main components: a visualization layer, a control system layer, and a control network layer, as shown in Figure 1. The visualization layer allows the LED modules to be placed by default on the base system in a 15x15 checkerboard pattern. A four-layer PCB (printed circuit board) was designed to control a total of 255 (LED or other digital) modules. On top of this, the 3D-printed diorama is positioned, and LEDs are used to represent a normal state and an abnormal state in different color schemes. As an option, AR technology was used to express this effect more vividly.

FIGURE 1. THREE MAIN COMPONENTS OF THE EXERCISE PLATFORM



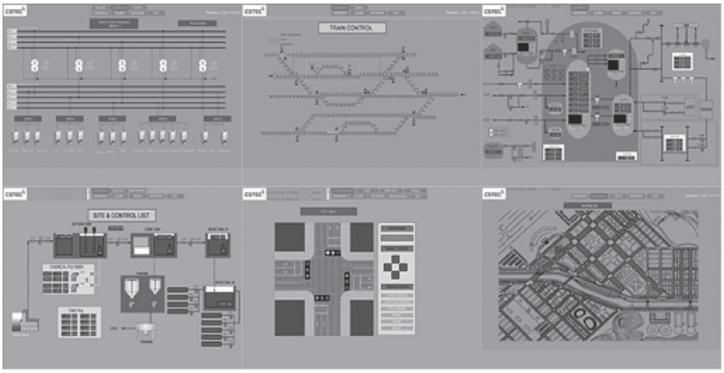
To design the control system layer, six critical infrastructure sectors (a power grid, a nuclear power plant, a water purification plant, railroad control, airport control, and traffic light control) were selected and implemented among the major national critical information infrastructure sectors designated by the Korean government. After an in-depth analysis of the control networks of each field, we derived common elements to be the focus of the development. Two PLC models from two different PLC vendors (one from a local Korean vendor and the other from a European global vendor, considering the geographic locations at which the target exercises take place) were chosen to meet several requirements, such as supported network protocols, power supply voltage, device size, usability in the actual field, and budget limitations, among others. In order to add reality by performing the actual physical operations, some typical actuators, such as a mechanical relay, a magnetic switch, and a motor with a turning plate, were connected to and controlled by the PLCs, making physical sound or moving effects. There is one master switch with which to select the operating PLC. The visualization layer unit and the control system layer unit are designed so that they can be connected and separated easily and stably through the D-sub connector for power supply and communication (see Figure 2).

FIGURE 2. DIAGRAM AND IMPLEMENTATION OF THE VISUALIZATION LAYER AND CONTROL SYSTEM LAYER



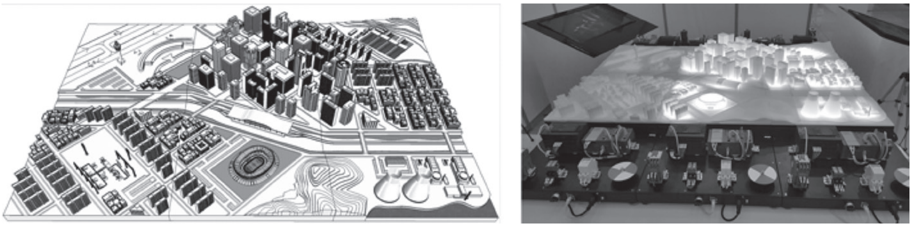
The control network layer is not a visible part of the platform, given its implementation in the virtualized game network hosted by remote cloud exercise servers. Connecting the platform to the game server was designed simply and easily as the plug-and-play level with one Ethernet interface. The control network is configured to provide a virtual environment that includes common control system components such as an HMI, an engineering workstation, a historian DB, a patch management system (PMS), and office computers. After conducting multiple on-site visits and an in-depth analysis, and consulting with field experts, we developed a highly advanced exercise environment and realistic cybersecurity incident scenarios so that the exercise participants can experience situations very similar to those in the real world. We made every effort to achieve high-quality results in all six selected fields. Common software or functionalities are shared and reusable code is recycled as much as possible. However, the PLC logic and HMI design that characterize each field are implemented independently to ensure a high degree of similarity to actual systems in the field (see Figure 3).

FIGURE 3. THE HMIS OF SIX DIFFERENT CRITICAL INFRASTRUCTURE SECTORS



A 3D-printed diorama is designed and produced for a private residential area and a commercial area in which people live, centering on a base site symbolizing each field. Completing these six sectors and integrating them into one large city naturally alludes to the extensibility of the system across critical infrastructure sectors (see Figure 4).

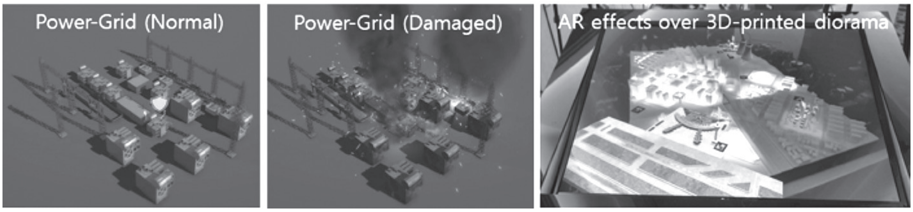
FIGURE 4. THE PROTOTYPE DESIGN AND THE FINAL 3D-PRINTED RESULT OF A SMART CITY DIORAMA COMPOSED OF SIX SECTORS



To make the visualization more dramatic, an additional feature is developed to automatically recognize the six critical infrastructure sectors and launch real-time live graphics using AR technology on the diorama. As shown in Figure 5, we designed the AR visual effects to show a normal state of each sector, its damaged state, and the state transition between them (due to RT's successful attack or BT's successful restoration of the damaged system) for each sector.[3]

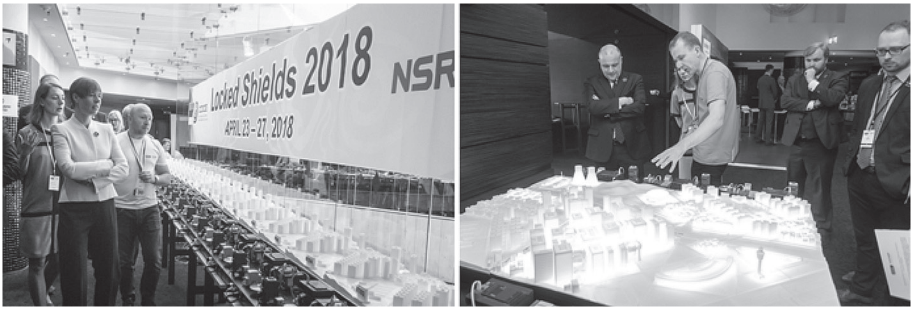
[3] Initially, showing the exercise progress using the AR was considered. However, there was also a concern that more than necessary information for RTs or BTs can be provided for them to experience a realistic cyber conflict during the exercise. Therefore, AR was designed to provide only the amount of information that can be experienced and obtained in reality. A situational awareness tool was developed independently for exercise operators or observers.

FIGURE 5. AR VISUAL EFFECT DESIGN ON THE 3D-PRINTED DIORAMA OF THE PLATFORM



Based on this platform design, CCE 2017 deployed three smart cities (for a total of 18 simulation systems in six areas) to serve as the core network which must be defended by the BT against the RT's campaigns. In LS18, considering fairness across the BTs, 24 complete sets of water treatment plant systems were developed and given to each BT. In addition, one smart city, composed of six different areas, is constructed to provide a demo for the observers (see Figure 6).[4]

FIGURE 6. LS18 SET-UP OF THE WATER TREATMENT PLANTS FOR 22 BTS AND THE LS18 SMART CITY DEMO



During both events, the exercise platform attracted attention as the highlight and it was evaluated to have contributed greatly to the success of both events. Most importantly, we provided each BT with a separate, advanced and realistic ICS/SCADA network environment in which technical hands-on exercises could be conducted. It also enabled the running of a new strategic game scenario of drinking water pollution during a cyber warfare situation. We demonstrated the platform's easy deployment and good mobility during all the processes of preparing and conducting the three exercises of CCE 2017, the LS18 test-run, and the LS18 main execution. Before and after the LS18 events, all systems required long-haul shipping between Estonia and South Korea, but no durability issues arose.

[4] One of the practical but important goals of hosting a large-scale cyber exercise is to raise the cyber security awareness of high-level decision-makers and to increase their interest and investment in cyber security. The enhanced visualization feature of the platform is effective in providing such an impact to achieve the goal.

5. DISCUSSION OF POSSIBILITIES, LIMITATIONS, AND FUTURE IMPROVEMENTS

The platform can revolutionize the national-level cyber exercise process. It is difficult to provide exercise scenarios and environments that are tailored to the needs and tastes of everyone because there are many organizations participating in large-scale national-level exercises and their situations are all different. Generalization is widely used to resolve this issue. Hence, there may be criticism that the scenarios are not specific and do not reflect reality. Customized exercises are great, but can be very costly and may not be suitable for large-scale exercises.

A recent report [34] regarding the Grid Security Exercise (GridEx) IV in the United States highlights an attempt to develop a new exercise process. Six months before the main execution, basic scenarios were given to participating national institutions. Each institution developed its own exercise scenario following the needs of the field and carried out a local exercise in synch with the overall exercise plan. This represents a highly desirable approach, and the question arises whether the proposed platform could be introduced to a similar process. It may be possible for customized exercise environments based on the direct needs and reality from the field to be designed and developed in a distributed manner.

As discussed earlier, one of the most important characteristics of the platform is the accumulation of knowledge. Due to the existence of the exercise platform, knowledge can accumulate around the common elements of the cyber-threat environment of each institution. Through the platform, a portfolio of various national cyber-physical battlefields can be built.

There were some critical reviews of the platform by those who felt that it might oversimplify reality. Reality is a highly relative concept. The concepts of ‘verisimilitude’ or ‘suspension of disbelief’ must be considered.[5] When planning and preparing the exercise, it is necessary to provide trainees with elements that make the exercise situation appear real; if this is done, trainees will be willing to suspend their disbelief within the framework of the narrative provided and accept an impossible mission to protect society. Therefore, having the actual systems used in the field environment, apart from its possibility, does not guarantee a realistic exercise experience. The trainees can feel a greater sense of reality in a simple world that is seamlessly connected. Though there will always be aspects to be improved, we feel that the proposed platform was sufficiently detailed and complete, while implementing the critical elements of a CPB to provide practical real-life experience to the trainees.

[5] *Verisimilitude* has its roots in both the Platonic and Aristotelian dramatic theory of *mimesis*, the imitation or representation of nature [35]. This leads to the idea of ‘*(willing) suspension of disbelief*’, a term coined by Samuel Taylor Coleridge [36]. Although these concepts are originally developed for literary work, they are widely used in any kind of storytelling, including (serious) game design [37], [38].

The future plan is to build a specific and interesting portfolio that will demonstrate the potential of the developed platform. Without becoming mired in unproductive discussions focusing on technical implementation issues, we will select any areas that meet the exercise goals and create the best cyber-threat scenarios in the future. We will secure a variety of interesting CPB deployments.

One possibility is the logical implementation of cross-sector dependencies between multiple critical infrastructure sectors [33], [39]–[41]. Another possibility is to include electronic warfare with cyber exercises [42]–[44]. This will be more appropriate for high-level wargame-like table-top exercises and the use case of the platform may be limited to visualization effects of electronic warfare impacts. Whether it is possible or desirable to integrate cyber warfare and electronic warfare scenarios with very different attributes into one exercise depends on the choices made by exercise planners. Nonetheless, it is clear is that there is a demand for this type of exercise and that this platform has the potential to be used even in these extreme cases. Another possibility is to use sensor modules to construct an IoT-enabled cyber-physical system, such as an IoT-enabled smart grid [45]–[48] or an industrial IoT system [49]–[51]. The possibilities are endless. This platform will provide a basis for accumulated knowledge and technologies as long as we continuously innovate.

6. CONCLUSION

In this study, we proposed a means by which to construct a cyber-physical battlefield platform for large-scale cyber exercises. The main goal is to develop a platform that maximizes the coverage to encompass various design considerations, such as the target exercises, scalability, mobility, reality, extensibility, domain or vendor independency, and visualization technologies of physical facilities and their damage as caused by cyber attacks. The three main components of the platform are the control system layer, the virtual control network layer, and the visualization layer. The HW-based control system layer and the virtualized control network layer are used to simulate the control system operating in the actual field realistically, based on an in-depth analysis of the field. A checkerboard-shaped visualization layer created for a modular design is one of the most noticeable differences of this ICS/SCADA platform.

This platform played a significant role in enhancing the effectiveness of the exercises at the two events of CCE 2017 and LS18. In particular, it was demonstrated that the platform has scalability and extensibility in that a complete CPB was provided to each participating BT and six different critical infrastructure sectors were simulated based on the same platform. These were a power grid system, a nuclear power plant, a water treatment plant, a railroad control system, a traffic light control system, and an airport control system. The goals of developing a practically complex ICS/SCADA security exercise environment that can integrate technical hands-on missions successfully with high-level table-top exercise scenarios and challenge each trainee with a real-life cyber crisis experience that will check their readiness and strengthen their capability were all achieved. We claim that this platform can be a fundamental tool that can foster continuous innovation and the accumulation of knowledge pertaining to national cybersecurity readiness assessment and capability-building activities.

Acknowledgments

The authors would like to thank Dr. Rain Ottis and Dr. Adrian Venables from TalTech and Raimo Peterson from NATO CCD COE for their valuable comments and interesting discussions on topics covered in this paper.

REFERENCES

[1] FEMA, "Program Manual: Radiological Emergency Preparedness (REP)," The Federal Emergency Management Agency (FEMA), USA, FEMA P-1028, 2016.

[2] R. M. Clark and S. Hakim, *Cyber-Physical Security: Protecting Critical Infrastructure at the State and Local Level*, vol. 3. Springer, 2016.

[3] Kate O'Flaherty, "Cyber Warfare: The Threat From Nation States," 03-May-2018. .

[4] Tim Johnson, "The Battlefields of Cyberwarfare Include Infrastructure and Industry, and May Become Deadly," 02-Jul-2018. .

[5] M. N. Schmitt, *Tallinn Manual 2.0 on the international law applicable to cyber operations*. Cambridge University Press, 2017.

[6] M. N. Schmitt, "Computer network attack and the use of force in international law: thoughts on a normative framework," *Colum J Transnatl L*, vol. 37, p. 885, 1998.

[7] M. N. Schmitt, "Cyber Operations and the Jus in Bello: Key Issues," *Intl Stud Ser US Nav. War Col*, vol. 87, p. 89, 2011.

[8] E. T. Jensen, "Computer Attacks on Critical National Infrastructure: A Use of Force Invoking the Right of Self-Defense," *Stan J Intl L*, vol. 38, p. 207, 2002.

[9] A. C. Foltz, "Stuxnet, Schmitt Analysis, and the Cyber Use of Force Debate," Air War College Maxwell Air Force Base United States, 2012.

[10] H. Holm, M. Karresand, A. Vidström, and E. Westring, "A survey of industrial control system testbeds," in *Secure IT Systems*, Springer, 2015, pp. 11–26.

[11] B. Green, A. Lee, R. Antrobus, U. Roedig, D. Hutchison, and A. Rashid, "Pains, gains and PLCs: ten lessons from building an industrial control systems testbed for security research," in *10th {USENIX} Workshop on Cyber Security Experimentation and Test ({CSET} 17)*, 2017.

[12] H. Gao, Y. Peng, K. Jia, Z. Dai, and T. Wang, "The design of ics testbed based on emulation, physical, and simulation (eps-ics testbed)," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2013, pp. 420–423.

[13] R. Candell, K. Stouffer, and D. Anand, "A cybersecurity testbed for industrial control systems," in *Proceedings of the 2014 Process Control and Safety Symposium*, 2014.

[14] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," in *Cyber-physical Systems for Smart Water Networks (CySWater), 2016 International Workshop on*, 2016, pp. 31–36.

[15] E. Korkmaz, A. Dolgikh, M. Davis, and V. Skormin, "Industrial control systems security testbed," in *11th Annual Symposium on Information Assurance*, 2016.

[16] I. Ahmed, V. Roussev, and G. Richard III, "SCADA Testbed for Security and Forensics Research," University of New Orleans, New Orleans, United States, 2017.

[17] M. Almgren et al., "RICS-el: Building a National Testbed for Research and Training on SCADA Security (Short Paper)," in *International Conference on Critical Information Infrastructures Security*, 2018, pp. 219–225.

[18] S. Adepu, N. K. Kandasamy, and A. Mathur, "EPIC: An Electric Power Testbed for Research and Training in Cyber Physical Systems Security," in *Computer Security*, Springer, 2018, pp. 37–52.

[19] Q. Qassim *et al*., "A survey of scada testbed implementation approaches," *Indian J. Sci. Technol*., vol. 10, no. 26, 2017.

[20] E. Skoudis, "How to build a completely hackable city in five steps: And why you should build your skills in this arena," *Pen Test Hackfest*, 2013.

[21] E. Skoudis, "NetWars: CyberCity." [Online]. Available: https://www.sans.org/netwars/cybercity. [Accessed: 01-Jan-2019].

[22] CYBERGYM, "Cyber Training and Technologies Arena as a Solution." [Online]. Available: https://www.cybergym.com/arena-as-a-solution/. [Accessed: 01-Jan-2019].

[23] US ICS-CERT and Idaho National Lab., "ICS Cybersecurity (301) Course." [Online]. Available: https://ics-cert.us-cert.gov/Training-Available-Through-ICS-CERT.

[24] J. K. Daoud, "Multi-PLC Exercise Environments for Training ICS First Responders," Air Force Institute of Technology, 2017.

[25] Cyber Security Training and Exercise Center, "Cyber Crisis Defense Training." [Online]. Available: http://www.cstec.kr/cstec/eng/. [Accessed: 01-Jan-2019].

[26] J. Davis and S. Magrath, "A survey of cyber ranges and testbeds," Defence Science and Technology Organisation Edinburgh (Australia) Cyber and Electronic Warfare Div, 2013.

[27] B. Hallaq, A. Nicholson, R. Smith, L. Maglaras, H. Janicke, and K. Jones, "CYRAN: a hybrid cyber range for testing security on ICS/SCADA systems," in *Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications*, IGI Global, 2018, pp. 622–637.

[28] Department of Homeland Security, "Cyber Storm: Securing Cyber Space." [Online]. Available: https://www.dhs.gov/cyber-storm. [Accessed: 01-Jan-2019].

[29] National Security Research Institute, "Cyber Conflict Exercise 2018." [Online]. Available: http://www.cstec.kr/cce2018/eng.html. [Accessed: 01-Jan-2018].

[30] NATO CCDCOE, "NATO Won Cyber Defence Exercise Locked Shields 2018," 27-Apr-2018. [Online]. Available: https://ccdcoe.org/nato-won-cyber-defence-exercise-locked-shields-2018.html. [Accessed: 01-Jan-2019].

[31] NATO CCDCOE, "Cyber Defence Exercise Locked Shields 2013: After Action Report," 2013.

[32] E. D. Knapp and J. T. Langill, Industrial Network Security: Securing critical infrastructure networks for smart grid, SCADA, and other Industrial Control Systems. Syngress, 2014.

[33] ENISA, "Communication network dependencies for ICS/SCADA Systems," TP-06-16-344-EN-N, Dec. 2016.

[34] North American Electric Reliability Corporation, "Grid Security Exercise GridEx IV: Lessons Learned," Mar. 2018.

[35] Wikipedia.org, "Verisimilitude (fiction)." [Online]. Available: https://en.wikipedia.org/wiki/Verisimilitude_(fiction).

[36] Wikipedia.org, "Suspension of disbelief." [Online]. Available: https://en.wikipedia.org/wiki/Suspension_of_disbelief.

[37] J. Thompson, B. Berbank-Green, and N. Cusworth, *Game design: Principles, practice, and techniques-the ultimate guide for the aspiring game designer*. John Wiley & Sons, 2007.

[38] S. De Castell and J. Jenson, "OP-ED serious play," *J Curric. Stud*., vol. 35, no. 6, pp. 649–665, 2003.

[39] P. Pederson, D. Dudenhoeffer, S. Hartley, and M. Permann, "Critical infrastructure interdependency modeling: a survey of US and international research," *Ida. Natl. Lab*., vol. 25, p. 27, 2006.

[40] F. Petit *et al*., "Analysis of critical infrastructure dependencies and interdependencies," Argonne National Lab.(ANL), Argonne, IL (United States), 2015.

[41] EPSA Analysis: J. Phillips, M. Finster, J. Pillon, F. Petit, and J. Trail, "State Energy Resilience Framework (Argonne, IL: Argonne National Laboratory, December 2016)," ANL/GSS-16/4.

[42] D. C. Schleher, *Electronic warfare in the information age*. Artech House, Inc., 1999.

[43] C. Wilson, "Information operations, electronic warfare, and cyberwar: Capabilities and related policy issues," 2007.

[44] M. C. Libicki, "The convergence of information warfare," *Strateg. Stud. Q*., vol. 11, no. 1, pp. 49–66, 2017.

[45] C. Bekara, "Security issues and challenges for the IoT-based smart grid," *Procedia Comput. Sci*., vol. 34, pp. 532–537, 2014.

[46] S. Tonyali, K. Akkaya, N. Saputro, A. S. Uluagac, and M. Nojoumian, “Privacy-preserving protocols for secure and reliable data aggregation in IoT-enabled Smart Metering systems,” *Future Gener. Comput. Syst.*, vol. 78, pp. 547–557, 2018.

[47] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, *Internet of Things security and forensics: Challenges and opportunities*. Elsevier, 2018.

[48] M. A. Ferrag, L. A. Maglaras, H. Janicke, and J. Jiang, “A survey on privacy-preserving schemes for smart grid communications,” *ArXiv Prepr. ArXiv161107722*, 2016.

[49] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, “Security and privacy challenges in industrial internet of things,” in *Design Automation Conference (DAC), 2015 52nd ACM/EDAC/IEEE*, 2015, pp. 1–6.

[50] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, “Security and privacy in cyber-physical systems: A survey of surveys,” *IEEE Des. Test*, vol. 34, no. 4, pp. 7–17, 2017.

[51] C. Alcaraz, R. Roman, P. Najera, and J. Lopez, “Security of industrial sensor network-based remote substations in the context of the internet of things,” *Ad Hoc Netw.*, vol. 11, no. 3, pp. 1091–1104, 2013.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Resilience of Cyber-Physical Systems: an Experimental Appraisal of Quantitative Measures

Giuseppina Murino, Alessandro Armando, Armando Tacchella[1]
Dipartimento di Informatica, Bioingegneria e Ingegneria dei Sistemi (DIBRIS)
Università degli Studi di Genova –
Viale Causa 13, 16145 –
Genova, ITALY
giuseppina.murino@edu.unige.it
alessandro.armando@unige.it
armando.tacchella@unige.it

Abstract: Cyber-Physical Systems (CPSs) interconnect the physical world with digital computers and networks in order to automate production and distribution processes. Nowadays, most CPSs do not work in isolation, but their digital part is connected to the Internet in order to enable remote monitoring, control and configuration. Such a connection may offer entry-points enabling attackers to gain control silently and exploit access to the physical world at the right time to cause service disruption and possibly damage to the surrounding environment. Prevention and monitoring measures can reduce the risk brought by cyber attacks, but the residual risk can still be unacceptably high in critical infrastructures or services. *Resilience* – i.e., the ability of a system to withstand adverse events while maintaining an acceptable functionality – is therefore a key property for such systems. In our research, we seek a *model-free, quantitative*, and *general-purpose* evaluation methodology to extract *resilience indexes* from, e.g., system logs and process data. While a number of resilience metrics have already been put forward, little experimental evidence is available when it comes to the cyber security of CPSs. By using the model of a real wastewater treatment plant, and simulating attacks that tamper with a critical feedback control loop, we

[1] The authors wish to thank Leonardo S.p.A. for its financial support. The research herein presented is partially supported by project NEFERIS awarded by the Italian Ministry of Defense to Leonardo S.p.A. in partnership with the University of Genoa. This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 830892 for project SPARTA.

provide a comparison between four resilience indexes selected through a thorough literature review involving over 40 papers. Our results show that the selected indexes differ in terms of behavior and sensitivity with respect to specific attacks, but they can all summarize and extract meaningful information from bulky system logs. Our evaluation includes an approach for extracting performance indicators from observed variables which does not require knowledge of system dynamics; and a discussion about combining resilience indexes into a single system-wide measure is included.

Keywords: *cyber-physical systems security, critical infrastructure protection, situational awareness and security metrics*

1. INTRODUCTION

A cyber-physical system (CPS) intertwines physical processes, hardware, software, and communication networks [1]. Examples of CPSs include water treatment plants, power plants and distribution networks, industrial plants, transportation vehicles, and smart buildings. The number of security incidents affecting CPSs has been steadily increasing over the past few years – see, e.g., [2]. The bottom line is that CPSs connected to the Internet can be the root cause of disruption in services, damage to equipment or severe impairment of human activities. Malicious acts most often exploit the weakness of the "*red dot*" representing the virtual place of convergence between Information Technology (IT) and Operation Technology (OT): exploitation of the former provides attack vectors, while exploitation of the latter makes kinetic impacts possible. Detecting weaknesses, fixing them and monitoring critical events in CPSs are compelling and heavily investigated matters, but we must also acknowledge that, in spite of all the efforts made to secure CPSs, interconnected systems may never be fully secure.

In this scenario, the concept of *resilience* emerges as an additional target, complementary to prevention and protection from attacks, but no less important. This line of thought is pervasive in the Presidential Policy Directive 21 [3] about the security of critical infrastructure, which defines resilience as "*[…] the ability to […] withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents*". More recently, the term cyber resilience has been coined to identify specifically "*the ability to continuously deliver the intended outcome despite adverse cyber events*" [4], and this is the interpretation to which we adhere in the following. More specifically, we believe that stakeholders like CERTs (Computer

Emergency Response Teams), management authorities, regulators, and local and national government branches could be interested in a resilience evaluation framework possessing the following properties:

- *Model-free*. Accurate mathematical models of real-world scale CPSs are very difficult to obtain and maintain. Therefore, the assessment of resilience should not require a detailed description of the system dynamics, e.g., in the form of system equations or other formal models, but rather it should be possible to rely on monitored process data and events only.
- *Quantitative*. A synthetic measure (or index) must be provided that describes as faithfully as possible the amount of damage that a system can tolerate before becoming unstable or irreversibly damaged, or before exhibiting potentially dangerous behaviors.
- *General-purpose*. The way in which the resilience index is computed, starting from performance indicators, should be applicable, in principle, to as wide a class of systems as possible, in order to achieve economy of scale in the deployment of the framework.

We propose an evaluation methodology that fulfills all the requirements cited above to extract resilience indexes from, e.g., system logs, control process data, and SIEM (Security Information and Event Management) tool logs. While several proposals exist in the literature, many of them do not meet the requirements we seek and, for those that do, little or no experimental evidence about their adequacy to account for resilience against cyber attacks is available. In order to start bridging this gap, out of a literature analysis consisting of 47 research papers and surveys, we selected four indexes that can be applied to quantify resilience independently from system dynamics and structure. Using the model of a real wastewater treatment plant, and simulating attacks that tamper with a critical feedback control loop inside the plant, we compare the indexes considering different attack hypotheses on a daily basis using Monte Carlo simulations. The computation of the indexes is oblivious of specific features of the system, but critically depends on the selection of performance indicators to extract system performances out of the evolution of monitored data. Our results show that the distributions of the selected indexes across the simulation of different attacks differ in terms of behavior and sensitivity, but they all extract meaningful information from bulky system logs.

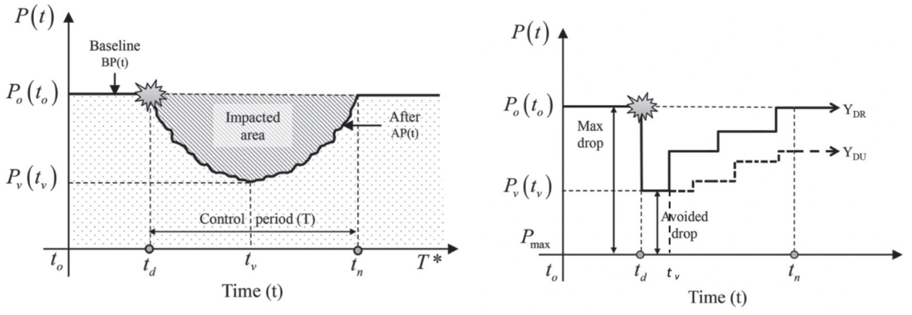
To sum up, the main contributions of the paper are:

- Comparison of four resilience indexes obtained from a thorough literature analysis involving over 40 research papers, in order to ensure model freedom and generality.

- An approach that does not require a mathematical model of system dynamics to extract performance indicators from observed variables.
- A discussion and a proposal about combining resilience indexes obtained from several process variables into a single system-wide measure.

The rest of the paper is structured as follows. In Section 2, we introduce the basic terminology. We succinctly review the related literature and we introduce the indexes we selected for evaluation, including some of the motivation behind their choice. In Section 3, we introduce our wastewater treatment facility case study and we describe the model that we devised in Matlab/Simulink® including its simulation under attack-free conditions. In Section 4, we describe the experimental models, including attack modalities, extraction of performance indicators and a discussion about the combination of resilience indexes. In Section 5, we present some results related to the case study according to the experimental setup described in Section 4. A brief discussion of the results is contained in Section 6, and we conclude the paper in Section 7 with some final remarks.

FIGURE 1: GENERIC RESILIENCE EVALUATION SCENARIO (LEFT) FOCUSING ON THE DIFFERENCE BETWEEN BASELINE PERFORMANCE BP(T) AND AFTER-IMPACT PERFORMANCE AP(T) OVER A CONTROL PERIOD T. GENERIC RESILIENCE EVALUATION SCENARIO (RIGHT) FOCUSING ON THE MAXIMUM AND AVOIDED PERFORMANCE DROPS DURING THE ADVERSE EVENT. NOTATION AND PICTURES FROM [5].



2. BACKGROUND AND RELATED WORK

The definitions and notation that we use are mostly borrowed from [5]. The plot in Figure 1 (left) is presented to describe a generic resilience evaluation scenario. The coordinates are time (x-axis) and performance (y-axis), *BP(t)* is the *Baseline Performance* and represents the performance of the system under normal conditions, whereas *AP(t)* is the *After-impact Performance* and represents the performance of the system after the impact of some disruptive event. Such an event is assumed to happen at time t_d (*disruption time*) and end at time t_n (*return to normality time*), where T= t_n-t_d is defined as the *control period* in [5]. A further point of interest is t_v (*lowest performances time*) where the system reaches the minimum level of performance after disruption. The period *T** is defined as the *observation period* and the condition *T*>T* holds. The plot in Figure 1 (right) introduces the notion of *maximum performance drop* (Max drop) and *avoided performance drop* (Avoided drop) which represent, respectively, how much performance can be lost before the system ceases to be functional and how much performance is left when the system reaches the minimum level of functionality after the attack and before the recovery. With reference to Figure 1 (left), the first resilience index that we consider is introduced by [6] and is defined as

$$\psi_A = \int_{t_d}^{t_n} \frac{AP(t)}{T}\, dt$$

The index ψ_A considers the area of the curve *AP(t)* normalized over the control period *T*, i.e., the residual normalized performance of the system during the disruption. Clearly, the higher the value, the closer to normal operating conditions, and the greater the resilience of the system. The advantage of this index is that it does not require establishing a baseline and it can be readily applied to any performance indicator computed on process data. The main disadvantage is that it assumes knowledge of the control period which, in the majority of cyber attacks, is not known and is difficult to estimate.

An index that overcomes such limitations, but that does require the establishment of a baseline performance, is introduced by [7], [8] and [9]. It is defined as

$$\psi_B = \frac{\int_{t_0}^{T^*} AP(t)dt}{\int_{t_0}^{T^*} BP(t)dt}$$

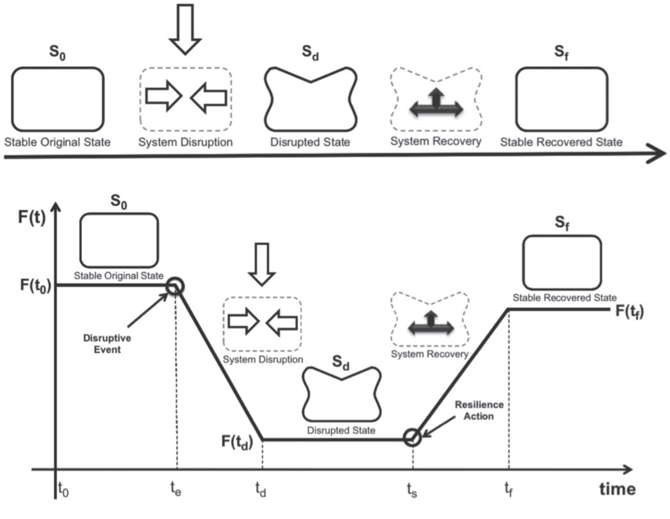
This index is the ratio of the areas enclosed by the curves *AP*(*t*) and *BP*(*t*). It ranges from 0 to 1, where the former is the limit case in which the disruptive event occurs at time t_0 and the system immediately loses its functionality, so that $AP(t)=0\ \forall t\in[t_0;T^*]$. The latter is the limit case in which no functionality is lost, i.e., $AP(t)=BP(t)\ \forall t\in[t_0;T^*]$. Both ψ_A and ψ_B consider the overall evolution of the system during (a subinterval of) the observation period. However, in [10] an index based on the values of max drop and avoided drop is put forward:

$$\psi_C = \frac{Avoided\ drop}{Max\ drop} = \frac{P_v(t_v)\text{-}P_{max}}{P_o(t_o)\text{-}P_{max}}$$

In this case, the evolution of the curves $AP(t)$ and $BP(t)$ are not relevant to establishing the value of the index, since only their extreme values are taken into account. While it is sufficient to consider only specific points in time to compute ψ_C, the evolution of system performances over the control period is completely disregarded.

Besides the above-mentioned contributions, our literature analysis included several other papers that we do not list here owing to a lack of space. References that are worth mentioning are [11], which helped us frame the problem of resilience evaluation, and [12], which provided us with an extensive bibliography to which we refer for further reading about the topic. Since our case study relates to wastewater treatment, we also considered a number of references related to the resilience of water/wastewater treatment plants, including [13], [14] [15] and [16], but we could not find additional candidates for evaluation that met our requirements. In particular, all the indexes proposed in the water/wastewater literature are specific to a given topology and system structure and are difficult to generalize to other plants.

FIGURE 2. PICTORIAL EVOLUTION OF THE STATE OF A SYSTEM UNDER ATTACK (TOP) AND RELATIONSHIP BETWEEN STATE EVOLUTION AND PERFORMANCE OF THE SYSTEM COMPUTED BY A FIGURE OF MERIT (FOM) FUNCTION (BOTTOM). NOTATION AND PICTURES FROM [17].



Considering the fact that the resilience indexes of our choice are based on performance indicators, the question of how to compute such indicators arises. In other words, while it is relatively easy to monitor process variables, the performance of the system cannot always be monitored directly, and should be inferred from collected data. In Figure 2, we present two plots excerpted from [17], wherein a resilience-oriented general-purpose and model-free method to derive performance indicators from state variables is presented. The plot on the top of Figure 2 represents pictorially the evolution of the state of the system during a disruptive event. The deformed box represents the state of the system under duress, and it is meant to show that the impact on state variables can involve several of them at the same time. Nevertheless, as it is shown in the plot at the bottom of Figure 2, we must relate the evolution of state variables to some "bathtub" curve which resembles the curve $AP(t)$ of Figure 1 (left). The proposal of [17] is to introduce a *Figure of Merit* (FOM) function, i.e., a function $F{:}S \rightarrow \mathbb{R}$ which maps any state s ∈S to a corresponding performance indicator. In general, mappings such that the condition $F(s)>F(s')$ holds whenever the performance of the system in state s is better than in state *s'* should work. In [17] no details on how to derive such a function are given, because this is a system-specific process.

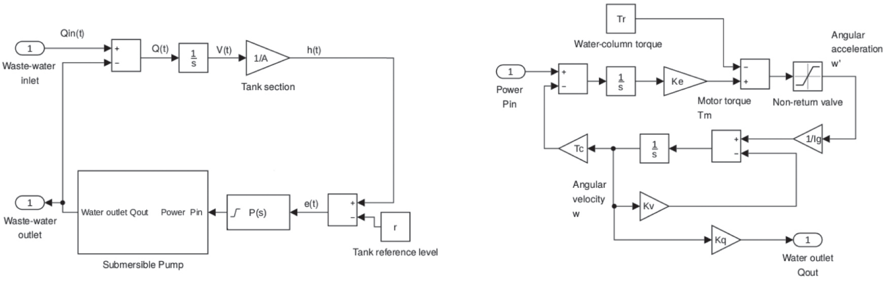
3. CASE STUDY: WASTEWATER TREATMENT FACILITY

A. Brief Description

The facility[2] performs sewage treatment using MemJet™/MemPulse™ MBR (micro-membranes) technology and ensures depollution and dumping at sea of urban wastewater produced by domestic and economic activities in an international tourist area encompassing a marine reserve. The facility handles an estimated maximum of 36,000 people, roughly equivalent to a wastewater supply of 250 liters per person, per day. The maximum output reaches up to 1,200 cubic meters/hour of purified wastewater. The plant is heavily automated: all biological, chemical and mechanical processes are controlled and monitored by a SCADA system connected through the Internet with a remote monitoring center located in the headquarters of the utility company running the plant. The plant consists of a pre-treatment compartment, responsible for filtering large solids – e.g., rags, plastics, nappies, grit and floating materials, oils and fats – before feeding a balancing reservoir. From here, the pre-treated input flow is pumped into the biological compartment where, passing through a denitrification (anoxic) process and a transition into nitrification-oxidation tanks, the oxygenated mixed liquor flows into the MBR reactor for solid-liquid separation and subsequent discharge of the effluent at sea. This is a physical-biological process, which requires precise software-based regulation in order not to wear out micro-membranes and to avoid outputting untreated liquor. The maximum mass flow rate through of MBR tanks – a reference for the whole process − is 900 cubic meters per hour.

[2] Name and location of the facility cannot be disclosed for security reasons.

FIGURE 3. MATLAB/SIMULINK® MODELS OF THE NITRIFICATION-OXIDATION (NO) TANK SUBSYSTEM (LEFT) AND OF THE "SUBMERSIBLE PUMP" COMPONENT (RIGHT). THE ACTUAL UNIT IS DRIVEN BY AN ASYNCRHONOUS MOTOR WITH 15KW OF RATED POWER CONTROLLED THROUGH AN INVERTER. IN THIS SIMPLIFIED MODEL WE ASSUME THAT THE INPUT SIGNAL IS THE POWER DELIVERED TO THE MOTOR AS COMPUTED BY A PROPORTIONAL REGULATOR.



B. Modeling and Simulation

In order to achieve a realistic, yet manageable, case study, we decided to model only the main wastewater cycle. Furthermore, we focus on the nitrification-oxidation process (tank NO) which is upstream from the final purification process (tank MBR) and thus is critical for the performance of the whole cycle. In Figure 3 (left) we show the detailed Matlab/Simulink® model of the tank NO. As we can see from the diagrams, we have assumed a simplified (first order) linear model, whereby the total volume *V(t)* of fluids contained in the tank is obtained by integrating the net inlet mass flow rate *Q(t)* which, in turn, is obtained by subtracting the outlet mass flow rate *Qout(t)* from the tank inlet *Qin(t)*. While the latter is an input to the NO subsystem, the tank outlet is controlled by electrical pumps driven by a proportional regulator tracking a given set point *r* on the height of the tank. The detail of the motor/pump model is given in Figure 3 (right). Also in this case, we assumed a (second order) simplified linear model of an asynchronous drive, whereby the pump rotation generates both viscous friction and counter-motion force, which simulates the asynchronous drive frequency lag.

Two key nonlinearities in the model are *(a)* the saturation of the control signal between 0 and 15KW, which corresponds to the actual range of power within which the pump operates and *(b)* the presence of a non-return valve which does not allow the pump to reverse its operation. The goal of the regulator is to avoid the tank becoming too full, so as to avoid triggering emergency bypasses, or too empty, so as to avoid impairing the chemical process undergone in the NO tank. Both events are undesirable because bypasses dump untreated sewage liquor in the sea, whereas incomplete chemical processing of wastewater may cause failures in subsequent steps. For this reason, we decided to focus our study on this part, on the hypothesis that an attacker may gain virtual access to the facility network and compromise this feedback loop and thus also the inlet flow to the MBR tank. As a yardstick for the calculation of resilience indexes, we simulate the plant without assuming external attack attempts in a Monte Carlo setting. To achieve this, we consider historical data made available from the managing utility to simulate regular sewage inlet. Random variates of the daily inlet profile under conditions of maximum utilization are obtained by adding (band-limited) Gaussian white noise with deviations of 20%. In the following, we call *baseline scenario* the simulation obtained by running the plant without attacks.

4. MODELING: SIMULATING ATTACKS, PERFORMANCE INDICATORS AND SYSTEM-WIDE RESILIENCE

A. Attack Scenarios

To develop attack scenarios, we must consider the effects that an attacker may induce by gaining system access. Conceptually, feedback control loops are at the core of every CPS, and an attacker gaining access to the control system can alter them in three ways: *(a)* by changing the set point, *(b)* by altering the feedback signal, and *(c)* by changing the regulator parameters. To illustrate, consider the control loop that keeps the level of the NO tank close to the desired level shown in Figure 3 (left). Here, attack *(a)* corresponds to changing the desired tank level *r*, attack *(b)* corresponds to altering the actual tank level feedback *h*, and attack *(c)* corresponds to changing the proportional gain of the regulator *P(s)*. In practice, an attacker may decide to perform all such actions and in more than one part of the system, as well as other disruptive actions – blocking the functionality of components or flooding them with requests. Some of these attacks can be prevented or detected by SIEM tools, but attacks on feedback loops can be subtle and destructive. As an example, the pump keeping the NO tank at level can be exercised more than necessary by fooling the controller about the tank level in a small, but persistent way. Such an attack pattern – similar to the one staged by the famous Stuxnet virus [18] – is very difficult to detect, but it reduces the residual life of the pump and thus it is worth evaluating its impact on resilience.

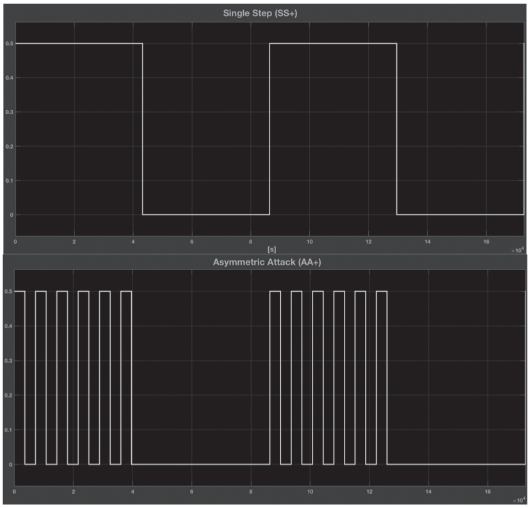
In our simulations we assume that an attacker may alter the set point of the regulator by subtracting a disturbance – attack *(a)*. Under this hypothesis and given the structure of the feedback loop, this attack is equivalent to an alteration of the feedback signal – attack *(b)*. We did not consider attack *(c)* as well as multiple or blocking attacks, but our evaluation framework is able to handle them without modifications. We can obtain several attack scenarios by changing:

- The *duration*, i.e., the control period (in seconds) T=t_n-t_d, as defined in Section 2.

- The a*mplitude* Δ_a, i.e., how much the reference signal is changed.
- The *frequency* f_a; when the disturbance is periodically zeroed every $^1/f_a$ seconds during *T*.

In Figure 4 we show an example assuming T=12 hours and Δ_a=0.5 meters. The plot on top depicts the case in which the duration of the disturbance is held fixed during the attack: we call this *positive single step attack* scenario (SS+), and we foresee also a negative counterpart SS- (*negative single step attack*). The plot on the bottom depicts the case in which the attack signal has a period of two hours (f_a in the order of 10^{-4} Hertz): we call this *positive asymmetric attack scenario* (AA+) and *negative asymmetric attack scenario* (AA-) its counterpart. We also combine the two attacks in a *symmetric attack scenario* (SA), wherein the disturbance ranges from Δ_a to -Δ_a with frequency f_a. In Section 5 we report results obtained by running these scenarios with different values of T, and Δ_a.

FIGURE 4. CHANGES TO THE NO TANK REFERENCEL BROUGHT BY THE HACKER ATTACK. SINGLE STEP POSITIVE (TOP) AND ASYMMETRIC POSITIVE (BOTTOM). THE PLOTS DEPICT TWO ATTACKS LASTING 12 HOURS EACH OVER A TOTAL TIME OF 48 HOURS. THE PERIOD OF THE ASYMMETRIC ATTACK IS 2 HOURS.



B. Building Performance Indicators Through FOM Functions

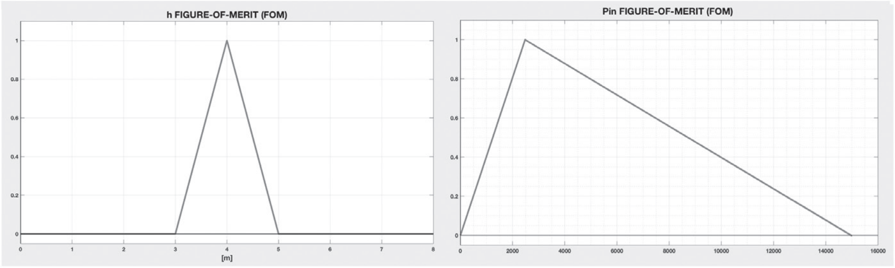
The resilience indexes presented in Section 2 rely on performance indicators, and suitable FOM functions must be provided to map observed variables to the performance space. Considering our case study, the variables that we observe are the following:

- The height of the NO tank *h*; this is the state variable whose reference point is subject to the attack, and it is thus the main focus of our investigation.
- The power delivered to the pump *Pin*; among the effects of a successful and silent cyber-attack, wearing the pump and reducing its residual life is a concrete possibility.
- The outlet mass flow rate *Qout*; the mass flow rate through membranes in the MBR tank, which is downstream from the NO tank, must be regulated precisely, lest the purification process malfunctions or even ceases to work.

As for the definition of FOMs, we can make some observations:

- FOM functions are of the form $F{:}D\rightarrow\mathbb{R}$, where D is the domain of the observed variable, but without loss of generality we can restrict our FOMs in the range [0;1], where 0 and 1 represent minimum and maximum performance, respectively.
- We posit that, when an observed variable x is close to some desirable value(s) x_{good}, then $F(x)\cong 1$, whereas if x is close to undesirable value(s) x_{bad}, then $F(x)\cong 0$.
- $F(x)$ should behave monotonically with respect to the distance from x_{bad} and x_{good}: it must decrease when getting close to x_{bad} and increase when getting close to x_{good} – a concept we borrow from [19].

FIGURE 5. FIGURE-OF-MERIT (FOM) FUNCTIONS FOR TWO OUT OF THREE OBSERVED VARIABLES RELATED TO THE NO TANK: TANK HEIGHT H (TOP) AND POWER SIGNAL TO THE PUMP PIN (BOTTOM). EACH FOM FUNCTION TAKES AS INPUT AN OBSERVED VARIABLE AND RETURNS AN ADIMENSIONAL FIGURE BETWEEN 0 (WORST PERFORMANCE) AND 1 (BEST PERFORMANCE).



We now consider Figure 5, where we represent FOM function for NO tank height (top) and power delivery to the pump (bottom). We do not show the one for outlet mass flow rate, but it similar to the ones shown in Figure 5. The shape of the functions is the simplest satisfying the constraints outlined above, where a linear decay in performance is assumed when variables are getting away from desirable values. More specifically, for each observed variable we identify (un)desirable values as follows:

- The reference value of tank height *h* is 4 meters, therefore we consider h_{good}=4; the tank can tolerate some amount of overshooting of the reference level, but heights of five meters and more may cause spilling; therefore, we set h_{bad}=5 and, symmetrically, h_{bad}=3.
- Under normal conditions, the power delivery to the motor is *Pin*≅3 KW, therefore we set Pin_{good} to the average value under normal operations; the pump operates within 0 to 15KW, which means that delivering power always close to 15KW reduces its residual life, whereas values close to 0 mean that the pump is switched off or works at reduced power; therefore, we set Pin_{bad}=0 and Pin_{bad}=15000.
- Under normal conditions, the outlet mass flow rate is *Qout*≅0.05 m^3/s, therefore we set $Qout_{good}$ to the average value that the variable assumes under normal daily operations. Attempting to deliver more than 0.3 m^3/s mass flow rate to the MBR tank as well as shutting down the flow completely might damage the membranes; therefore, we can set $Qout_{bad}$=0 and $Qout_{bad}$=0.3.

In Figure 5, we show FOM functions assuming linear decay of performances. We remark that this choice is arbitrary and other possibilities exist which are compatible with our assumptions, e.g., quadratic or cubic decay to penalize small changes with respect to x_{good} less than large ones, or RBF (radial basis function) profiles to smooth the decay and avoid discontinuities at the boundaries.

C. A Discussion About System-wide Resilience Indexes

The introduction of FOM functions for each observed variable *h*, *Pin* and *Qout*, enables us to compute resilience indexes related to each variable separately. In our comparison this is fine because we have a relatively limited scope of investigation – the feedback control of the NO tank – and we wish to compare the behavior and the sensitivity of the indexes we consider. However, it can be desirable to build indexes that summarize the performance of the system as a whole, instead of relying on many separate figures. This is especially true when the size of the system grows, and so does the number of observed variables. Keeping in mind that we seek a model-free and general-purpose approach, we can consider three possibilities to extend resilience indexes to a system-wide measure:

- Use a FOM function that maps all the observed variables into a single performance indicator; in our case, this would amount to devising a vector function *F*(*h,Pin,Qout*) to summarize the change of the observed variables into a single performance index.
- Construct a system-wide performance indicator out of scalar FOM functions; in our case, this would amount to combining *F*(*h*), *F*(*Pin*) and *F*(*Qout*) into a single measure, e.g., a linear combination of the three $F(h,Pin,Qout)=\alpha F(h)+$

$\beta F(Pin)+\gamma F(Qout)$, where $\alpha,\beta,\gamma \in [0;1]$ and $\alpha+\beta+\gamma=1$ are *weights* determining the contribution of FOMs.

- Finally, one may either come up with a definition of resilience that accommodates a vector as a performance indicator, or combine resilience indexes computed with scalar performance indicators on single variables; in our case, one may consider, e.g., that a worst-case estimation of the resilience of the whole system can be obtained by considering the smallest index computed according to $F(h)$, $F(Pin)$ and $F(Qout)$.

The first approach is quickly ruled out as the number of observed variables increases. As long as the definition of the FOM function relies on a manual process, defining hyper-surfaces that are meant to respect the given constraints is untenable. One may consider using optimization or machine-learning techniques in order to devise suitable $\mathbb{R}^n \to [0;1]$ mappings (n number of observed variables), but the complexity of the procedure should be factored in. The second approach provides a simplification of the first one, and it remains amenable to manual configuration as long as the number of FOM functions to combine remains small. Scalable linear optimization and relatively simple machine-learning techniques can be used when the number of variables to combine is growing, and hierarchical composition is a possibility. Also, the definition of each single FOM will remain an explainable scalar-to-scalar function. Clearly, the choice of weights to combine the FOM functions is critical for the assessment of resilience, because underestimating or overestimating impacts of a specific FOM may obscure relevant effects in the evaluation of the global resilience index. The third option shares the same issues as the first one when it comes to finding a vector-based index, whereas the combination of different resilience measures is the only approach for which some literature exists. In particular, in [6], the authors propose a method to combine different indexes based on the assumption that they are computed from independent systems. This proposal is not applicable to our case, because the indexes are part of a single feedback control loop. In this case, our proposal is to apply a "weakest link" rule, and estimate the resilience of the overall system considering the resilience index with the smallest median among the ones we compute.

5. EXPERIMENTAL ANALYSIS

We briefly recapitulate the definitions that we have introduced so far to put them in context for our experimental setup. Starting from the resilience indexes that we define in Section 2, let $F:\mathbb{R}\to 0;1$ be one of the FOM functions introduced in Section 4-B, and $x \in \{h,Pin,Qout\}$ be one of the observed variables, where $x(t)$ denotes its value under normal operations and $x_a(t)$ denotes its value under attack scenarios. We consider four resilience indexes defined as follows:

$$\psi_A = \int_{t_d}^{t_n} \frac{F(x_a(t))}{T}\, dt \qquad \psi_B = \frac{\int_0^{T^*} F(x_a(t))dt}{\int_0^{T^*} F(x(t))dt} \qquad \psi_C = \frac{\min_{t\in T^*} F(x_a(t))}{\min_{t\in T^*} F(x(t))} \qquad \psi_D = \int_0^{T^*} \frac{F(x_a(t))}{T^*}\, dt$$

The indexes ψ_A and ψ_B are exactly those defined in Section 2, under the hypothesis that t_0=0. The index ψ_C is computed assuming that the worst-case estimation of the maximum performance drop is the minimum performance of the system under normal operating conditions for a given observation period *T** and that the avoided drop is the minimum performance of the system under attack. Finally, the index ψ_D is obtained from ψ_A by changing the span of the integral from $T=t_n-t_d$ to *T**. The idea behind ψ_D is that, while in our simulations the control period *T* is known, in practice it might be difficult to estimate. On the other hand, the observation period *T** is always chosen by design: in all our experiments, *T**=24 hours.

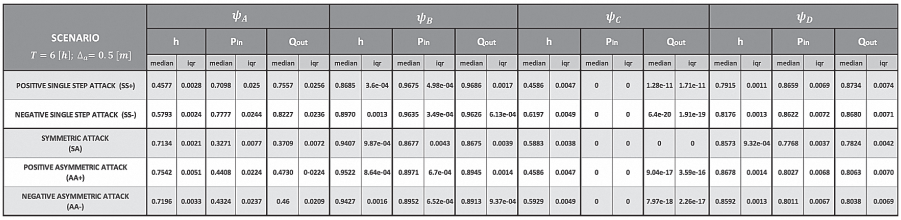
As far as the attack is concerned, we consider all the scenarios defined in Section 4-A, namely single step positive and negative attacks, denoted SS+ and SS-, symmetric attack, denoted SA, and asymmetric positive and negative attacks, denoted AA+ and AA-. For each such attack, we build a factorial experiment with different levels of *T*, Δ_a and f_a. In particular we consider:

- *T*={6,12,18}, i.e., the attack always starts at midnight and lasts 6 to 18 hours.
- Δ_a={0.25,0.5,075}, i.e., the attacker can change the tank reference level from 25 to 75 centimeters.
- f_a={1/3600,1/7200,1/10800}, i.e., the attack period can be one, two, or three hours.

The main reason behind the choice of these values is to increase the probability that the attack on the system remains silent. Indeed, decreasing the period of the attack ($^1/_{f_a}$) below two hours can trigger fast oscillatory system dynamics (e.g., in the pumps) that are unusual in the normal operation of the facility and thus can be identified as anomalous. Also, attempting to change the tank reference level beyond one meter can cause overflow alarms to be triggered. Finally, the attack period is kept at a fraction of the observation period, knowing that longer attack periods imply higher chances of being uncovered. As mentioned in Section 3-B, all the scenarios are simulated on a daily basis, obtaining a different value of the performance indexes that we average over the number of days – one hundred in all of our experiments – for which the simulation runs.[3]

[3] All our experiments run on a PC equipped with an Intel 2.6Ghz Dual Core i7 CPU, 32GB of RAM and running Matlab/Simulink® ver. 2018a on Mac Os Sierra.

TABLE 1. RESILIENCE INDEXES COMPUTED FOR ALL OBSERVED VARIABLES CONSIDERING FIVE DIFFERENT ATTACK SCENARIOS.



SCENARIO $T = 6\ [h];\ \Delta_a = 0.5\ [m]$	ψ_A						ψ_B						ψ_C						ψ_D					
	h		P_{in}		Q_{out}		h		P_{in}		Q_{out}		h		P_{in}		Q_{out}		h		P_{in}		Q_{out}	
	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr	median	iqr
POSITIVE SINGLE STEP ATTACK (SS+)	0.4577	0.0028	0.7098	0.025	0.7557	0.0256	0.8685	3.6e-04	0.9675	4.98e-04	0.9686	0.0017	0.4586	0.0047	0	0	1.28e-11	1.71e-11	0.7915	0.0011	0.8659	0.0069	0.8734	0.0074
NEGATIVE SINGLE STEP ATTACK (SS-)	0.5793	0.0024	0.7777	0.0244	0.8227	0.0236	0.8970	0.0013	0.9635	3.49e-04	0.9626	6.13e-04	0.6197	0.0049	0	0	6.4e-20	1.91e-19	0.8176	0.0013	0.8622	0.0072	0.8680	0.0071
SYMMETRIC ATTACK (SA)	0.7134	0.0021	0.3271	0.0077	0.3709	0.0072	0.9407	9.87e-04	0.8677	0.0043	0.8675	0.0039	0.5883	0.0038	0	0	0	0	0.8573	9.32e-04	0.7768	0.0037	0.7824	0.0042
POSITIVE ASYMMETRIC ATTACK (AA+)	0.7542	0.0051	0.4408	0.0224	0.4730	0-0224	0.9522	8.64e-04	0.8971	6.7e-04	0.8945	0.0014	0.4586	0.0047	0	0	9.04e-17	3.59e-16	0.8678	0.0014	0.8027	0.0068	0.8063	0.0070
NEGATIVE ASYMMETRIC ATTACK (AA-)	0.7196	0.0033	0.4324	0.0237	0.46	0.0209	0.9427	0.0016	0.8952	6.52e-04	0.8913	9.37e-04	0.5929	0.0049	0	0	7.97e-18	2.26e-17	0.8592	0.0013	0.8011	0.0067	0.8038	0.0069

In Table 1 we show the results for T=6 and Δ_a=0.5. Each row of the table is related to an attack scenario. Columns are divided into four groups, corresponding to the resilience indexes, and each group reports the median and interquartile range (iqr) of the resilience index computed using a specific variable and related FOM function. The choice of median and iqr as measure of center and spread, respectively, are motivated by the fact that they are more robust to outliers and the presence of skewed distributions. A glance at the table reveals the following facts:

- The iqr is always at least one order of magnitude smaller than the median except when the median is 0 as in ψ_C; this indicates that indexes are not very sensitive to the random variation of the input flow.
- The index ψ_D is more conservative than ψ_A; this is to be expected, because the former averages the effects of the attack over the whole observation period.
- Considering observed variable *h*, the lowest resilience values are obtained for the SS+ attack; this is because the attack signal is *subtracted* from the reference level, and thus throughout the duration of the attack the tank is seen by the controller to be emptier than in reality; in AA+ and SA attacks this is not true, because the attack signal oscillates and, on average, the controller keeps the tank level closer to normal.
- Considering observed variables *Pin* and *Qout*, the worst figures are obtained for the SA attack because the performance of the pump and the mass flow rate output are far from 1 only during transient regimes induced by the "steps" in the attack signals; therefore, in SS+ and SS- attacks, the height of the tank remains "off balance" while the pump and the mass flow rate output stabilize to levels corresponding to normal operation.

We analyzed the data shown in Table 1 considering 36 distributions obtained with SA, AA+ and AA- attacks. We preliminary tested each distribution for normality with the Shapiro-Wilk test, and groups of distributions across attack modality for homoskedasticity (equal variance) with the Levene test (non-parametric version). The results can be summarized as follows:

- the null hypothesis of the Shapiro-Wilk test (values being normally distributed) cannot be rejected at the 5% confidence value for all but a few distributions, e.g., ψ_A computed on state variable *h* for SA, and the distributions of ψ_C for state variables *Pin* and *Qout*.
- considering the distributions of single resilience indexes computed for specific state variables, and comparing them across different attacks, the null hypothesis of the Levene test (variances being equal) can be rejected at the 5% confidence value for all the groups we consider with the single exception of ψ_C for state variables *h* and *Pin*.

Given the above results, we compare the distributions across attack modalities with a multiple pairwise Mann-Whitney U-test (non-parametric alternative to t-test) using Bonferroni's correction for p-values. Overall, the results of this test confirm that the qualitative observations we made above hold true.

For example, in the case of ψ_A considering variable *h*, and attacks SA, AA+ and AA-, the null hypothesis that two samples obtained from different attacks are coming from the same distributions can be rejected at the 5% confidence level in all cases. For lack of space, data obtained with T=12,18 hours, Δ_a=0.25,0.75 meters and f_a={1/3600,1/10800} are not reported, but similar considerations apply also to these cases.

Using the rule proposed in Section 4-C, the overall resilience of the system under the various attacks can be estimated considering the minimum value for each index in a given row. For instance, if we consider ψ_A with T=6 and Δ_a=0.5, we would get a global index $\Psi_A = 0.4577$ (the value for *h*) in the SS+ attack, and $\Psi_A = 0.3271$ (the value for *Pin*) for the SA attack.

6. DISCUSSION

While our current results are not a ready-made tool for detecting or preventing cyber attacks, in principle some of the resilience indexes we propose could be deployed to support an intrusion detection tool, e.g., by letting the tool "learn" the baseline distribution of some resilience index during secure operation, and then relying on the tool to detect significant deviations from the baseline during normal operation. Our methodology consists of three steps:

- Identify the relevant state variables considering those available from process control logs.
- Build FOM functions considering (un)desirable values and making assumptions about the effects of variables change on system performance.
- Compute resilience indexes based on FOM functions.

We stress that any system is amenable to this analysis, therefore our methodology is general-purpose. It is also model-free, because identifying state variables does not require knowing system dynamics in detail; also, identifying (un)desirable values requires behavioral knowledge of the process being carried out by the system but does not require the mathematical model of the system. The advantage of relying on our methodology, with respect to standard intrusion detection applied to single process variables, is that our resilience indexes are built and tested to provide statistically significant deviations when anomalies affect the system, and can also be used to summarize the combined effect of several process variables at once. More generally, if a simulator of the CPS under scrutiny is available, one can test and tune resilience indexes to achieve desired properties by means of controlled experiments, and the indexes engineered through simulations will be deployable on the implemented system without further adaptations. For systems in which simulation is not an option, computing indexes is still possible by relying on process data and system logs, while testing and tuning could be performed by replaying historical data.

One key issue arising in practice is the ability of the selected indexes to tell naturally occurring faults from cyber attacks. Given our current approach, a statistically significant deviation in resilience indexes for the wastewater facility can be produced, e.g., by a faulty pump or a stuck-at-level tank sensor. However, naturally occurring faults exhibit predictable patterns, whereas cyber attacks, in general, do not. Therefore, hints about the cause of an anomaly could come from comparison between several indexes, including those obtained simulating possible faults. While we have not yet developed a procedural way to diagnose symptoms of decreasing resilience indexes, we can observe that the behavior of the system in case of SA, AA+/- attacks can hardly be traced back to a physical anomaly: a change in resilience indexes, that are known to be sensitive to those attacks, will indicate that the system is being compromised with high probability.

7. CONCLUSIONS AND FUTURE WORK

We have improved on the current state of the art in resilience evaluation by providing experimental data showing that it is possible to summarize the resilience of a system through numerical indexes that ensure model freedom and generality. Our approach, based on FOM functions computed from observed variables, does not require a mathematical model of system dynamics, but only knowledge of (un)desired values for process variables. We have provided a discussion and preliminary experimental evidence about combining resilience indexes obtained from several process variables. Future work will include furthering our investigation into the combination of several FOM functions or resilience indexes in systems with several observed variables and more complex hierarchical structures. We plan to analyze data from logs of real systems and validate the results obtained with simulation to provide tools for security monitoring for critical infrastructure.

REFERENCES

[1] E. A. Lee, "Cyber Physical Systems: Design Challenges" in *11th {IEEE} International Symposium on Object-Oriented Real-Time Distributed Computing {(ISORC} 2008), 5-7 May 2008, Orlando, Florida, {USA}*, 2008.

[2] G. Loukas, *Cyber-physical attacks: A growing invisible threat*, Butterworth-Heinemann, 2015.

[3] B. Obama, "Presidential Policy Directive 21 (PPD21): Critical infrastructure security and resilience" *Washington, DC*, 2013.

[4] F. Björck, M. Henkel, J. Stirna and J. Zdravkovic, "Cyber Resilience-Fundamentals for a Definition." in *WorldCIST (1)*, 2015.

[5] N. Yodo and P. Wang, "Engineering resilience quantification and system design implications: a literature survey" *Journal of Mechanical Design*, vol. 138, n. 11, p. 111408, 2016.

[6] C. S. Renschler, A. E. Frazier, L. A. Arendt, G. P. Cimellaro, A. M. Reinhorn and M. Bruneau, A framework for defining and measuring resilience at the community scale: The PEOPLES resilience framework, MCEER Buffalo, 2010.

[7] D. G. Dessavre, J. E. Ramirez-Marquez and K. Barker, "Multidimensional approach to complex system resilience analysis" *Reliability Engineering \& System Safety*, vol. 149, pp. 34-43, 2016.

[8] M. Ouyang, L. Dueñas-Osorio e X. Min, "A three-stage resilience analysis framework for urban infrastructure systems" *Structural Safety*, vol. 36, pp. 23-31, 2012.

[9] A. Shafieezadeh and L. I. Burden, "Scenario-based resilience assessment framework for critical infrastructure systems: Case study for seismic resilience of seaports" *Reliability Engineering \& System Safety*, vol. 132, pp. 207-219, 2014.

[10] A. Rose, "Economic resilience to natural and man-made disasters: Multidisciplinary origins and contextual dimensions" *Environmental Hazards*, vol. 7, n. 4, pp. 383-398, 2007.

[11] Y. Y. Haimes, "On the definition of resilience in systems" *Risk Analysis: An International Journal*, vol. 29, n. 4, pp. 498-501, 2009.

[12] S. Hosseini, K. Barker and J. E. Ramirez-Marquez, "A review of definitions and measures of system resilience" *Reliability Engineering \& System Safety*, vol. 145, pp. 47-61, 2016.

[13] P. Juan-Garcìa, D. Butler, J. Comas, G. Darch, C. Sweetapple, A. Thornton and L. Corominas, "Resilience theory incorporated into urban wastewater systems management. State of the art" *Water Research*, vol. 115, pp. 149-161, 2017.

[14] S. N. Mugume, D. E. Gomez, G. Fu, R. Farmani and D. Butler, "A global analysis approach for investigating structural resilience in urban drainage systems" *Water Research*, vol. 81, pp. 15-26, 2015.

[15] S. Panguluri, W. Phillips and J. Cusimano, "Protecting water and wastewater infrastructure from cyber attacks" *Frontiers of Earth Science*, vol. 5, n. 4, pp. 406-413, 2011.

[16] M. Schoen, T. Hawkins, X. Xue, C. Ma, J. Garland and N. J. Ashbolt, "Technologic resilience assessment of coastal community water and wastewater service options" *Sustainability of Water Quality and Ecology*, vol. 6, pp. 75-87, 2015.

[17] D. Henry and J. E. Ramirez-Marquez, "Generic metrics and quantitative approaches for system resilience as a function of time" *Reliability Engineering \& System Safety*, vol. 99, pp. 114-122, 2012.

[18] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war" *Survival*, vol. 53, n. 1, pp. 23-40, 2011.

[19] A. Armando and A. Coletta, “Security Monitoring for Industrial Control Systems” in *Security of Industrial Control Systems and Cyber Physical Systems - First Workshop, CyberICS 2015 and First Workshop, {WOS-CPS} 2015, Vienna, Austria, September 21-22, 2015, Revised Selected Papers*, 2015.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Detection of Malicious Remote Shell Sessions

Pierre Dumont
Department of Information Technology
and Electrical Engineering
ETH Zürich / Kudelski Security
Zürich / Lausanne, Switzerland
pierre.dumont@kudelskisecurity.com

Roland Meier
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

David Gugelmann
Exeon Analytics
Zürich, Switzerland
david.gugelmann@exeon.ch

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Abstract: Remote shell sessions via protocols such as SSH are essential for managing systems, deploying applications, and running experiments. However, combined with weak passwords or flaws in the authentication process, remote shell access becomes a major security risk, as it allows an attacker to run arbitrary commands in the name of an impersonated user or even a system administrator. For example, remote shells of weakly protected systems are often exploited in order to build large botnets, to send spam emails, or to launch distributed denial of service attacks. Also, malicious insiders in organizations often use shell sessions to access and transfer restricted data.

In this work, we tackle the problem of detecting malicious shell sessions based on session logs, i.e., recorded sequences of commands that were executed over time. Our approach is to classify sessions as benign or malicious by analyzing the sequence of commands that the shell users executed. We model such sequences of commands as n-grams and use them as features to train a supervised machine learning classifier.

Our evaluation, based on freely available data and data from our own honeypot infrastructure, shows that the classifier reaches a true positive rate of 99.4% and a true negative rate of 99.7% after observing only four shell commands.

Keywords: *malware, botnets, machine learning, attribution, digital forensics, digital trust, authentication*

1. INTRODUCTION

The rise of cloud and Internet of Things (IoT) infrastructure makes it crucial to access computing services and devices remotely for configuration, maintenance or deployment purposes. Recent numbers show that the secure shell protocol (SSH), which is the state-of-the-art method for remote shell login, is available on over 21 million publicly accessible devices [1]. This number does not include devices that provide SSH access only from the internal network or via VPN, which is the common practice for enterprise and home networks.

Ensuring the security of remote shell sessions is not a trivial task. While the SSH protocol itself is believed to be secure in terms of implemented cryptographic primitives, malicious actors still manage to gain access to Internet-facing SSH servers.

Outdated software, weak or stolen credentials and lack of multi-factor authentication are frequent ways for malicious actors to gain access to a remote device. Recent examples include the Mirai botnet, where attackers gained access to 600,000 devices [2].

Systems that are only available internally (e.g., to employees of a company) are generally better protected because (i) they cannot be attacked from outside the company network; (ii) the administrators can enforce strong authentication schemes; and (iii) the tasks that each user is allowed to do on these systems are usually well-defined. However, even if access is restricted and strong authentication schemes make it impossible to steal credentials, internal systems are not spared from attacks because studies show that many attacks come from insiders [3].

The operator of an infrastructure therefore needs a way to differentiate between legitimate and malicious actions without trusting the identity or authenticity of the logged-in user, because failing to block malicious actors can have severe consequences including downtime, data loss, and reputation loss.

In this paper, we present a system that analyzes commands within shell sessions and classifies them as benign or malicious. We leverage the fact that remote shell sessions leave traces by logging the executed commands.

Problem statement and research questions. We address the following problem statement:

Solely based on the commands that are executed in a shell session, we aim at building a classifier capable of quickly distinguishing between benign and malicious sessions. More specifically, we answer the following research questions:

1. Do individual commands contain enough information for a classification between malicious and benign purposes?
2. How many commands need to be analyzed to accurately identify malicious remote shell sessions?
3. Can we detect attackers who try to obscure their commands?

Approach. Our approach is to perform the classification using supervised machine learning, trained on logs from real (benign) users and malicious logs from real attackers. As features, we use sequences of commands of variable length (i.e., n-grams built from entries in session logs). This allows us to capture the context in which a command was executed.

Challenges. The main challenges that we face in this research are the following:

- Attackers tend to be unpredictable and commands can be ambivalent in their purpose depending on the context. This requires us to define and extract features that capture the context in which a command was executed.
- Attackers can mix malicious commands with harmless commands in order to obscure their intentions. This requires our features to be meaningful even if the context in which commands are executed is obscured.
- Attackers can cause considerable damage with only a few commands. This requires our classifier to output reliable results after a short time (i.e., after analyzing a few commands).

Contributions. The main contributions of this paper are:

- A fully implemented binary classifier using machine learning algorithms to differentiate between malicious and benign shell commands (Section 4.).
- A thorough evaluation of the results to show the usability of our classifier (Section 5.).
- Case studies to illustrate use-cases for our classifier (Section 6.).

2. RELATED WORK

To the best of our knowledge, this work is the first that addresses the problem of classifying malicious remote shell sessions based on the executed commands. Existing works show how attackers can gain access to SSH-enabled devices and how to identify malicious commands in source code. Below, we reference the most relevant publications in these areas.

In [4], Song et al. study users' typing patterns when logging in on SSH sessions to guess their passwords. Other papers (e.g., [5], [6]) analyze SSH from a network perspective and focused on the threat of SSH brute-forcing. Using honeypots, the authors of [7] and [8] study how attackers operate after they gain access to an SSH-enabled device. In contrast to their work, our approach is agnostic to how attackers manage to gain remote access to a system and thus also works for malicious insiders that have legitimate access to a system.

In [9], Shabtai et al. present a broad survey of machine learning classifiers for detecting malicious code. Many of the techniques employed to classify malicious executable files can be found in other papers (e.g. n-grams [10] [11]). In contrast to identifying malicious instructions in a program, our approach works in real time and does not require the full session (i.e. the entire program). Therefore, our approach is useful for immediately intercepting ongoing remote shell sessions, which is different from analyzing static program code.

3. BACKGROUND AND DATA SOURCES

In this section, we define our attacker model and introduce the data sources that we used to train and evaluate the classifier.

A. Attacker Model

In this paper, we focus on commands executed in the UNIX-like shell of a system that can log executed commands. A server with a shell provides a command-line interface to users either on a physical terminal or on a remote interface. To access the shell remotely, the most commonly used protocol is Secure Shell (SSH). SSH servers run software such as OpenSSH to authenticate users and provide interactive terminals. SSH provides a secure encrypted channel with user authentication over passwords or certificates.

We consider the threat of an attacker who has bypassed the SSH authentication system and is therefore able to login remotely. This could be achieved by exploiting SSH software vulnerabilities, performing a man-in-the-middle attack, or by acquiring the necessary password/certificate by guessing or data theft. Alternatively, the attacker could be a malicious user with legitimate access to the system. The goal of our work is to detect such attackers as quickly as possible by analyzing their behavior.

B. Data Sources

To train and evaluate our classifier, we need data from both benign and malicious users of shell sessions. In this section, we describe how we collected publicly available logs of benign sessions and how we used our own infrastructure to collect logs of malicious sessions.

1) Public Bash History Logs

By default, the Bash shell (the most widely used shell) and most other shells log the commands that a user executes. In the example of Bash, all executed commands are by default written to a file in the user's home directory. To obtain access to a large amount of such history files, we leverage the fact that many developers, intentionally or not, publish their history files in public GitHub repositories [12]. Using GitHub's API [13], we were able to collect 3,146 non-empty Bash history files containing a total of 973,621 commands and 234,063 unique commands. Since these commands originate from real users and it is unlikely that attackers would publish their files, we labeled the commands contained in these files as benign.

2) Honeypot Logs

Honeypots are systems that appear to be vulnerable to some attacks. They are typically deployed by security researchers to trick attackers into revealing their malicious code, allowing the defenders to learn about their latest methods of operation.

For our purposes, we run Cowrie, a medium-interaction SSH/Telnet honeypot [14]. After running our honeypot for 5 months (June – October 2017), we collected data from over 320,000 remote shell sessions and a total of 2.35 million executed commands (5,316 unique commands). Because no benign user would log in to an unknown device, it is safe to assume that all these sessions were established with malicious intent by automated scripts and bots. Thus, we labeled the commands in these sessions as malicious.

4. CLASSIFYING REMOTE SHELL SESSIONS

In this section, we describe how we reached our goal of differentiating between malicious and benign sequences of commands. After explaining how we identified useful features, we describe two variations of our classifier: we start with a one-command classifier that classifies one command at a time, and generalize it to an N-command classifier that classifies N subsequent commands at a time.

A. Feature Selection

Whether a command is malicious or not depends on three factors:

- The program that is executed: e.g. "rm" to delete files or directories

- The parameters: e.g. “-rf ./*” to recursively delete everything within the current directory
- The context: e.g. the current directory or previously executed commands.

For example, executing “rm -rf ./*” in a user’s “downloads” directory is likely not malicious but rather used to clean up old files. On the other hand, executing “rm -rf ./*” in a server’s root directory would make the server unusable and is probably done with malicious intent (or by accident). In the following, we describe how we built features for machine learning based on this information. With these features, we wanted to cover the first two factors (program and parameters). For the third factor (context), we used sequences of commands to feed the classifier (cf. Section 4. B.3).

To compute features that contain programs and parameters, there are two extreme approaches: taking the entire command, program and parameters: e.g. “rm -rf ./*”, as one feature or treating the program and parameters as independent features, e.g. “rm”, “-rf”, “./*”. However, both extremes are unpromising because they either fail to recognize similar commands – “rm -rf ./*” has similar effects as “rm -rf ./” – or fail to capture dependencies between the program and its parameters: “rm -rf ./*” is very different from “ls -rf ./*”. Therefore, we followed a middle course and used so-called n-grams as features. N-grams are sub-sequences of n items from a sequence of items. N-grams are widely used in natural language processing and existing work shows that they are also useful for classifying malicious executables [15].

We generate the n-grams by splitting a command at each whitespace, e.g. “rm -rf ./*” decomposes into the sequence [“rm”, “-rf”, “./*”]. This sequence then has 1-grams [“rm”, “-rf”, “./*”],2-grams [“rm -rf”, “-rf ./*”] and 3-gram “rm -rf ./*”.

In Section 4. B.1) we explain how we optimized the value of n for our classifier.

B. Classifier

In this section, we explain how our classifier works. We start by explaining the motivation behind our choice of k-NN as the classification algorithm to use and continue by describing two versions of our classifier: the first classifies one command at a time while the second works on multiple commands simultaneously.

1) Model Selection

For our classifier, we need an algorithm that fulfills the following requirements:

- It is suitable for clustering high-dimensional data
- It is fast enough to classify commands in real time

- It produces explainable results (e.g. to allow manual inspection by a security analyst)
- It provides high accuracy.

There are a large number of binary classification algorithms (e.g. Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), or random forests) that are used in the security domain [16] [17]. After initial experiments with various algorithms, we decided to use the k-Nearest Neighbor (k-NN) algorithm, which achieves good, explainable, and quick results (cf. Section 5.) with a simple model. One major downside of k-NN is its (comparably) high memory requirement because the trained model must contain a large number of data points.

As explained in Section 4. A, we use n-grams consisting of program names and parameters as features for the classification. More precisely, we define a parameter, M and simultaneously use n-grams up to length M. That is, for M = 3, we use 1-grams, 2-grams, and 3-grams as features for the machine learning algorithm.

For each n-gram length, we only consider a fixed number of the most frequent n-grams. This is done to reduce the complexity of the model and avoid infrequent commands that might cause overfitting.

Each n-gram is represented as one dimension in k-NN classification, thus the higher this number is, the larger the space for classification.

2) 1-Command Classifier

The 1-command classifier works on a single command. This comes with the advantages that it is fast and easy to run but – because whether a command is malicious or benign often depends on the commands that were executed before and after – it achieves limited accuracy.

As stated above, we use the k-Nearest Neighbor (k-NN, with the default parameters specified by sklearn[1]) method to classify a command and thus a command is assigned to a label (malicious or benign) of the k-nearest commands.

3) N-Command Classifier

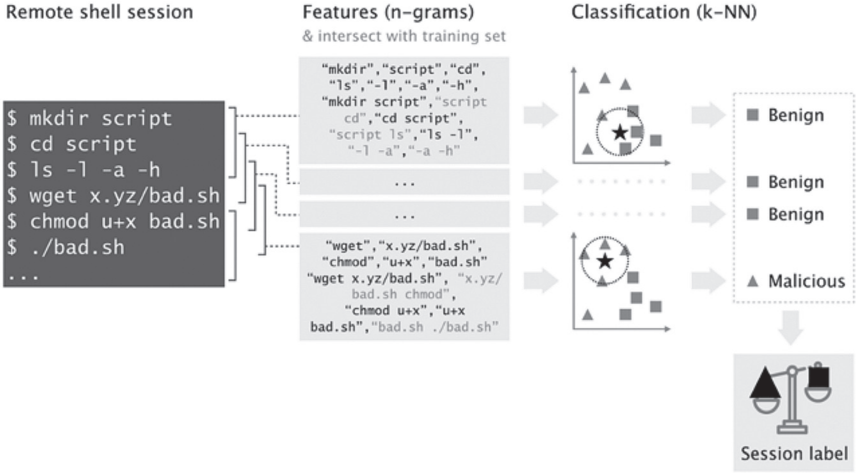
While the 1-command classifier is useful for quickly classifying individual commands, it lacks the ability to capture the context in which a command was executed (i.e. the preceding and succeeding commands); hence, we generalize it to an N-command classifier.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

As the name suggests, the N-command classifier analyzes N subsequent commands together. In contrast to the 1-command classifier (which builds the n-grams for each command individually), it uses a sliding window over the entire shell session and computes the top n-grams for each sequence of N subsequent commands. Similar to the 1-command classifier, it subsequently uses k-NN to predict whether a sequence of commands is malicious or not.

Figure 1. illustrates the N-command classifier with an example.

FIGURE 1. N-COMMAND CLASSIFIER SYSTEM OVERVIEW. FROM SESSION LOGS, WE PARSE N SUBSEQUENT COMMANDS (EXAMPLE: N = 3), COMPUTE THE N-GRAMS FOR N = 1...M (M = 2), REMOVE N-GRAMS THAT DO NOT EXIST IN THE TRAINING DATA, AND APPLY K-NN (K = 4) TO OBTAIN THE PREDICTED LABEL.



```
$ mkdir script
$ cd script
$ ls -l -a -h
$ wget x.yz/bad.sh
$ chmod u+x bad.sh
$ ./bad.sh
...
```

Remote shell session
Features (n-grams)
& intersect with training set
Classification (k-NN)
"mkdir","script","cd", "ls","-l","-a","-h", "mkdir script","script cd","cd script", "script ls","ls -l", "-l -a","-a -h"
...
...
"wget","x.yz/bad.sh", "chmod","u+x","bad.sh" "wget x.yz/bad.sh", "x.yz/bad.sh chmod", "chmod u+x","u+x bad.sh","bad.sh ./bad.sh"
Benign
Benign
Benign
Malicious
Session label

4) Classifying Sessions

To label a session, an operator could specify an arbitrary policy that uses a mix of predictions of individual commands and sequences of variable lengths to determine an action. This would allow an operator to implement custom policies such as to report, terminate or manually inspect sessions depending on the types of commands that they contain.

Since k-NN produces explainable models, the operator is even able to manually inspect the closest commands or sequences that were used to classify ongoing sessions. This would allow for a short feedback loop as the operator is capable of quickly dismissing false positives.

5. EVALUATION

In this section, our classifier is evaluated based on real-world data. After explaining our evaluation methodology, we present the resulting performance of our classifier. We show that the 1-command classifier can quickly identify malicious commands and that the N-command classifier improves the results even for small values of N. Finally, we evaluate the impact of an attacker who tries to obscure their intent.

A. Methodology

In this section, we provide details about the datasets, parameters, and performance metrics used.

1) Datasets

The data used in this paper originates from two data sources: publicly available shell session logs and our honeypot logs, both described in Section 3. B. In Table I, we summarize key information about the datasets used for the following experiments.

TABLE I: INFORMATION ABOUT THE DATASETS USED FOR THE EVALUATION

	Training	**Testing**
Malicious dataset	Honeypot logs (Aug 2017)	Honeypot logs (Sept 2017)
Total commands	837,176	1,408,905
Unique commands	2,774	2,760
Benign dataset	Random sample from public Bash history files	Random sample from public Bash history files
Total commands	29,710	39,749
Unique commands	11,051	11,539

2) Parameters and Metrics

In Table II, we list all the parameters that influence the performance of our algorithm, explain their meaning and specify the evaluated values.

TABLE II: PARAMETERS USED FOR THE EVALUATION

Parameter	**Explanation**	**Evaluated values**
M	Maximum n-gram length (the features are n-grams for n=1, …, M)	1, …, 5
N	Number of subsequent commands analyzed together	1, …, 10
T	Number of n-grams to keep for each n-gram length n	[50, 100, 250, 500, 750]
k	Number of neighbors (for k-NN)	[1, 3, 5, 8]

We interpret commands that are predicted to be malicious as "positive" and commands that are labelled benign as "negative". We then use standard statistical terms and metrics for binary classification:

- True Positive (TP): a malicious command was classified as malicious
- True Negative (TN): a benign command was classified as benign
- False Positive (FP): a benign command was classified as malicious
- False Negative (FN): a malicious command was classified as benign.

In the experiments, we typically report the True Positive Rate (TPR = TP / (TP + FN)), also known as recall, which denotes the proportion of correctly identified malicious commands, and the True Negative Rate (TNR = TN / (TN + FP)), also known as specificity, which denotes the proportion of correctly identified benign commands. The False Positive Rate (FPR = 1-TPR) and the False Negative Rate (FNR = 1-TNR) can be derived from these values.

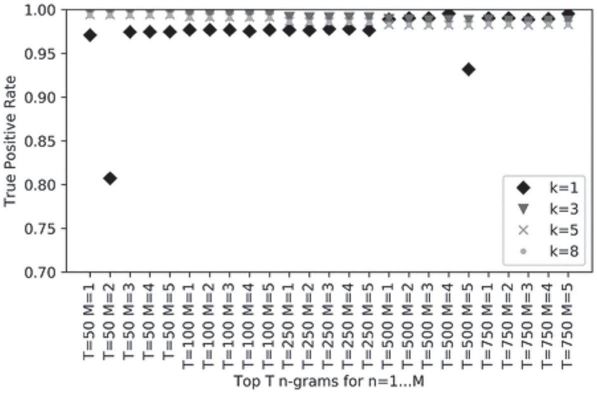
The accuracy of the classifier computes to acc= $\frac{TP+TN}{TP+TN+FP+FN}$.

B. Classification with the 1-Command Classifier

In this experiment, we evaluate the performance of the 1-command classifier, a special case of the N-command classifier with N = 1. This classifier is useful for fast classification, but has limited performance because it does not consider the context in which a command was executed.

We use the datasets described in Table I to train and test the classifier with different values for parameters T, M, and k (cf. Table II). In Figures 2, 3, and 4, we plot the resulting TPR, TNR, and accuracy, respectively.

FIGURE 2. TRUE POSITIVE RATE FOR DIFFERENT VALUES OF T, M, AND K FOR THE 1-COMMAND CLASSIFIER.

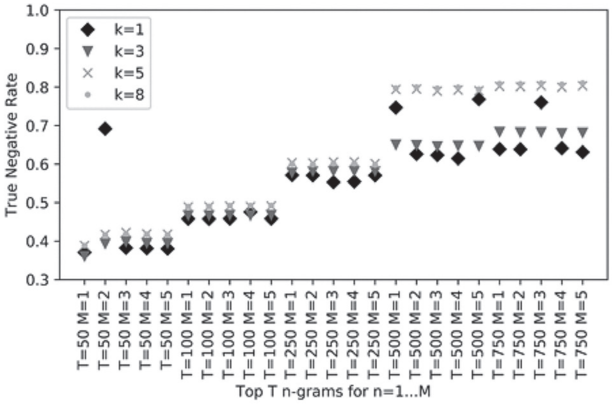


As Figure 2 shows, T and M have little impact on the identification of malicious commands and the 1-command classifier achieves a TPR of more than 0.98 in most configurations. An explanation for the small impact of T and M is the fact that most of the attacks against our honeypot are automated bots using scripts with low variability (i.e. they often run the same commands). Adding too many features by increasing T and M likely overfits the data, which explains the downward trend as both values increase.

Except for k = 1, which is expected to be unstable as it relies only on the closest neighbor, the number of chosen neighbors has little impact on the TPR. The observation that the results are good and stable for small values of k (e.g. k = 3) shows again the low variability; however, as we show below, a higher value for k is required for a good TNR.

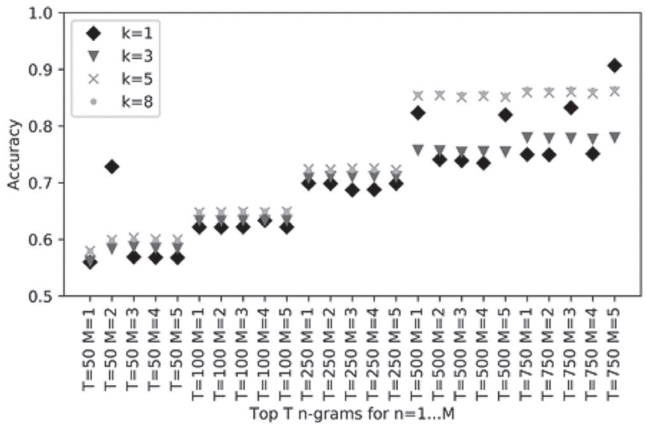
In Figure 3, we analyze the True Negative Rate and show that the 1-command classifier obtains a TNR of up to 0.81 for the parameters used. The results for TNR show the opposite trend compared to the TPR: the higher T and M are, the better the performance of the 1-command classifier. This is due to the high variability of benign commands as they were taken randomly from many different users. For the best TNR, k should be chosen between 5 and 8.

FIGURE 3. TRUE NEGATIVE RATE FOR DIFFERENT VALUES OF T, M, AND K FOR THE 1-COMMAND CLASSIFIER.



Combining the results for TPR and TNR shows that the best results (in terms of accuracy as seen in Figure 4) are achieved for T = 750, M = 5, and k = 8 (TPR = 0.983, TNR = 0.800, accuracy = 0.860). However, because these parameter values result in high resource requirements, we use the slightly sub-optimal parameter values T = 500 and M = 3 for evaluating the N-command classifier in the next section. Note that the 1-command classifier using these parameters still achieves TPR = 0.982, TNR = 0.796, accuracy = 0.855 and is therefore only about 0.5% worse than the optimal configuration (w.r.t. accuracy for k = 8) while requiring approximately 50% less computation time (15 instead of 30 minutes to generate the testing features table), and 50% less storage (750 MB instead of 1.5 GB for the k-NN model for a specific k).

FIGURE 4. ACCURACY FOR DIFFERENT VALUES OF T, M, AND K FOR THE 1-COMMAND CLASSIFIER.



C. Classification with the N-Command Classifier

In this experiment, we evaluated the performance of the N-command classifier for different values of N and k. In contrast to the 1-command classifier evaluated above, the N-command classifier considers the context in which a command was executed by analyzing sequences of N commands (cf. Section 4. B.3).

Figures 5, 6, and 7 show plots of the TPR, TNR, and accuracy, respectively for the N-command classifier depending on N and k. As motivated in the previous section, we fix the values for T and M (T = 500, M = 3). The results show that the TPR as well as the TNR and the accuracy are significantly better for N > 3 compared to the 1-command classifier (note that the results for N = 1 correspond to the 1-command classifier). This confirms our intuition that considering the context in which a command was executed is essential for an accurate classification. We further notice that the value of k has little impact on the TPR and TNR for N > 3.

FIGURE 5. TRUE POSITIVE RATE OF THE N-COMMAND CLASSIFIER AS A FUNCTION OF N (FOR T = 500 AND M = 3). N > 3 SIGNIFICANTLY IMPROVES THE TPR COMPARED TO THE 1-COMMAND CLASSIFIER.

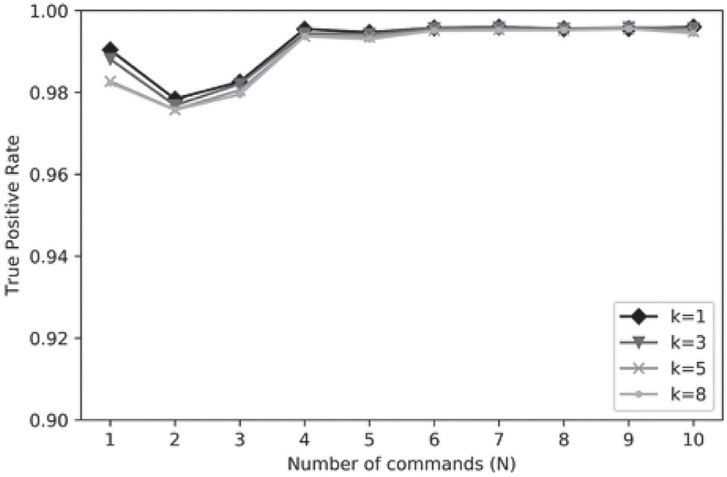


FIGURE 6. TRUE NEGATIVE RATE OF THE N-COMMAND CLASSIFIER AS A FUNCTION OF N (FOR T = 500 AND M = 3). N > 3 SIGNIFICANTLY IMPROVES THE TNR COMPARED TO THE 1-COMMAND CLASSIFIER.

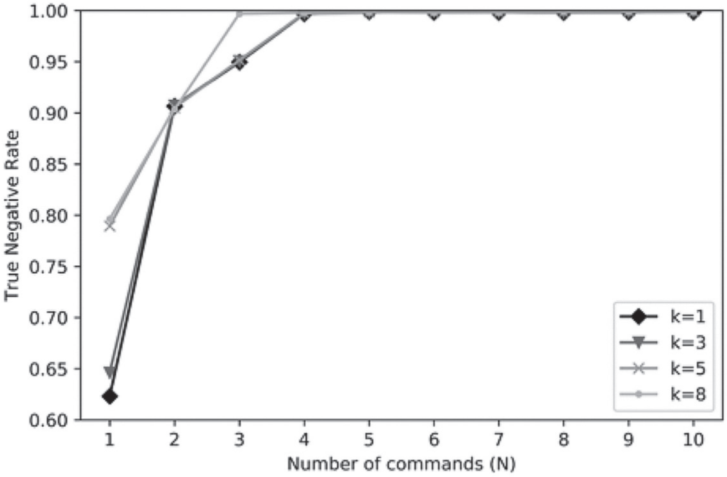
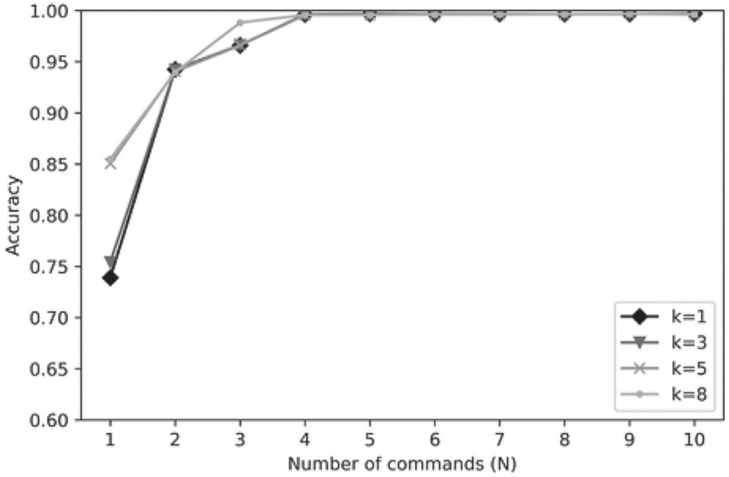


FIGURE 7. ACCURACY OF THE N-COMMAND CLASSIFIER AS A FUNCTION OF N (FOR T = 500 AND M = 3). N > 3 SIGNIFICANTLY IMPROVES THE ACCURACY COMPARED TO THE 1-COMMAND CLASSIFIER.



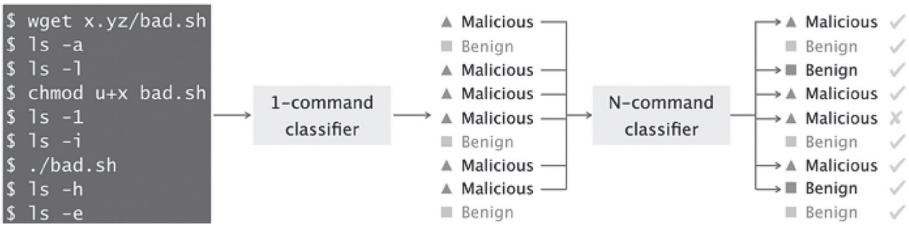
Since the N-command classifier needs to wait until a user has executed at least N commands, it is desirable to have N as small as possible. Therefore, N = 4 is an optimal choice. At this point (and for k = 3), the classifier achieves TPR = 0.994, TNR = 0.997 and accuracy = 0.996. Increasing N only slightly improves the results (by approximately 0.2% until N = 10) and requires waiting for more commands. Note that increasing the number of commands has no noticeable effect on the time it takes to classify those sequences.

D. Robustness Against Obscuring Attempts

While the previously evaluated N-command classifier works well because it considers the context in which a command was executed, it is vulnerable to attackers that try to obscure the context. In particular, an attacker can interleave their malicious commands with benign commands. For example, assuming that the following commands are malicious [“wget x.yz/bad.sh”, “chmod u+x bad.sh”, “./bad.sh”], an attacker could add (presumably) non-malicious commands in between and run [“wget x.yz/bad.sh”, “ls”, “ls -a”, “ls -l”, “ls -al”, “chmod u+x bad.sh”, “ls”, “ls -a”, “ls -l”, “ls -al”, “./bad.sh”] instead. In this example, the benign “ls” commands hide the contextual information from the N-command classifier for N <= 5 because each sequence of 5 commands contains at most one malicious command.

As a promising solution, we propose to use the 1-command classifier in combination with the N-command classifier. That is, (i) we apply the 1-command classifier to identify and remove benign commands inserted by an attacker; and (ii) classify the sequences of the remaining commands using the N-command classifier. We illustrate this approach in Figure 8 and describe it in detail in the following paragraphs.

FIGURE 8. THE 1-COMMAND CLASSIFIER IS USED AS A FILTER AND ONLY COMMANDS CLASSIFIED AS MALICIOUS BY THE 1-COMMAND CLASSIFIER ARE PROVIDED TO THE 4-COMMAND CLASSIFIER. THUS, THE 4-COMMAND CLASSIFIER IS PRESENTED WITH A SEQUENCE OF COMMANDS THAT IS AN APPROXIMATION OF THE ORIGINAL MALICIOUS SEQUENCE.



```
$ wget x.yz/bad.sh
$ ls -a
$ ls -l
$ chmod u+x bad.sh
$ ls -1
$ ls -i
$ ./bad.sh
$ ls -h
$ ls -e
```

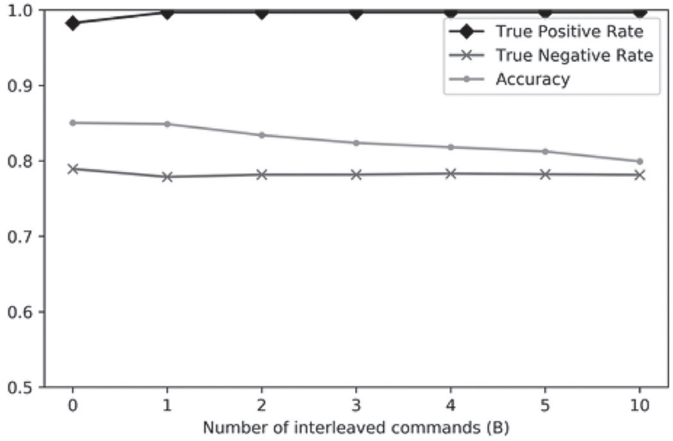
In the first step, we run the 1-command classifier on all commands to identify potentially malicious commands. From previous experiments, we know that the 1-command classifier has a TNR of approximately 80%. Therefore, after removing all commands that the 1-command classifier labeled as benign, we end up with approximately 20% of the benign commands (false positives) and 98% of the malicious commands (true positives).

In the second step, we apply the 4-command classifier on the remaining commands. The intuition behind this is that the 4-command classifier has a small FPR and can therefore compensate for the high FPR of the previously applied 1-command classifier.

Because the 4-command classifier analyzes sequences of 4 commands in a sliding window, each command is contained in up to 4 such sequences and gets a set of up to 4 predictions. To determine the final prediction of an individual command, we simply select the one that appeared most often within this set.

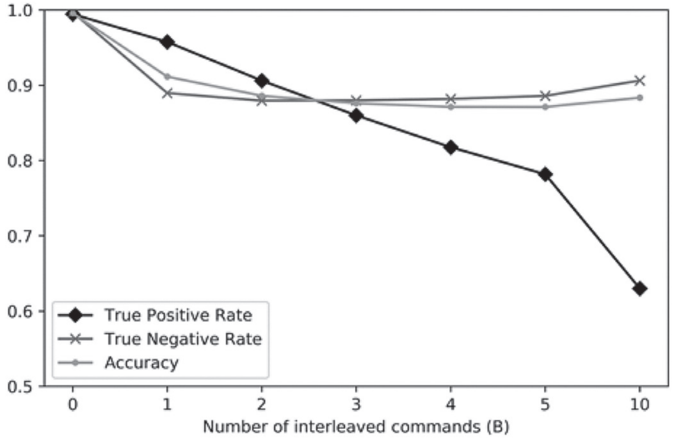
To evaluate this approach, we interleave B benign commands between any two malicious commands from our malicious testing dataset. The performance of the 1-command classifier on the interleaved dataset is equal to the results described in Section 5. B (as expected because the 1-command classifier does not depend on the context; the minor differences can be explained by the larger number of benign commands in the dataset) and are shown in Figure 9.

FIGURE 9. PERFORMANCE OF THE 1-COMMAND CLASSIFIER ON A MALICIOUS DATASET INTERLEAVED WITH B BENIGN COMMANDS (T = 500, M = 3, K = 5). THE RATHER LOW TRUE NEGATIVE RATE, WHICH IMPLIES THAT SOME BENIGN COMMANDS ARE MISTAKENLY CLASSIFIED AS MALICIOUS, IS NOT AN ISSUE SINCE THE 1-COMMAND CLASSIFIER IS ONLY USED FOR PRE-FILTERING BEFORE PROVIDING THE COMMANDS TO THE N-CLASSIFIER FOR FINAL CLASSIFICATION.



In Figure 10, we plot the results for the 1-command classifier combined with the 4-command classifier. As expected, a drop in accuracy occurs as benign commands are interleaved. However, even after interleaving with B = 5 benign commands, the 4-command classifier has an accuracy of almost 90%. This is due to the higher number of benign commands being classified correctly as B increases.

FIGURE 10. PERFORMANCE OF THE 1-COMMAND CLASSIFIER COMBINED WITH THE 4-COMMAND CLASSIFIER ON A MALICIOUS DATASET INTERLEAVED WITH B COMMANDS (T = 500, M = 3, K = 5).



Since the 1-command classifier lets through approximately 20% of benign commands, more benign commands get through as B increases. Therefore, the 4-command classifier has more sequences containing one or more benign commands to classify, reducing its performance.

We conclude from these results that combining the 1-command classifier and the 4-command classifier makes sense to minimize false positives. However, using the 1-command classifier directly minimizes false negatives.

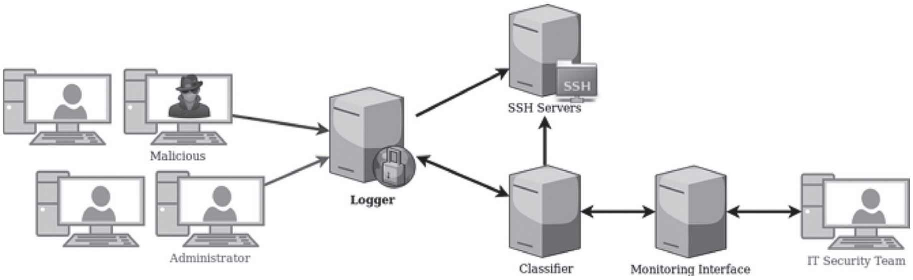
6. USE CASES

In this section, we propose two use cases for which our system is useful: (i) as a passive monitoring tool to protect SSH servers, and (ii) for continuous authentication.

A. Passive Monitoring

The primary use case of our system is for passive monitoring of SSH servers. As illustrated in Figure 11, this tool would be composed of the following components: (i) a *logger* of SSH sessions to record executed commands (e.g. using an SSH transparent proxy, by adapting an SSH server like OpenSSH [18], or by using a tool such as Snoopy Logger [19]); (ii) a *classifier* that receives the stream of commands recorded by the logger and applies our system on them. It should also update the models as they become stale over time; and (iii) a *monitoring interface* for the IT security team to configure the system and observe triggered alerts. However, due to the nature of the malicious data we used (bots targeting a honeypot) automated attacks are more likely to be detected.

FIGURE 11. USING OUR SYSTEM FOR PASSIVE MONITORING OF SSH SERVERS.



B. Continuous Authentication

As another use case, an extended version of our system could be used for continuously authenticating users. Instead of classifying sessions as malicious or not, the system

]

could check whether the commands in the current session are similar to those in past sessions of the same user. This setting is similar to previous works [20] [21], which shows that it is possible to differentiate between users based on keystroke dynamic analysis. Applying this approach to SSH logs would allow the detection of attackers who use stolen credentials of legitimate users. In this case, models should be trained per user.

7. DISCUSSION

In this section, we discuss the answers to our research questions and possible extensions as well as limitations of our system.

A. Research Questions

In the following paragraphs, we answer the three research questions from Section 1.

Do individual commands contain enough information for a classification between malicious and benign purposes?

The results for our 1-command classifier show that it is possible to distinguish between malicious and benign commands from our datasets with a TPR of more than 98% and a TNR of more than 79%. While this is a good starting point, our N-command classifier proves that analyzing multiple commands can lead to significantly better scores.

How many commands need to be analyzed to accurately identify malicious remote shell sessions?

Our evaluation of the N-command classifier shows that 4 commands are enough to accurately distinguish between our two datasets (with TPR 99.4% and TNR 99.7%). Compared to the 1-command classifier, the TNR therefore increases by approximately 20% when analyzing 4 commands (and thus is not significantly increased when analyzing more than 4 commands).

Can we detect attackers who try to obscure their commands?

We observed that when attackers interleave their malicious commands with benign ones, the accuracy falls. However, by combining the 1-command classifier and the 4-command classifier, the overall accuracy is still over 90%. We discuss additional countermeasures as well as other strategies for attackers in the next section.

B. Evading the Classifier

In this paper, we analyzed individual commands and sequences of commands.

However, attackers who know that such a system is being used could employ multiple techniques to evade it.

The first technique that we reproduced here was to interleave benign commands between each malicious command. As expected, the classification performs significantly worse than without interleaved commands. One of the possible solutions against this is to ignore commands that do not change context during the classification (i.e., non-malicious commands such as "ls" could no longer be used to interleave malicious commands). Another approach is to combine our system with keystroke dynamics analysis [20] [21] to detect compromised sessions or accounts.

Another circumvention strategy is to write malicious commands in a script file and execute it or to define aliases for malicious commands. In this case, we would still be able to use the classifier presented in this work, although it would require a tool to analyze the script file or the aliases before executing it. This is feasible for shell commands, but rather difficult for arbitrary programming languages, as one would have to classify commands on the system API level.

C. Classifying Unseen Commands

The classification performance depends on the data quality. We require as many classified commands and scripts as possible to train the classifier. Unknown commands might get decomposed into the n-grams that we previously encountered. However, for truly unseen commands, a default classification to malicious could be used, as most of the benign commands would have been observed after enough time. By design, our system is particularly successful in settings where similar malicious commands are executed frequently, as is the case for automated bots.

8. CONCLUSION AND FUTURE WORK

In this paper, we presented a machine learning-based system to distinguish between malicious and benign shell commands and sessions. Our classifier works solely based on the commands that a remote shell user executes and is able to identify benign and malicious commands with more than 99% true positive and true negative rate after observing 4 commands.

We have shown that our approach works very well in a passive setting where the attacker is not aware of our system's presence. From simulations with a sophisticated attacker who tries to circumvent our system by obfuscating malicious activities, we observed a decrease in the accuracy of our system and discussed possibilities to counteract such attackers.

We see potential future work particularly in two directions: (i) combining our approach with additional measures to identify attackers who try to circumvent the system; and (ii) training our approach on a per user basis to decide whether two remote shell sessions originate from the same user (which would allow us to use our approach for continuous authentication).

REFERENCES

[1] "SSH servers search on shodan.io," [Online]. Available: https://www.shodan.io/search?query=ssh. [Accessed 2 December 2018].

[2] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Level, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas and Y. Zhou, "Understanding the Mirai Botnet," in *26th USENIX Security Symposium*, Vancouver, BC, Canada, 2017.

[3] CA Technologies, "Insider Threat 2018 Report," 2018.

[4] D. X. Song, D. Wagner and X. Tian, "Timing Analysis of Keystrokes and Timing Attacks on SSH," in *10th USENIX Security Symposium*, Washington, D.C., USA, 2001.

[5] L. Hellemons, L. Hendriks, R. Hofstede, A. Sperotto, R. Sadre and A. Pras, "SSHCure: A Flow-Based SSH," in *6th International Conference on Autonomous Infrastructure, Management and Security, AIMS'12*, 2012.

[6] A. Sperotto, R. Sadre, P. d. Boer and A. Pras, "Hidden Markov Model Modeling of SSH Brute-Force Attacks," in *Integrated Management of Systems, Services, Processes and People in IT*, pp. 164-176.

[7] D. Ramsbrock, R. B. and M. Cukier, "Profiling attacker behavior following SSH," in *37th Annual IEEE/IFIP International Conference on Dependable Systems (DSN)*, 2007.

[8] E. Alata, V. Nicomette, M. Kaâniche, M. Dacier and M. Herrb, "Lessons learned from the deployment of a high-interaction honeypot," in *European Dependable Computing Conference (EDCC06)*,, Coimbra, Portugal, 2006.

[9] A. Shabtai, R. Moskovitch, Y. Elovici and C. Glezer, "Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey," in *Inf. Secur. Tech*., 2009.

[10] D. Reddy. and A. Pujari, "N-gram analysis for computer virus detection," *Journal in Computer Virology,* vol. 2, p. 231–239, 2006.

[11] J. Choi, H. Kim, C. Choi and P. Kim, "Efficient malicious code detection using n-gram analysis and SVM," in *14th International Conference on Network-Based Information Systems*, 2011.

[12] "Github search for bash history files, requires to be connected," [Online]. Available: https://github.com/search?l=Shell&q=filename%3Abash_history&type=Code. [Accessed 25 November 2018].

[13] "Github API," [Online]. Available: https://developer.github.com/v3/search/#search-code.

[14] "Cowrie Honeypot Github repository," [Online]. Available: https://github.com/cowrie/cowrie. [Accessed 25 November 2018].

[15] J. Z. Kolter and M. A. Maloof, "Learning to detect and classify malicious executables in the wild," *Journal of Machine Learning Research*, pp. 2721-2744, 2006.

[16] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Commun. Surveys Tutorials*, vol. 10, pp. 56-76, 2008.

[17] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016.

[18] "OpenSSH," [Online]. Available: https://www.openssh.com/. [Accessed 10 December 2018].

[19] "Snoopy Logger Github," [Online]. Available: https://github.com/a2o/snoopy. [Accessed 1 March 2019].

[20] F. Monrose and A. Rubin, "Authentication via keystroke dynamics," in *4th ACM Conf. Computer and Communications Security*, 1997.

[21] F. Bergadano, D. Gunetti and C. Picardi, "User authentication through keystroke dynamics," in *ACM Transactions on Information and System Security 5*, 2002.

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

BlackWidow: Monitoring the Dark Web for Cyber Security Information

Matthias Schäfer
Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
schaefer@cs.uni-kl.de

Markus Fuchs
SeRo Systems
Kaiserslautern, Germany
fuchs@sero-systems.de

Martin Strohmeier
Cyber-Defence Campus
armasuisse
Thun, Switzerland
martin.strohmeier@armasuisse.ch

Markus Engel
SeRo Systems
Kaiserslautern, Germany
engel@sero-systems.de

Marc Liechti
Trivo Systems
Bern, Switzerland
marc.liechti@trivo.ch

Vincent Lenders
Cyber-Defence Campus
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Abstract: The Dark Web, a conglomerate of services hidden from search engines and regular users, is used by cyber criminals to offer all kinds of illegal services and goods. Multiple Dark Web offerings are highly relevant for the cyber security domain in anticipating and preventing attacks, such as information about zero-day exploits, stolen datasets with login information, or botnets available for hire.

In this work, we analyze and discuss the challenges related to information gathering in the Dark Web for cyber security intelligence purposes. To facilitate information collection and the analysis of large amounts of unstructured data, we present BlackWidow, a highly automated modular system that monitors Dark Web services and fuses the collected data in a single analytics framework. BlackWidow relies on a Docker-based micro service architecture which permits the combination of both pre-existing and customized machine learning tools. BlackWidow represents all extracted data and the corresponding relationships extracted from posts in a large knowledge graph, which is made available to its security analyst users for search and interactive visual exploration.

Using BlackWidow, we conduct a study of seven popular services on the Deep and Dark Web across three different languages with almost 100,000 users. Within less than two days of monitoring time, BlackWidow managed to collect years of relevant information in the areas of cyber security and fraud monitoring. We show that BlackWidow can infer relationships between authors and forums and detect trends for cybersecurity-related topics. Finally, we discuss exemplary case studies surrounding leaked data and preparation for malicious activity.

Keywords: *Dark Web analysis, open source intelligence, cyber intelligence*

1. INTRODUCTION

The Dark Web is a conglomerate of services hidden from search engines and regular Internet users. Anecdotally, it seems to the uneducated observer that anything that is illegal to sell (or discuss) is widely available in this corner of the Internet. Several studies have shown that its main content ranges from illegal pornography to drugs and weapons [1], [2]. Further work has revealed that there are many Dark Web offerings which are highly relevant for the cyber security domain. Sensitive information about zero-day exploits, stolen datasets with login information, or botnets available for hire [2], [3] can be used to anticipate, discover, or ideally prevent attacks on a wide range of targets.

It is difficult to truly measure the size and activity of the Dark Web, as many websites are under pressure from law enforcement, service providers, or their competitors. Despite this, several web intelligence services have attempted to map the reachable part of the Dark Web in recent studies. One crawled the home pages of more than 6,600 sites (before any possible login requirement), finding clusters of Bitcoin scams and bank card fraud [4]. Another study found that more than 87% of the sites measured did not link to other sites [5]. This is very different from the open Internet, both conceptually and in spirit: in contrast, we can view the Dark Web as a collection of individual sites or separated islands.

In the present work, we introduce BlackWidow, a technical framework that is able to automatically find information that is useful for cyber intelligence, such as the early detection of exploits used in the wild, or leaked information. Naturally, analyzing a part of the Internet frequented by individuals who are trying to stay out of the spotlight is a more difficult task than traditional measurement campaigns conducted on the Surface Web.

Thus, a system that seeks to present meaningful information on the Dark Web needs to overcome several technical challenges – a large amount of unstructured and inaccessible data needs to be processed in a scalable way that enables humans to collect useful intelligence quickly and reliably. These challenges range from scalability and efficient use of resources over the acquisition of fitting targets to the processing of different languages, a key capability in a globalized underground marketplace.

Yet, contrary to what is sometimes implied in media reports, few underground forums and marketplaces use a sophisticated trust system to control access outright, although some protect certain parts of their forums, requiring a certain reputation [6]. We successfully exploit this fact to develop an automated system that can gather and process data from these forums and make them available to human users.

In this work, we make the following contributions:

- We present and describe the architecture of BlackWidow, a highly automated modular system that monitors Dark Web services in a real-time and continuous fashion and fuses the collected data in a single analytics framework.
- We overcome challenges of information extraction in a globalized world of cyber crime. Using machine translation techniques, BlackWidow can investigate relationships between forums and users across language barriers. We show that there is significant overlap across forums, even across different languages.
- We illustrate the power of real-time intelligence extraction by conducting a study on seven forums on the Dark Web and the open Internet. In this study, we show that BlackWidow is able to extract threads, authors and content from Dark Web forums and process them further in order to create intelligence relevant to the cyber security domain.

The remainder of this work is organized as follows. Section 2 provides the background on the concepts used throughout, while Section 3 discusses the challenges faced during the creation of BlackWidow. Section 4 describes BlackWidow's architecture before Sections 5 and 6 respectively present the design and the results of a Dark Web measurement campaign. Section 7 discusses some case studies, Section 8 examines the related work and finally Section 9 concludes this paper.

2. BACKGROUND

In this section, we introduce the necessary background for understanding the BlackWidow concept. In particular, we provide the definitions and also explain the underlying technological concepts relating to the so-called Dark Web and to Tor Hidden Services.

A. The Deep Web and Dark Web

The media and academic literature are full of discussions about two concepts, the Dark Web and the Deep Web. As there are no clear official technical definitions, the use of these terms can easily become blurred. Consequently, these terms are often used interchangeably and at various levels of hysterics. We provide the most commonly accepted definitions, which can also be used to distinguish both concepts.

1) The Deep Web

The term 'Deep Web' is used in this work to describe any type of content on the Internet that, for various deliberate or non-deliberate technical reasons, is not indexed by search engines. This is often contrasted with the 'Surface Web', which is easily found and thus accessible via common search engine providers.

Deep Web content may, for example, be password-protected behind logins; encrypted; its indexing might be disallowed by the owner; or it may simply not be hyperlinked anywhere else. Naturally, much of this content could be considered underground activity, e.g., several of the hacker forums that we came across for this work were also accessible without special anonymizing means.

However, the Deep Web also comprises many sites and servers that serve more noble enterprises and information, ranging, for example, from government web pages through traditional non-open academic papers to databases where the owner might not even realize that they are accessible over the Internet. By definition, private social media profiles on Facebook or Twitter would be considered part of the Deep Web, too.

2) The Dark Web

In contrast, the Dark Web is a subset of the Deep Web which cannot be accessed using standard web browsers, but instead requires the use of special software providing access to anonymity networks. Thus, deliberate steps need to be taken to access the Dark Web, which operates strictly anonymously both for the user and the service provider (e.g., underground forums).

There are several services enabling *de facto* access to anonymity networks, for example the Invisible Internet Project (IIP) or JonDonym [7]. However, the so-called

‘Hidden Services’ provided by the Tor project remain the most popular *de facto* manifestation of the Dark Web. In the next section we provide a detailed technical explanation of Tor’s Hidden Service feature, which formed the basis of the analysis done by BlackWidow.

B. Tor Hidden Services

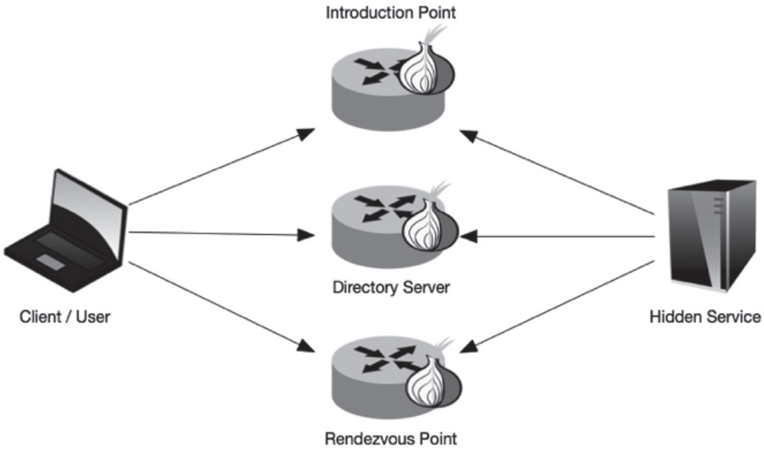
Tor, originally short for The Onion Router, is a project that seeks to enable low-latency anonymous communication through an encrypted network of relays. Applying the concepts of onion routing and telescoping, users obtain anonymity by sending their communication through a so-called *Circuit* of at least three relay nodes.

As Tor is effectively a crowdsourced network, these relays are largely run by volunteers. The network has been an important tool for many Internet users who depend on anonymity, from dissidents to citizens in countries with restricted Internet access. However, there have been many vulnerabilities found and discussed in the literature which could lead to deanonymization of Tor users. As it is not desired to authenticate the identity of every Tor relay, it is widely considered possible that state actors such as intelligence agencies run their own relay nodes, by which they may exploit some of these vulnerabilities in order to deanonymize users of interest [8]. Despite these potential threats, Tor is the best-known and most popular way to hide one’s identity on the Internet.

Besides enabling users to connect to websites anonymously, Tor offers a feature called *Hidden Services*. Introduced in 2004, it adds anonymity not only to the client but also to the server, also known as responder anonymity. More concretely, by using such Hidden Services, the operator of any Internet service (such as an ordinary web page, including forums or message boards, which we are interested in for this work) can hide their IP address from the clients perusing the service. When a client connects to the Hidden Service, all data is routed through a so-called *Rendezvous Point*. This point connects the separate anonymous Tor circuits from both the client and the true server [9].

Figure 1 illustrates the concept: overall, there are five main components that are part of a Hidden Service connection. Besides the Hidden Service itself, the client and the Rendezvous Point, it requires an *Introduction Point* and a *Directory Server*.

FIGURE 1. GENERAL ILLUSTRATION OF THE TOR HIDDEN SERVICE CONCEPT.



The former are Tor relays, which forward management information necessary to establish the connection via the Rendezvous point and are selected by the Hidden Service itself, which is necessary to connect the client and the Hidden Service at the Rendezvous point. The latter are Tor relay nodes, where Hidden Services publish their information and which are then communicated to clients in order to learn the addresses of the Hidden Service's introduction points. These directories are often published in static lists and are in principle used to find the addresses for the web forums used in BlackWidow.

It is unsurprising that Tor Hidden Services are a very attractive concept for all sorts of underground websites, such the infamous Silk Road or AlphaBay and due to their popularity form in effect the underlying architecture of the Dark Web.

3. CHALLENGES IN DARK WEB MONITORING

The overarching main issues in analyzing the Dark Web for cyber security intelligence relate to the fact that a vast amount of unstructured and inaccessible information needs first to be found and then processed. This processing also needs to be done in a scalable way that enables humans to collect useful intelligence quickly and reliably. In the following, we outline the concrete challenges that needed to be overcome in developing BlackWidow.

A. Acquisition of Relevant Target Forums

The first challenge is the identification of target forums that are relevant to our operation, i.e. those that contain users and content relating to cyber security intelligence. Due to the underground nature of the intended targets, there is no curated list available that could be used as input to BlackWidow. Intelliagg, a cyber threat intelligence company, recently attempted to map the Dark Web by crawling reachable sites over Tor. They found almost 30,000 websites; however, over half of them disappeared during the course of their research [1], illustrating the difficulty of keeping the information about target forums up to date.

Combined with the mentioned previously fact that 87% of Dark Web sites do not link to any other sites, we can deduce that the Dark Web is more a set of isolated short-lived silos than the classical Web, which has a clear and stable graph structure. Instead, only loose and often outdated collections of URLs (both from the surface Internet as well as Hidden Services) exist on the Dark Web. Consequently, a fully automated approach to overcome this issue is infeasible and a semi-manual approach must initially be employed.

B. Resource Requirements and Scalability

Several technical characteristics of the acquired target forums require the use of more significant resource inputs. As is typical in analyzing large datasets obtained from the Dark Web, it is necessary to manage techniques which limit the speed and the method of access to the relevant data [10].

Such techniques include the deliberate (e.g., artificial limiting of the number of requests to a web page) and the non-deliberate (e.g., using active web technologies such as NodeJS, which break the use of faster conventional data collection tools). Typically, these issues can be mitigated by expending additional resources. Using additional virtual machines, bandwidth, memory, virtual connections or computational power, we can improve the trade-off with the time required for efficient data collection. For example, by using several virtual private networks (VPNs) or Tor circuits, it is possible to parallelize the data collection in case there is a rate limit employed by the target.

Surprisingly, a factor not challenging our resources was the habit of extensively vetting the credentials or ‘bona fides’ of forum participants before allowing access. A sufficient number of the largest online forums are available without this practice, which enabled data collection and analysis without having to manually circumvent such protection measures. However, since we did encounter at least some such forums (or parts of forums), our approach could naturally be extended to them, although this would require significant manual resource investment.

C. Globalized Environment

As cyber security and cyber crime have long become a global issue, underground forums with relevant pieces of information are available in practically all languages with a significant number of speakers. Most existing studies of Dark Web content have focused on English or another single language (e.g., [2]). However, the ability to gather and combine information independent of the forum language broadens the scope and the scale of BlackWidow significantly. By employing automated machine translation services, we are able to not only increase the range of our analysis but also detect relationships and common threads and topics across linguistic barriers and country borders.

Naturally, this approach comes with several downsides. For example, it is not possible to employ sentiment or linguistic analysis on the translated texts nor is the quality of state-of-the-art machine translation comparable to the level of a human native speaker. However, given BlackWidow's aims of scalable and automatic intelligence gathering, these disadvantages can be considered an acceptable trade-off.

D. Real-Time Intelligence Extraction

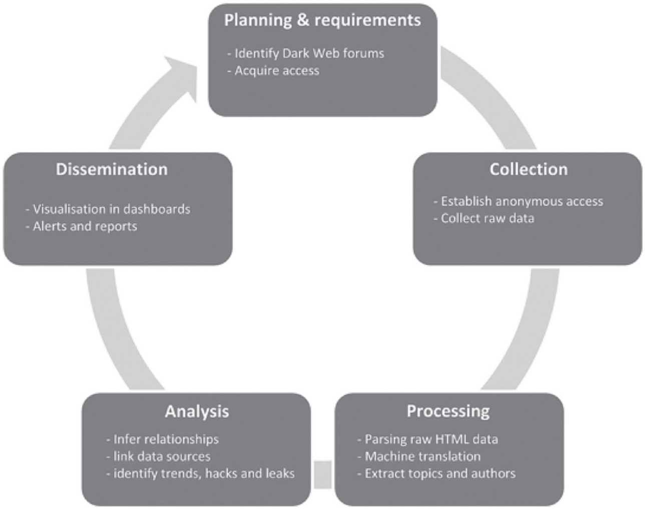
Beyond the previous issues, BlackWidow focuses in particular on the challenges posed by the nature of a real-time intelligence extraction process. Whereas previous studies have collected data from the Dark Web for analytical purposes, they have typically concentrated on a static environment. In contrast to collecting one or several snapshots of the target environment, BlackWidow aims to provide intelligence and insights much faster. Real-time capability is a core requirement for the longer-term utility of the system, due to the often very limited lifetime of the target forums.

To enable these functionalities, a high grade of automation is required, from the collection to the live analysis of the data. After the initial bootstrapping of sources and creating a working prototype, it is imperative that the processes require less manual input beyond normal human oversight tasks.

4. ARCHITECTURE OF BLACKWIDOW

In this section, we describe the basic architecture of BlackWidow. We largely abstract away from the exact technologies used and focus on the processing chain and the data model that enabled us to analyze the target forums in real time. Figure 2 shows the processing chain, including five phases defined as a recurrent cycle. The phases of the cycle are highly inspired by the conceptual model of the intelligence cycle [11]. Like the intelligence cycle, theses phases are continuously iterated to produce new insights.

FIGURE 2. BLACKWIDOW PROCESS CYCLE.



A. Planning and Requirements

The key focus of BlackWidow is on automation; manual work should only be needed for the integration of target forums in the initial planning and requirements phase, while all other phases are highly automated.

1) Identifying Dark Web Forums

The first suitable target forums are identified by hand to bootstrap the process and overcome the challenges described in Section 3.A. After obtaining a foothold, BlackWidow then aims to analyze the content of these forums in order to obtain further links and addresses to other targets in a more automated fashion in later iterations.

2) Gaining access

Since most forums require some sort of login to access the site, BlackWidow needs personal accounts to authenticate on each site. The way to acquire such logins differs on each site. While certain sites only request new users to provide a valid email address, others have higher entry barriers with reputation systems, measures of active participation, or even requiring users to first buy credits.

B. Collection

After the planning and requirements phase, all steps are fully automated. The collection phase deals with establishing anonymous access to the forums over Tor and the collection of raw data.

1) Establishing anonymous access to forums

We establish anonymous gateways to the identified forums using Docker containers, Tor to access Hidden Services and Virtual Private Networks (VPN) for regular Deep Web sites. Here, it is necessary to add custom functions to BlackWidow, which emulate typing and clicking behavior in order to log in automatically and subsequently detect whether the gateway has successfully logged into the target or not.

2) Collection of raw data

For the actual collection of the forum content and metadata, we employed the node.js headless Chrome browser puppeteer [12] as a crawler within the Docker containers. While it requires more resources than other collection methods, it more closely emulates the behavior of real forum users, meaning it more easily avoids defensive action by the Dark Web marketplace operators. In order to improve the speed, the collection is distributed across multiple containers and parallelized.

C. Processing

The processing phase deals with parsing the collected raw HTML data from the previous phase, translating the content into English and extracting the entities of interest to feed a knowledge graph.

1) Parsing raw HTML data

Since BlackWidow retrieves data over a headless browser, the data to process is in the Hypertext Markup Language (HTML). Extracting structured information from HTML data can be quite challenging depending on the layout of the forums. BlackWidow implements a standard HTML parser that we adapt to the layout of each forum. While this approach may seem expensive at first, many forums have a similar layout such that the same parsers can be reused for different forums. The output of the HTML parser for each page is a structured file including only the text information from the HTML page.

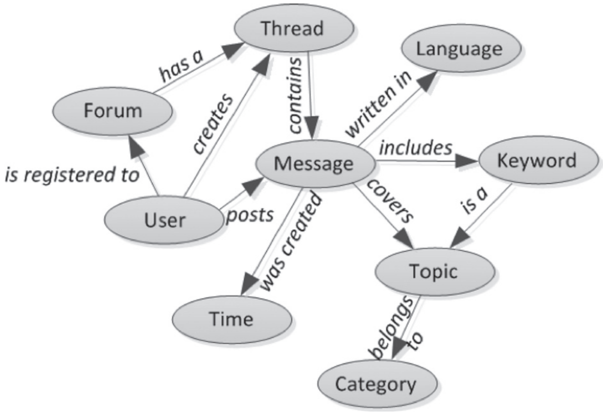
2) Translation of raw data in foreign languages

As much of the collected raw data contains content in several languages, we used automated machine translation to convert all non-English content into English. Through the use of Google's translation API, we obtain state-of-the-art translations, which enable the more complex data modeling and relationship analysis over forums in different languages in the follow-up phases.

3) Information extraction

To extract relevant information from the translated text from the gateways, we developed so-called extractors in Scala, which were also processed in a distributed fashion using the Apache Spark analytics framework. BlackWidow extracts information about the forum writers and their content, i.e. the titles of forum threads and the posted messages. It then constructs a knowledge graph that connects threads, actors, messages and topics. Figure 3 shows the underlying data model of the knowledge graph of BlackWidow. The collected raw data and the knowledge graph is then put into Elasticsearch, a search engine based on Lucene [13]. As a tool for data exploration, it reads structured data and interprets timestamps and locations.

FIGURE 3. DATA MODEL REPRESENTING THE KNOWLEDGE GRAPH IN BLACKWIDOW.



D. Analysis

While inferring simple relationships between messages and authors is a relatively easy task given the HTML structure of the forums, other types of relationships and information extraction steps for the knowledge graph require advanced data analysis techniques. BlackWidow's goal is to automatically find relationships and trends across different threads and forums; the following processing steps are thus executed in this phase.

1) Infer user relationships

Relationships between users are mainly inferred in BlackWidow through the analysis of threads, since users of Dark Web forums barely link to each other explicitly as in classical social networks such as Facebook or LinkedIn [6]. A thread is always created by a single user and many different users then start posting messages on this thread. BlackWidow infers a relationship between users by ordering all messages in the same thread by their message times. We define a relationship from user A to user B if user B posted a message after user A in the same thread. The intuition is that user B is interacting with user A when he replies to his messages.

2) Identify topics

While messages in forums are commonly structured in threads and categories, it is not always obvious to see which threads cover the same topics. To facilitate trend analysis across different threads and forums, BlackWidow automatically identifies topics by means of automatic topic modeling. BlackWidow implements unsupervised text clustering techniques based on Latent Dirichlet Allocation (LDA) to classify messages into groups. These groups are then assigned to higher-level categories of interests such as botnets, databases, exploits, leaks and DDoS.

3) Identify cyber security trends

To identify cyber security trends, BlackWidow fuses the messages, topics and categories from the different forums and computes aggregated time series. These time series form the basis to identify trending topics, e.g. when the time series experiences a high growth or decline over short periods. Long-term trends are also detectable given that all collected messages are time-stamped and thus provide information over the whole lifetime of the forum.

E. Dissemination

Finally, it is important to disseminate the extracted information so that it can be easily processed by human intelligence analysts. To serve this purpose, BlackWidow supports various types of data visualizations and data query interfaces for exploratory analysis. For example, customized Kibana dashboards provide real-time views of the processed data that is stored in the Elasticsearch database. These dashboards can be generated and customized easily by the users allowing different views depending on the question of interest.

Finally, users may realize that some data is missing or that the additional forums should be integrated. The cycle of BlackWidow's architecture supports users to refine the planning and data collection requirements, thus closing the loop of the intelligence process.

5. STUDY DESIGN

After describing the architecture of BlackWidow, we now explain the goals of the study conducted for this paper. The study was designed to show the power and effectiveness of our automated data extraction and analysis efforts for the Dark Web.

A. Information Extraction

Forum contents are usually structured hierarchically. Users provide or exchange information by posting messages, known as "posts". Collections of posts belonging to the same conversation are called threads. Threads can be separated by categories such as “Drugs”, “Exploits”, or “Announcements”. Besides the actual message, posts also provide meta information on the author (e.g., username, date of registration) and the exact date and time when the message was posted.

While posts are certainly the most interesting source of information in a forum, it is worth taking other parts of the forum into account for information retrieval as well. For example, most forums have a publicly available list of members which provides links to the profiles of all users registered in the forum. By additionally crawling the public profiles of all registered users, it is possible to gather information on passive users and the overall community as well. User profiles often provide useful information, such as registration date and time of last visit.

To extract all this information from the HTML-based forum data collected by BlackWidow, we implemented HTML parsers for each forum based on jsoup. Although forums generally have a very similar structure, the underlying HTML representations differ significantly depending on the platform. The consequence is that for each different forum platform (e.g., vBulletin), a separate forum parser is required.

For this analysis, we limit our implementation to parsing posts and user profiles. Our parsers transform the HTML-based representation of posts and user profiles into a unified JSON-based format. More specifically, each post is transformed into a JSON object with attributes forum, category, thread, username, timestamp and message. Objects from non-English forums are extended with the English translations of categories, threads and messages. User profiles are parsed into JSON objects with attributes forum, username, registration date and (where available) last visit date.

B. Forum Selection

For the purpose of this study, we collected data from seven forums as a proof of concept, as the manual integration of new forums can require significant time investment. At the time of writing, roughly one year after collecting the data, only four of the scanned forums are still online, confirming the short lifetime and high volatility of such forums. Overall, three of the seven forums were only accessible in the Dark Web and four were Deep Web forums. The languages used in the forums were Russian, English and French. An overview over the considered forums and the most popular categories (by number of posts) is provided in Table 1.

TABLE 1: OVERVIEW OF THE FORUMS CONSIDERED IN OUR ANALYSIS.

#	Type	Language	Top Categories	Online as of 12/ 2018
Forum 1	Deep Web	English	News, Porn, Software, Drugs	Yes
Forum 2	Deep Web	Russian	Marketplace, Electronic Money, Hacking	No
Forum 3	Dark Web	French	Drugs, News, Porn, Technology	No
Forum 4	Dark Web	Russian	Marketplace, General Discussions, Hacking, Security	Yes
Forum 5	Deep Web	English	Gaming, Leaks, Cracking, Hacking, Monetizing Techniques, Tutorials	Yes
Forum 6	Dark Web	French	News, Frauds, Conspiracy Theories, Drugs, Crime	No
Forum 7	Deep Web	Russian	Software, Security & Hacking, DDoS Services, Marketplace	Yes

6. STUDY RESULTS

A. Target Analysis

The size of each forum can be determined either in the number of posts or in the number of users. Both metrics for the crawled forums are shown in Figure 4 and 5. Forum 5 has by far the largest community with 67,535 registered users, while Forum 3 has (also by a considerable margin) the most content with over 288,000 posts. Forum 3 is also the forum with the most active community in terms of average posts per user. On average, each user had posted 22.74 messages in Forum 3. In contrast, the community of Forum 5 seemed to consist largely of passive users, since for each user, there were only 2.28 messages, roughly one tenth of those in Forum 3.

FIGURE 4. NUMBER OF USERS EXTRACTED FROM EACH FORUM.

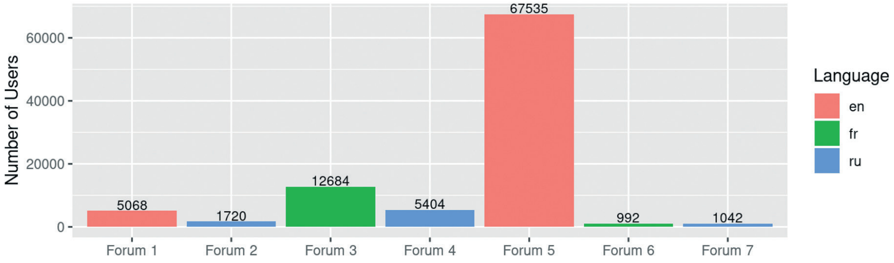
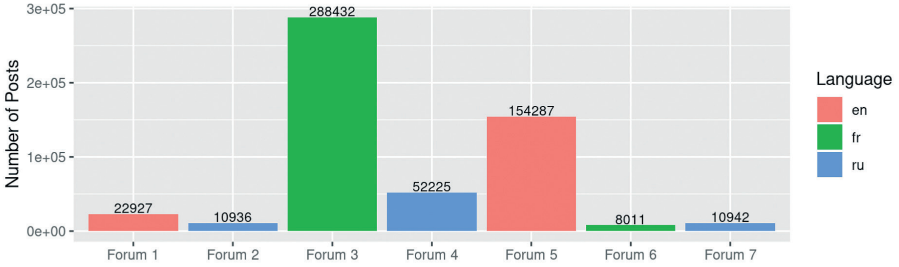


FIGURE 5. NUMBER OF POSTS EXTRACTED FROM EACH FORUM.



We hypothesize that the extremely large number of passive users in Forum 5 comes from the fact that the forum is a Deep Web forum, meaning that it does not require users to use additional software (such as the Tor browser) to sign up. As a consequence of this significantly lower technical hurdle, is can be accessed much more easily than Dark Web forums and is therefore open to a broader, less tech-savvy audience.

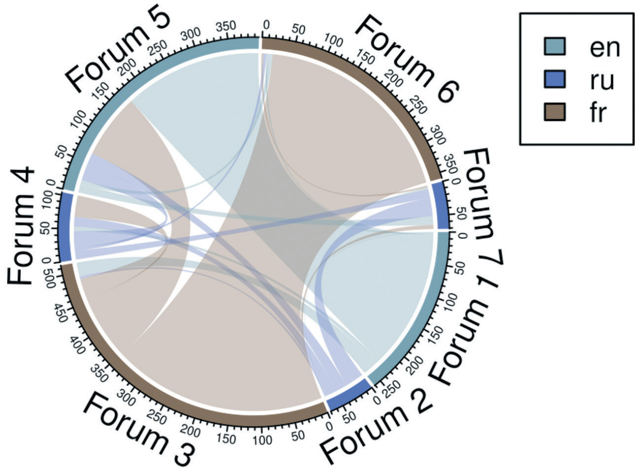
B. Forum Relationships

In order to get some insights on the relationships between the forums, we compared the sets of usernames of the forums. More specifically, we were interested in the intersections of these sets to see whether these forums host separate communities or whether there are significant overlaps. Surprisingly, those usernames that appeared most often were very specific, suggesting that they actually belonged to the same person. In fact, generic usernames such as "admin" or "john" were very rarely seen. Instead, users tended to individualize their usernames, for example by using *leetspeek*,[1] most likely as a means of anonymous branding. This tendency benefits the social network analysis conducted in this section since it provides us with reliable information about individual users, even across forums.

The result of this analysis is depicted as a chord diagram in Figure 6. Unsurprisingly, there are significant overlaps across forums in the same language. More interesting, however, is the fact that Forum 5, the forum with the largest community, has significant overlaps with most other forums, even if they are in a different language. By looking at these intersections as information dissemination channels, Forum 5 certainly provides the best entry point to spread information across the deep and dark side of the web.

[1] A system of modified spelling, whereby users replace characters with resembling glyphs.

FIGURE 6. RELATIONSHIPS BETWEEN THE FORUMS IN TERMS OF COMMON USERS.



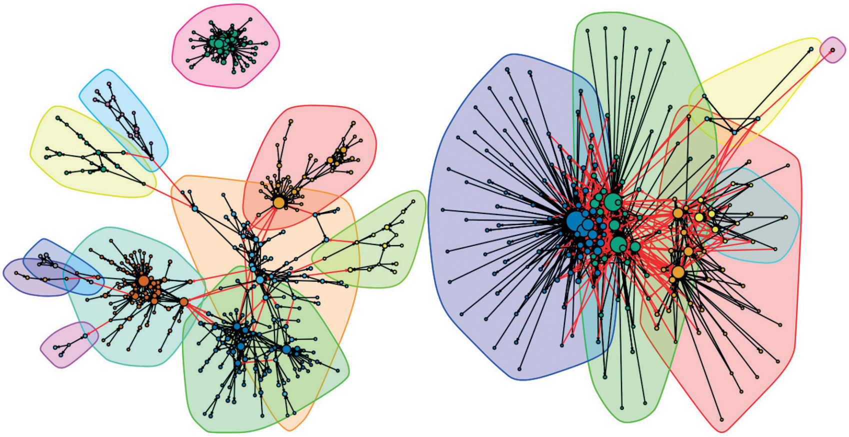
C. Author Relationships

In order to analyze the internal relationships between users of forums, we first need to establish a reasonable definition of user relationships. While there are clearly defined relationships in social networks such as Facebook or Twitter, forum users do not have natural links such as friendships or followers. Given the hierarchical structure of forums, however, we can identify users with common interests by looking at the threads in which they are active together. We therefore define the relationships between two users in a forum by the number of threads in which both users posted messages.

Based on the creation timestamp of each post, we can also add a direction to this relationship by acknowledging which user merely reacted to a post of another user; i.e., which user posted a message in the same thread at a later point in time. This directed relationship will help us distinguish information or service providers from consumers. This is possible since a common communication pattern, for example in forum-based marketplaces, is that someone shares data or services in a new thread and interested users must post a reply (e.g., “thank you”) in order to access the shared content.

After these relationships were established, we used the Walktrap community-finding algorithm [14] with a length of 4 to determine sub-communities in the forums. These sub-communities evolve naturally since forums often cover many different unrelated topics. For example, users interested in drugs might not be interested in hacking and vice versa, resulting in two sub-communities.

FIGURE 7. NETWORK SHOWING THE RELATIONSHIPS BETWEEN USERS OF FORUM 4 (LEFT) AND 5 (RIGHT).



The result of this analysis for Forums 4 and 5 is shown in Figure 7. The vertices in the graphs represent the individual users, while the (directed) edges show the relationships as defined above. Each sub-community is indicated by a color. The size of each node in the network represents the number of incoming edges, i.e., its degree. In comparison, the structural differences of the communities of the two forums are clearly visible. Forum 4, which has a much smaller community, is much denser, meaning that there are many more relationships between users, even across the different sub-communities.

The network analysis enabled us to select sub-communities and identify their key users, i.e., the most active information or service providers. For instance, the completely separate sub-community in Forum 5 is a group of so-called skin gamblers, i.e., people who bet virtual goods (e.g., cryptocurrencies) on the outcome of matches or other games of chance. Another sub-community in Forum 5 deals with serial numbers of commercial products, with one user being a particularly active provider.

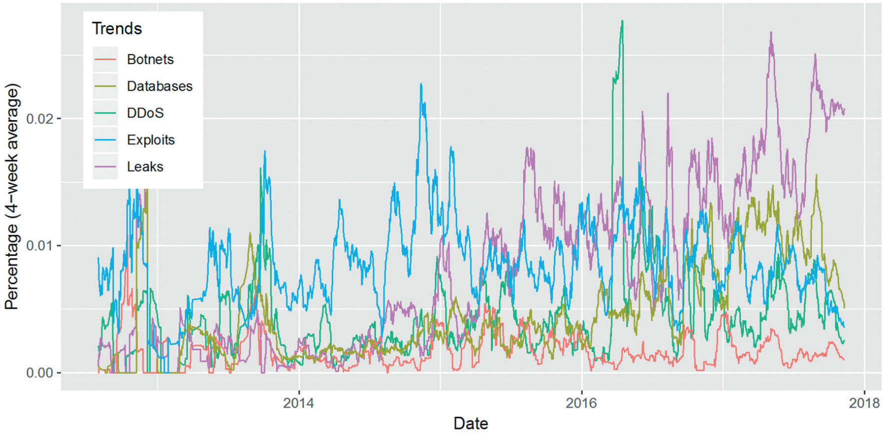
It is worth noting that, besides active providers, forum administrators and moderators also stick out in terms of node degree (activity) as they post a lot of administrative messages. For example, one user was very prominent in a sub-community and by manually checking his posts we found that he was a very active moderator who enforced forum rules very strictly and made sure transactions were being handled correctly. His power to enforce rules and certain behavior was established by a system of reputation, in which users must gain hard-earned reputation points, for example by posting free content or being an active community member over a long period of time. Once a certain reputation is earned by a user, it becomes much easier for her or him to sell products on the marketplace; or they can charge higher prices as the risk of scam for buyers is lower. This system provides administrators and moderators with a certain leverage, since a ban from the forum would mean a complete loss of hard-earned reputation.

7. CYBER SECURITY INTELLIGENCE EXAMPLES

After conducting our quantitative study, we now discuss some exemplary trends and case studies that we noticed using BlackWidow during its initial deployment in 2017 to collect and analyze forum datasets dating back to 2012.

A. Forum Trends

FIGURE 8. CYBERSECURITY TRENDS BETWEEN 2012 AND 2017 IN SEVEN FORUMS AS OBSERVED BY BLACKWIDOW.



It is possible to use BlackWidow's functionality to look at the popularity of different concepts over time, which can aid the intelligence analyst in finding sudden anomalies or identifying trends that suggest increased or decreased importance. Figure 8 shows the fraction of all posts for the five most popular identified cyber security categories over a time frame of five years from the end of 2012 to the end of 2017. The time series are generated using the running mean of the number of posts in the respective topics over time and normalized with respect to the overall activity in the considered period. The topic assignment is based on regular expressions and string matching.

From this, we can see a substantial change in the number of times that forum actors were discussing *leaks*, which increases roughly ten-fold in 2017 and outpaces the other groups in number of mentions significantly by the end of the period. Related to leaks, posts on *databases* seem to become increasingly popular, while talk of *exploits* remains more or less constant as a trend, with several peaks, e.g. at the end of 2014. *DDoS* and *botnets* are the least popular of the five; the significant DDoS peak in the beginning of 2016 was caused by one of the analyzed forums itself being the victim of a DoS attack.

B. Discovered Leaks and Exploits

During the course of our study, we encountered various data leaks consisting of usernames and passwords. As an example, BlackWidow crawled links to a list of half a million leaked Yahoo! accounts, a well-known dataset from a hack in 2014 (officially reported by Yahoo! in 2016). Perhaps surprisingly, these leaked datasets were accessible for free through direct links, highlighting again that security-relevant information can indeed be automatically collected by BlackWidow without interacting personally with criminal data brokers.

Exploits for various platforms were also found abundantly. Again, the open nature of the forums makes it possible to collect large amounts of exploits for free. While a systematic analysis on the quality and novelty of the individual exploits is outside the scope of this paper, we are confident that BlackWidow constitutes a very useful data source to better understand the cyber threat landscape and anticipate exploits that may be expected in the wild. Security professionals and defenders should therefore aim at analyzing such information to anticipate emerging threats.

8. RELATED WORK

Web forums inside and outside the Dark Web have been an active field of research in the recent past, with authors approaching them from a wide variety of angles, including cyber security and intelligence.

The closest works to ours relate to underground crawling systems. Pastrana et al. [6] recently built a system that looks at cyber crime outside the Dark Web. The authors discuss challenges in crawling underground forums and analyze four English-speaking communities on the Surface Web. In contrast, Nunes et al. [15] mine Dark Web and Deep Web forums and marketplaces for cyber threat intelligence. They show that it is possible to detect zero-day exploits, map user/vendor relationships and conduct topic classification on English-language forums, results that we have been able to reproduce with BlackWidow.

Benjamin et al. [16] explore cyber security threats in what they call the “hacker web”, with a focus on stolen credit card data activity but also potential attack vectors and software vulnerabilities. The authors extract data from carding shops, the Internet Relay Chat (IRC) and web forums, but do not investigate Tor Hidden Services.

In [17] and [18], the authors look at major hacker communities in the US and China, aiming to identify key players, experts and relationships in open web forums. They base their approach on a framework for automated extraction of features using text analytics and interaction coherence analysis. Similarly, Motoyama et al. [19] look at six different underground forums on the open web, providing a measurement campaign on historical data. The extensive quantitative data analysis covers features from the top content over the size of the overlapping user base to interactions and relationships between the users. However, their analysis is based on leaked SQL dumps of the forums, while BlackWidow is a framework that collects information in real time through the frontend of the forums.

Outside the academic literature, we find several commercial enterprises which aim to conduct automated analysis of cyber security intelligence from the Dark Web, among other sources. Two examples are provided by DarkOwl [20] and Recorded Future [21], which monitor the Dark Web in several languages and offer to detect threats, breached data and indicators of compromise.

To the best of our knowledge, this paper is the first to discuss real-time data collection in the Deep and Dark Web and the integration of external translation capabilities in a scalable way. Additionally, our results have been able to show that there is substantial overlap between actors across forums, even if they are not in the same language.

9. CONCLUSION

It is imperative in the current cyber security environment to have a real-time monitoring solution that works across languages and other barriers. We have shown in this paper that early detection of cyber threats and trends is feasible by overcoming several key challenges towards a comprehensive framework.

While we can be fairly certain that techniques similar to ours are being used by both governmental and private intelligence actors around the world, it is important to analyze their power in a more open fashion, giving rise to possible scrutiny and further development. By implementing BlackWidow as a proof-of-concept collection and analysis tool, we show that monitoring of the Dark Web can be done with relatively little resources and time investment, making it accessible to a broader range of actors in the future.

REFERENCES

[1] Intelliagg, “Deeplight: Shining a Light on the Dark Web. An Intelliagg Report,” 2016.

[2] M. W. Al Nabki, E. Fidalgo, E. Alegre and I. de Paz, “Classifying illegal activities on TOR network based on web textual contents,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017.

[3] A. Biryukov, I. Pustogarov and R.-P. Weinmann, “Trawling for tor Hidden Services: Detection, measurement, deanonymization,” in *IEEE Symposium on Security and Privacy (S&P)*, 2013.

[4] Hyperion Gray, “Dark Web Map,” [Online]. Available: https://www.hyperiongray.com/dark-web-map/. [Accessed 7 1 2019].

[5] V. Griffith, Y. Xu and C. Ratti, “Graph Theoretic Properties of the Darkweb,” *arXiv preprint arXiv:1704.07525*, 2017.

[6] S. Pastrana, D. R. Thomas, A. Hutchings and R. Clayton, “CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale,” in *Proceedings of the 2018 World Wide Web Conference*, 2018.

[7] A. Pescape, A. Montieri, G. Aceto and D. Ciuonzo, “Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web),” *IEEE Transactions on Dependable and Secure Computing*, 2018.

[8] K. Bauer, D. McCoy, D. Grunwald, T. Kohno and D. Sicker, “Low-resource routing attacks against Tor,” in *Proceedings of the ACM Workshop on Privacy in Electronic Society*, 2007.

[9] A. Biryukov, I. Pustogarov, F. Thill and R.-P. Weinmann, “Content and popularity analysis of Tor Hidden Services,” in *IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)* , 2014.

[10] I. Sanchez-Rola, D. Balzarotti and I. Santos, “The onions have eyes: A comprehensive structure and privacy analysis of Tor Hidden Services,” in *Proceedings of the 26th International Conference on the World Wide Web*, 2017.

[11] L. K. Johnson, Ed., *Handbook of intelligence studies*, Routledge, 2007.

[12] “Puppeteer,” [Online]. Available: https://pptr.dev. [Accessed 7 1 2019].

[13] Elastic, “Elasticsearch,” [Online]. Available: https://www.elastic.co/products/elasticsearch. [Accessed 7 1 2019].

[14] P. Pons and M. Latapy, “Computing communities in large networks using random walks.,” *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp. 191-218, 2006.

[15] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart and P. Shakarian, “Darknet and deepnet mining for proactive cybersecurity threat intelligence,” in *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016.

[16] V. Benjamin, W. Li, T. Holt and H. Chen, “Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops,” in *IEEE International Conference on Intelligence and Security Informatics (ISI),* 2015.

[17] A. Abbasi, W. Li, V. Benjamin, S. Hu and H. Chen, “Descriptive analytics: Examining expert hackers in web forums,” in *IEEE Joint Intelligence and Security Informatics Conference (JISIC)*, 2014.

[18] V. Benjamin and H. Chen, “Securing cyberspace: Identifying key actors in hacker communities,” in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2012.

[19] M. Motoyama, D. McCoy, K. Levchenko, S. Savage and G. M. Voelker, “An analysis of underground forums,” in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference,* 2011.

[20] “DarkOwl,” [Online]. Available: https://www.darkowl.com. [Accessed 7 1 2019].

[21] “Recorded Future,” [Online]. Available: https://www.recordedfuture.com. [Accessed 7 1 2019].

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Towards Measuring Global DDoS Attack Capacity

Artūrs Lavrenovs
NATO CCD COE
arturs.lavrenovs@ccdcoe.org

Abstract: In today's Internet, distributed denial-of-service (DDoS) attacks play an ever-increasing role and constitute a risk to any commercial, military or governmental entity that has a presence on the Internet or simply has an Internet connection. To address this threat on all levels, decision-makers have to rely on trustworthy information regarding attack capacity, sources, and the largest contributors. The lack of this information limits the ability of technicians, policymakers, and other relevant decision-makers to remediate the issue as efficiently as possible.

This research introduces a methodology for measuring the properties of individual devices participating in such attacks. These properties include rate limiting, amplification factor, and speed, which allows the calculation of each device's actual contribution to the attack capacity. This methodology was implemented as a proof of concept for the NTP protocol and the results indicate that it has promising potential. Individual measurements aggregated together provide insights into particular abused protocols: all the protocols together could provide the global DDoS attack capacity.

Keywords: *DDoS, attack capacity measurement, global DDoS attack capacity*

1. INTRODUCTION

Distributed denial of service (DDoS) attacks have been plaguing the Internet almost since its inception. Although the first large-scale network DDoS attack happened in 1999 [1], DDoS is still a serious and even growing threat to Internet-connected organizations. DDoS attacks have become almost daily news and created a large cybercrime industry offering DDoS attacks as a service as well as an immense cyber defense industry providing network filtering services, software, and hardware. Reasonable observers without any computer networking or cybersecurity background would assume that this issue has been and is currently being addressed properly to eliminate it at its root cause. The reality is that DDoS attacks have been on the rise with the increase in Internet connection speeds but mitigation efforts have only slowed down the total growth of the attacks.

DDoS attacks have become almost a household word, because when an online gaming platform or other popular resource goes offline because of continuous attacks, tens of millions of users are affected. Cybersecurity and network specialists are well aware of the attack properties, and many of the decision-makers are aware of the risks. Why, then, has this problem not been tackled in a global or at least a national manner? Due to the nature of the Internet, DDoS attacks transcend national borders and although they are illegal, there are no technical means to stop them at a national level. There must be a push for international policy from the highest-level decision-makers. These efforts cannot be made without being information-driven. The underlying causes are well-known but the question that is missing an answer is, what is the current status: total attack capacity, geographical regions and legal entities contributing the most? This information, presented in an easily digestible way together with associated risks, should be useful for non-technical decision-makers to justify taking action.

The kind of data needed and how to acquire it is investigated and the methodology for producing the missing information is proposed in this paper, resulting in the development of a proof of concept for the NTP protocol.

2. RELATED WORK

DDoS attacks are widely discussed and researched in academia. Although raw data is significantly less available to researchers than to commercial and other entities that receive DDoS attacks themselves, in some cases researchers make special agreements to access it. Attack detection and defense is a significantly explored topic; in the real world these solutions most often involve basic statistical analysis of incoming traffic. However, researchers are trying a wide range of old and new technologies like machine learning, software-defined networks, etc. to achieve better results. Less relevant research topics such as motivation, financial and criminal aspects are not reviewed here.

Protocol analysis is the default method for identifying protocols that could be abused in the future. Analysis of the protocol definition, documentation and source code of different implementations can allow researchers to identify new potentially abusable services. Some assumptions or previous research into the prevalence of the analyzed protocol must be made in advance to choose which of the many protocols to pick for analysis. If only a few devices with abusable services are found, then the overall impact for DDoS attack is negligible and malicious actors might not even bother exploiting it. Correct responsible disclosure mandates security researchers to report discovered vulnerabilities in advance to hardware and software vendors and other parties that would be responsible for issue remediation. In theory, this would preemptively mitigate the abuse of specific protocols, but the real situation is quite different. Research publications and vulnerability reports disclosed after the time period given to vendors still enable malicious actors to exploit reflection from devices that are presently not mitigated. One of the most prominent such cases was NTP DDoS. Rossow evaluated common UDP-based protocols and observed that most NTP implementations support a command to return client list that was a feature of the implementation and not defined in the protocol itself. The measured amplification factor was up to 4670, which was the largest of those measured in the research [2]. Because of the potential for abuse, the researcher conducted responsible disclosure to the security community and appropriate vendors. However, either directly due to this disclosure or inferred through released software fixes, malicious actors started exploiting it in the wild.

Attack analysis covers newly abused protocols or protocol features combining data from attack monitoring points, Internet scan data, backscatter, and other sources that present an integrated overview of the specific protocol. The success of mitigation efforts can be evaluated by notifying the system owners and continuously monitoring changes in the number of abusable devices. Czyz et al. investigated NTP DDoS in detail, additionally exploring unique protocol features that provide insight into victims [3]. This type of research does not contribute to overall DDoS attack capacity knowledge, as the produced estimates are for a fixed point in the past, possibly at the peak of the attacks, and quickly become outdated.

The ability to spoof the IP address of packets is the main cause of multiple types of attacks, including the most problematic: reflected DDoS. The Center for Applied Internet Data Analysis (CAIDA), based at the University of California’s San Diego Supercomputer Center, has been conducting research into the state of IP spoofing and continuously monitoring since 2008. The CAIDA spoofer project publishes updated and historical data from their measurements. In total, 22.6% of the IPv4 AS not using NAT were spoofable in July 2018, which corresponded to 14.3% of IP address blocks [4]. In general, countries in developing regions are found to be proportionally more spoofable than those in developed countries. However, in absolute numbers, the USA has most of the spoofable IP blocks.

A. Capacity Measurement

Measuring attack capacity is not sufficiently investigated. Currently, the only methodology for measuring the overall worldwide capacity for DDoS attacks is published in the scientific literature by Leverett et al. [5]. Researchers analyzed only reflected volumetric UDP DDoS attacks, thus closely relating to this research. More specifically, four protocols were analyzed – NTP, DNS, SSDP, and SNMP. Using this methodology, it was concluded that the total estimated DDoS attack capacity is 108.49 Tb/s. As the authors acknowledged themselves, this figure is limited by factors not explored in detail; thus, in reality, it is significantly lower. This figure does not take into account the ability of the AS network to handle all the capacity at same time, device load, existing bandwidth utilization, and device computational power that might not be able to handle producing responses to fully utilize the whole available network connection.

In addition to the total capacity estimate, additional avenues to present and visualize data for easier consumption by non-technical policy-makers were explored, e.g., a map of the world with the risk posed to others attached to each individual country. This visualization allowed the important discovery that developed countries actually possess higher DDoS attack capacity than developing countries. This finding points to the lack of a policy to, at the very least, mitigate DDoS attacks or its enforcement even in developed countries. Instead of pointing the finger at developing countries, this issue should be addressed internally and at an international level.

B. Industry Research

Case studies analyzing individual attacks are occasionally published online by commercial entities receiving or mitigating DDoS attacks. This usually happens when a new protocol has started getting abused or when previous attack records are broken. The motivation behind these case studies is to advertise the ability to handle DDoS attacks to gain more clients, and the details provided in the case studies are usually very restricted so as not to reveal any commercial information or weak points in the defenses. However, these case studies have become the main point of reference when discussing DDoS attack capacity. When the question is, ‘What is the maximum realistic DDoS attack capacity?’ the answer that follows usually refers to the latest or recent published attack case study. At the time of writing this paper, the case study by Arbor reported a maximum observed DDoS attack capacity of 1.7 Tbps caused by abusing Memcache [6].

Whenever a new service gets abused for DDoS attacks, a new scanning project presenting the results publicly is usually created. The creators of these projects are organizations and individuals working in networking or cybersecurity fields who are affected by the DDoS attacks but frequently prefer to remain anonymous. The main purpose of such projects is to advise the public in general and network owners that their networks contain systems that can be abused. It can be done by either emailing a notification message to network abuse addresses, notifying only persons who have signed up their network ranges or enabled the conduct of a network-range search in their database. The goal of these projects is to minimize the number of abusable devices as much as and as quickly as possible. Sometimes, these projects cooperate with researchers from academia by providing them with raw data, so that research can concentrate on data analysis instead of technical data gathering. On its own, this research is usually limited to scanning the Internet for all the devices using specific ports and protocols, grouping the results by AS and geographical attributes and presenting them in table and graph formats. If scans are repeated, then comparisons can be made between timespans and device count decline tendency can be identified. If scans are scheduled periodically, then the current situation can be ascertained. Many open ports exposed to the Internet are being scanned by The Shadowserver Foundation, which also includes more than 10 that are most commonly used for amplified reflected DDoS attacks [7]. The Open NTP and Open Resolver projects focus on a single protocol while CyberGreen goes the furthest by calculating and assigning risks.

From the opposite side, scanning activities can be detected and presented in real time and as historical data. One of these projects is NetworkScan Mon, which aggregates data by the source and destination attributes of IP packets and presents aggregated statistics which show that in July of 2018, there was not a single protocol abusable for DDoS attacks among the top 10 ports receiving scanning activities [8]. This indicates that DDoS is a specialized niche of cybercrime and because of the required 2-pronged execution, it is less attractive to cyber criminals as opposed to most popular scanned ports which are used by services that can be directly exploited.

It is possible to monitor DDoS attacks and extract some of the attack attributes by either passively monitoring network traffic at Internet Exchange points or maintaining a distributed set of honeypots that pretend to be exploitable network services. The DDoS Mon project provides insight into worldwide DDoS attack statistics and historical trends; in July of 2018 it reported an average of about 20,000 attacked IP addresses per day [9]. Attacked IPs do not necessarily equate to a single attack or target as systems under attack can have multiple IP addresses. However, no deeper analysis into grouping separate IP addresses into a single target was provided; hence, there is potential for separate research. In the same time period, the USA and China were the most attacked countries, HTTP port 80 and HTTPS port 443 were the most targeted ports, and websites using a .com top-level domain were targeted most often. Amplification and reflection-based attacks were most common, amounting to nearly 70% of the DDoS attacks by frequency, while the most commonly abused protocols were CLDAP, NTP, and DNS. These attack statistics have drawbacks because the number of some specific abused protocol services does not reflect their overall bandwidth contribution to the attack, which is the main property of DDoS attacks.

3. MEASURING DDOS CAPACITY

Most types of DDoS attacks have effective remedies available but volumetric DDoS attacks can exhaust the resources of the whole targeted network, thus affecting all the connected services. Specifically, Reflected Amplified Volumetric DDoS attacks are the most problematic type and the proposed methodology covers only this type of attack. To mitigate these attacks, the defender must absorb and process all the received network traffic by separating legitimate packets from the attack packets. The bandwidth capacity of the attacked network is limited, not only by contractual relations between the ISP and the attacked network but also by the chosen technology and network hardware.

There are two main causes of these types of attacks – the ability to spoof IP addresses, and network services that use the UDP protocol and can produce responses significantly larger than the received requests. Volumetric attacks generate higher bandwidth than attacked networks can process. These attacks are indirect and attacking traffic is produced by unsuspecting devices running abusable network services that generate significantly larger responses than requests. To measure DDoS capacity, these devices need to be identified and their properties extracted and measured to produce the whole picture.

A. Identifying Devices

To estimate the current status of attack capacity, it is sufficient to investigate only publicly known protocols that are being abused. A whole Internet scan should provide the set of abusable devices for the attacks. Depending on the protocol and implementation choices, the scan can be either a generic protocol request or abusable functionality itself. There are differences in the information that can be extracted from this data depending on the approach, e.g., if a scan is conducted using a generic request, then a ratio of abusable to all protocol-implementing devices can be established, which might be useful. If the generic request is not the same as that abused for the attacks, an additional checking request is required before conducting further measurement. At the end of this stage, a set of only abusable devices should be produced.

B. Detecting Rate Limiting

Attackers abusing network services rely on the fact that they do not have any rate limiting. Academic and industry research usually stops at identifying the devices or estimating amplification; there is no published research regarding real-world rate limiting among the identified potentially abusable devices. Technically, rate limiting can either be explicit or implicit. The former is preset in the service's software configuration file or hardcoded in the source code, while the latter is caused by OS, hardware, or network limitations.

Technically, rate limit measurement can be implemented in two ways – sending a burst of packets and verifying the count of received packets or by analyzing every pair of response and request packet sets. Because the measurement requests are the same in most protocols, it should produce exactly the same or a very similar response packet count-wise. By using packet count from the amplification measurement step, it is possible to divide the number of received packets with a packet amplification factor to determine if the resulting value is close enough to the number of sent requests. If it is, then there is no rate limiting or it is above the selected threshold; otherwise, the resulting value approximately corresponds to the rate limit.

More precisely, rate limiting can be measured when mapping sets of response packets to each appropriate request. To some extent, it allows the differentiation of packet loss from rate limiting; as rate limiting is implemented on a per response basis, it might allow identifying exactly from which request responses stopped coming. This method is also suitable for measuring rate limiting that is not on a per second basis by detecting at what request number responses stop and at what number they restart. This type of measurement can technically be implemented in two ways. The easiest way is that every request uses a different source port number, thus every response packet set will be received by a different port. However, in DDoS attacks, all the reflected packets usually target a single port. The harder way is to use the same port and try to differentiate between responses, but depending on the tested protocol, this might be unfeasible because all the sent requests must be the same. Different protocols might possibly produce better data using different measurement methods; from a methodological perspective, it does not matter which approach is implemented as long as advantages and disadvantages are considered for every tested protocol implementation. For the proof of concept, every measurement request expects a response to different incrementing port numbers while enforcing appropriate timeouts to maintain request and response matching.

C. Estimating Network Speed

The network speed of individual devices is one of the main pieces of information lacking in attack capacity estimates. The easiest solution is to use country or specific ISP average upload data gathered by research organizations, but the issue is that abusable devices are a small part of the networks and might not be representative in terms of speed.

An important question is: can the speed of individual devices be estimated from timestamps in the current measuring methodology? If for the start and end time the minimum and maximum values are selected, then a single delayed packet skews the calculation significantly. Speed can be calculated by adding up all the received protocol payload sizes and 42 bytes as transmission overhead for each received packet and dividing it by the time difference between the last and first packet; responses with one packet cannot be processed this way and should be ignored. Although speed calculated in this manner might not necessarily correspond to the speed of the network connection for the device, it could still be a good metric. The device might not be able to fill all the bandwidth capacity available to it using a specific measured protocol. The bandwidth could have been in use in other ways at the time of measurement or the speed could have decreased over a long distance.

D. Technical Concerns

There are significant technical concerns that might affect the quality of measurement and overall viability of the proposed methodology. The measured devices might be participating at the time of measurement in real attacks, thus the measurement would not accurately reflect their capacity. If attackers are measuring themselves and selecting specific most powerful abusable devices, then the total results might get significantly skewed.

Measurement traffic looks exactly like DDoS traffic because the types of requests and responses are the same as those used by attackers. In the real world, receiving or transit networks cannot judge if the request traffic is spoofed or legitimate. Thus, the way to mitigate DDoS to an extent is to block this traffic. We have observed measurement interference from automated solutions deployed across transit networks.

The location of the measurement server both geographically and in the network affects the data. The further away the measured device is, the higher the probability of mitigation solution interference, packet loss, and delays affecting the calculation of its contribution. The same time measurement from a single point produces a view from the perspective of a single specific victim.

E. Estimating Total Attack Capacity

It might be tempting to sum up all the abused protocol measured capacity values together to produce a single value of total worldwide DDoS attack capacity. In reality, there are two major and a wide range of minor factors that limit the attack capacity.

Every network has a limited upload bandwidth capacity that is available for outgoing DDoS attack traffic. A specific network's connection capacity is directly affected by the physical technology in use, router capability, free unused capacity of the uplink and contractual agreement with the ISP or transit provider. The issue is that it is not clear where to draw the border for every network and what the capacity of every network actually is. The easiest solution would be splitting the Internet by AS and using open information from IX monitoring projects and estimating private peering capacity. However, nothing precise is possible because a single AS can contain a large number of separate networks with their own limits that decrease estimate quality as well. Even if reasonable estimates per network basis are established, then the layer of limitation could move up to the transit provider level, as their routers are often not designed to handle maximum load through all the connections at the same time.

Another major factor is that a single device could be providing multiple abusable services simultaneously. In these cases, only the protocol providing higher bandwidth should be counted towards total attack capacity. It might be easy if the protocol measurements for each IP address happen within a small time frame (seconds or minutes), but this is not the case in the designed solution. The greater the time difference between measurements per IP, the less precise it becomes. IP address reachability is affected by dynamic addressing, operating hours, network anomalies and other factors. Properly addressing these factors is crucial for future multi-protocol measurements.

F. Legal and Ethical Considerations

Cybersecurity researchers often cross into gray areas and the legal basis for cybersecurity research is still evolving around the globe. There are three main legal and ethical aspects to consider for this research: scanning the Internet to find abusable devices, measuring discovered devices, and publishing the results.

Scanning the Internet is a common occurrence, it is performed by academic and commercial researchers as well as malicious parties. Although there is no common legal framework that addresses scanning, most non-malicious researchers follow the best practices [10], [11] laid out by the developers of zmap. This allows the minimization of negative impact on the scanned networks and devices but does not negate legal liability.

The significantly higher concern is the measuring stage for every discovered device, as it requires significant interaction with the devices by sending dozens of requests and measuring replies. Scanning for TCP protocol usually involves sending 1 request and a more detailed investigation might involve multiple requests to extract the properties of the device. DDoS capacity measurement relies on the ability to detect rate limiting, thus the number of requests should exceed commonly used rate limits. A large number of requests might interfere with the measured device or cause it to hang, which opens up legal liability. At the same time, devices with abusable protocols are already abused for real-world DDoS attacks, so if they are susceptible to overload, they might be continually affected and should not be serving a critical role.

Published security research always has a risk of being abused by malicious parties. Responsible disclosure minimizes impact, but in these cases of known abused protocols, it is not effective. Furthermore, there is no easy mitigation possible for the DDoS issue. The goal of the research is to present results and encourage positive long-term changes. All three discussed ethical and legal aspects are being evaluated for further research.

4. PRELIMINARY RESULTS

A proof of concept was developed to test the proposed methodology for the NTP protocol. NTP DDoS is known to be significantly mitigated and has a small set of abusable devices to minimize the potential negative impact of the research. The abused feature is a diagnostic command monlist and not a part of the protocol itself. This feature was enabled by default for the NTP server software distribution and produced BAF up to 4670 [2]. After discovery, it was quickly abused by attackers, and at the beginning of 2014 it caused record-breaking DDoS attacks up to 400 Gbps [12]. This command returns up to 440 bytes of payload per packet and up to 100 packets containing recent client data.

A. Scanning and Measuring NTP

Since abused command is not part of the protocol, specific monlist payload must be sent as a request. Different implementations and versions of NTP servers treat this command differently. The vast majority were observed to completely disregard the request without any reply and that is the general way it is expected to detect abusable functionality. However, there are multiple other types of responses stating that the command is not supported and standard time synchronization packets were received as well; these responses are undesired.

Scanning and measuring were conducted in August 2018; as there is no known common rate limit specific to the NTP monlist command, an aggressive 100 measurements per device were used. From the full Internet scan, 943,116 UDP responses to monlist were received, the majority were deemed undesired and only 92,990 devices were actually measured. Almost 63% of the measured devices, a majority of which were located in China, did not respond at all at the measurement stage, potentially indicating network issues, some DDoS protection mechanism or aggressive one-request limits.

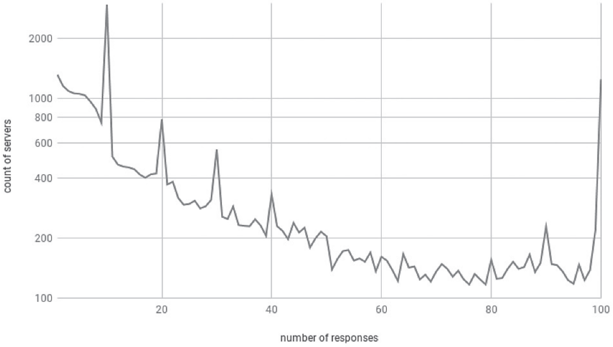
At the measurement stage, 33,325 devices responded with at least one valid monlist response containing an 80-byte payload; these devices are used for attack capacity calculation. Although some of the remaining devices provide some amplification, they are insignificant contributors to the attacks.

B. Rate Limiting

The number of responses for every request was originally expected to identify common patterns of rate limiting but the data produced just demonstrated a downward trend. This might be because it is not known in which order packets are received by the server and only monitoring the server's output on the wire would yield clear patterns. Aggregated data for the number of NTP servers per response count presented in Figure 1 portrays a much clearer picture. 1310 servers responded once (totaling 2 responses as one was received by zmap), to all requests responded only 1242 servers indicating sufficient computing power and network connection quality. However, most noteworthy are the clearly observable spikes at 10, 20, 30, 40, 80 and 90 responses.

There is nothing in the measurement system or network that relies on increments of 10 for sending or receiving packets. This indicates that some kind of rate limiting might be present, possibly set by humans. It is not necessarily explicitly defined in the configuration file of the NTP server software. It might be hardcoded as a limit inside the software or the system itself, especially for low-power embedded systems. This limit might also be present outside the devices, or it is possible that some rate limiting might be enforced by network devices in general or possibly targeting response payload known to be used mostly for DDoS attacks. It is not enforced by measurement network ISP, otherwise the full response spike would not be so significant. It is unlikely that this limit is enforced by a major IP transit provider, or that end-user networks apply these limits manually. Another possibility is that some network security solutions apply these limits automatically. Midsized ISPs are the most likely candidates that would manually create this type of limiting policy.

FIGURE 1: COUNT OF SERVERS PER COUNT OF RESPONSES

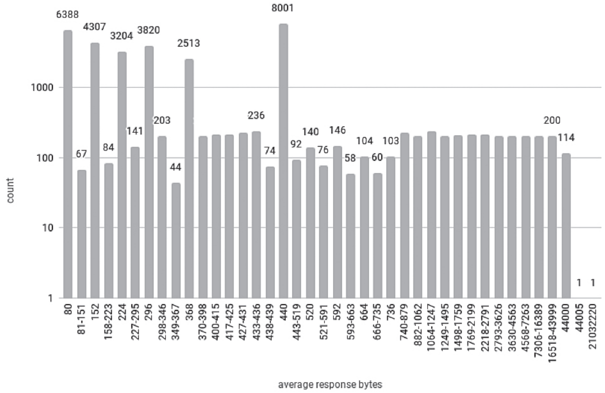


C. Response Size

Actual response size is an important metric as it allows us to calculate real-world BAF. Since the attacker's spoofable bandwidth is a limiting factor of the attack, the attacker would prefer the maximum amplification that abusable services offer. A small response is not necessarily limiting the total contribution to the attack, but it is definitely increasing network resource spending from the attacker. If no implicit or explicit rate limiting is present, then the server can utilize all the upload bandwidth available to it.

The NTP server distribution per average response size is provided in Figure 2. The average is calculated over received responses. If a single response is received, then its size will be the average. The most common values are displayed individually and uncommon values are grouped together; the highest values are the most significant ones. With an 80-byte payload, 6388 servers responded, all of which are monlist replies containing a single client entry. However, the most common response size is 440 bytes in 8001 cases, which corresponds to a single full packet monlist response. It is either an implementation issue or a mitigation effort fix for the configuration or the software itself to minimize the impact of the abuse. Only 114 servers provided maximum possible responses of 100 packets with a 440-byte payload totaling 44,000 byte responses without packet loss. Diagnostic information about a single client uses 72 bytes of response, 5 clients produce single full response packet and if there are more clients, additional response packets are generated in the same way.

FIGURE 2: NTP AVERAGE RESPONSE SIZE DISTRIBUTION



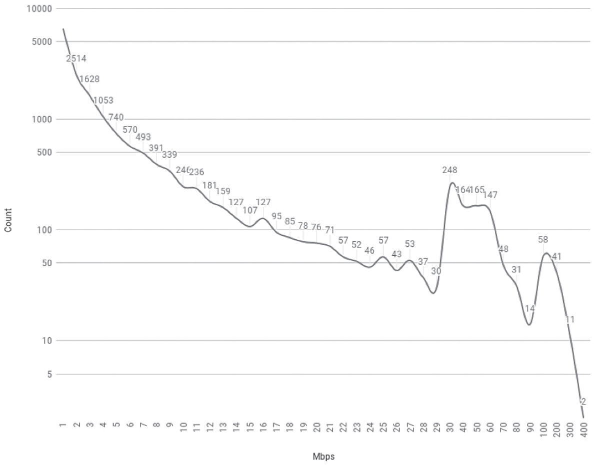
D. Response Speed

All the servers responding with an average speed above 0.5 Mbps can be considered significant contributors and are presented in Figure 3. In total 17,208 or 54% of the servers responded with speed above 0.5 Mbps, with the peak being at 0.5 – 1.5 Mbps and then decreasing; 73% of these servers had measured speeds below 5.5 Mbps.

A large portion of responses (569 servers) were received at speeds below 5 Kbps. These servers either have a low response rate and respond slowly or have a high response rate and take multiple seconds to respond from the first to last packet. Random sampling indicates that a significant portion of these devices have slow-speed wireless connections to the Internet. These servers are disregarded from further calculations. There is a spike in the number of servers responding with speed 5 – 15 Kbps; in total, 6896 servers responded with speeds between 5 and 100 Kbps. In total, 14,750 servers responded with speed below 0.5 Mbps. These are insignificant contributors to the overall DDoS attack capacity that could potentially be disregarded from the analysis set.

There are noticeable outliers with average speeds above 100 Mbps, most of which were identified as data centers and hosting providers providing virtual servers and dedicated servers for rent and supplying them with high-speed Internet connections with speeds of 100 Mbps – 1 Gbps or above. A top provider, OVH, with 15 reflectors is known for low prices and abuses. These devices are high contributors to the attack capacity.

FIGURE 3: NTP SERVERS RESPONDING WITH AVERAGE SPEEDS ABOVE 0.5 MBPS



E. Contributors to the Attack

Most NTP servers were located in the USA (8061), China (4689), Brazil (3320), Spain (2420), Turkey (1832), Indonesia (1432), Taiwan (1227), Vietnam (1226), Saudi

Arabia (1134) and Malaysia (1032). The USA is disproportionately represented in many scans, which might be surprising, but it is related to the historical availability of the Internet and a high number of legacy systems. The whole continent of Africa has very few amplifiers, about half of the countries have none. With the speed and cost of the Internet in Africa, it is expected that contribution to the total attack capacity is insignificant. Asia is a high contributor and many other network issues are caused by fast proliferation and growth of the Internet in these developing countries. A large connection count and fast speed, coupled with a lack of regulation and enforcement and general disregard for the best network management practices, all cause Asian countries to be breeding grounds for cybersecurity issues. However, as noted by existing research [5], pure count is not a good metric for estimating contribution to total attack capacity; the count needs to be balanced against upload bandwidth.

Bandwidth contribution is a significantly more important metric than overall server count. Compared to the count, significant differences can be observed – the USA and Spain contributed much more count-wise than capacity-wise. This might confirm that high-count ISPs might have low-power or low-speed embedded devices running the services without contributing significantly to the total attack capacity. China is the top contributor with about 42 Gbps total attack capacity, followed by the USA with 16 Gbps. Next are Russia and France which have low NTP server counts but very high network speeds. The rest of the top 10 are Asian countries and Brazil. Overall top contributors to the attack capacity are developed countries and developing countries with high Internet connectivity speeds.

F. NTP Attack Capacity

Summing all the calculated average speed together for the reflectors that provided more than one reply with calculated speed of at least 5 Kbps, the total speed of the NTP monlist DDoS attack was **134 Gbps**, generated by 31,389 servers. This value does not necessarily correspond to the real-world situation. There might have been competition for bandwidth with ongoing DDoS attacks. Real-world capacity could be significantly larger. Geographic distance decreases average speed as well, intermittent or permanent network quality issues would decrease measured speeds but not actual bandwidth.

There are no current estimates of NTP monlist attack capacity and no published recent attack case studies because the attack has long lost its peak capacity. It would allow the extraction of some empirical constant that potentially could be used as a multiplier for the measured capacity to produce a realistic estimate.

The real measured BAF for the 134 Gbps capacity can be calculated by dividing the total received bytes with the 100 payloads sent multiplied with payload length and server count. In this case, the real total measured **BAF** was **20.55**, which is significantly below the standard maximum of 2750. If an attacker disregards servers with large packet losses and small responses, then he can achieve attacks with multiple times larger BAFs. Whether or not attackers conduct such measurements is an open question for further research.

5. CONCLUSIONS

The proposed methodology is promising and covers aspects missing in existing ones. The implemented proof of concept produced an NTP DDoS capacity of 134 Gbps and is suitable for adaptation to different protocols. Significant technical, ethical and legal concerns were identified that require further investigation to determine if the research methodology is viable.

REFERENCES

[1] "History of DDoS Attacks," *Radware*, 13-Mar-2017. [Online]. Available: https://security.radware.com/ddos-knowledge-center/ddos-chronicles/ddos-attacks-history/. [Accessed: 04-May-2017].

[2] C. Rossow, "Amplification Hell: Revisiting Network Protocols for DDoS Abuse," in *Proceedings of the 2014 Network and Distributed System Security Symposium*, San Diego, CA, USA, 2014.

[3] J. Czyz, M. Kallitsis, M. Gharaibeh, C. Papadopoulos, M. Bailey, and M. Karir, "Taming the 800 pound gorilla: The rise and decline of NTP DDoS attacks," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014, pp. 435–448.

[4] Center for Applied Internet Data Analysis based at the University of California's San Diego Supercomputer Center, "State of IP Spoofing." [Online]. Available: https://spoofer.caida.org/summary.php. [Accessed: 23-Jul-2018].

[5] E. Leverett and A. Kaplan, "Towards estimating the untapped potential: a global malicious DDoS mean capacity estimate," *Journal of Cyber Policy*, vol. 2, no. 2, pp. 195–208, May 2017.

[6] C. Morales, "NETSCOUT Arbor Confirms 1.7 Tbps DDoS Attack; The Terabit Attack Era Is Upon Us," 05-Mar-2018. [Online]. Available: https://asert.arbornetworks.com/netscout-arbor-confirms-1-7-tbps-ddos-attack-terabit-attack-era-upon-us/. [Accessed: 09-Mar-2018].

[7] The Shadowserver Foundation, "The scannings will continue until the Internet improves." [Online]. Available: http://blog.shadowserver.org/2014/03/28/the-scannings-will-continue-until-the-internet-improves/. [Accessed: 30-Jun-2018].

[8] Qihoo 360 Technology Co., Ltd, "Scan volume per 10 minutes," *NetworkScan Mon*. [Online]. Available: http://scan.netlab.360.com/. [Accessed: 19-Jul-2018].

[9] Qihoo 360 Technology Co.,Ltd, "Insight into Global DDoS Threat Landscape," *DDoS Mon*. [Online]. Available: https://ddosmon.net/insight/. [Accessed: 20-Jul-2018].

[10] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-wide Scanning and Its Security Applications," in *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, Washington, D.C., 2013, pp. 605–620.

[11] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A Search Engine Backed by Internet-Wide Scanning," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver, CO, USA, 2015, pp. 542–553.

[12] M. Prince, "Technical Details Behind a 400Gbps NTP Amplification DDoS Attack," 13-Feb-2014. [Online]. Available: https://blog.cloudflare.com/technical-details-behind-a-400gbps-ntp-amplification-ddos-attack/. [Accessed: 25-Jan-2018].

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Silent Battles: Towards Unmasking Hidden Cyber Attack

Robert Koch
Fraunhofer FKIE
Bonn, Germany
Robert.Koch@fkie.fraunhofer.de

Mario Golling
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
Mario.Golling@Unibw.de

Abstract: When looking at the media, it can easily be seen that new cyber attacks are reported on a regular basis. The corresponding effects of these attacks can be manifold, ranging from downtime of popular services affected by a rather trivial Denial-of-Service attack, to physical destruction based on sophisticated cyber weapons. This can also range from single affected systems up to an entire nation (e.g., when the cyber incident has major influence on a democratic election). Some of these attacks have gained broader public attention only by chance. This raises the fundamental question: do some cyber activities remain hidden, even though they have a significant impact on our everyday lives, and how can such unknown cyber involvements be unmasked? The authors investigate this question in depth in this paper.

The first part of the paper analyzes the characteristics of silent battles and hidden cyber attacks – what needs to be considered on the way towards a better detection of hidden cyber attacks? After that, an evaluation of the current cyber security landscape is provided, summarizing what developments we can see and what we can expect. Based on this, the complexity of detecting hidden cyber attacks is discussed and we ask the question: why does detection fail and how can it be improved?

To investigate this question, the capabilities of the attackers are examined and are reflected in a 3-Layer Vulnerability Model. It is shown that a traditional Cyber Kill Chain is not sufficient to detect complex cyber attacks. Therefore, to improve the detection of hidden cyber attacks, a new detection model based on combining the 3-Layer Vulnerability Model and the Cyber Kill Chain is proposed.

Keywords: *cyber war, cyber defense improvement, attack model, detection requirements, detection transitions, Cyber Kill Chain, 3-dimensional detection model*

1. INTRODUCTION

Whether we consciously perceive it or not, whether we want to admit it or not, our everyday life is entangled with information technology (IT). Today, IT is a corner stone for our daily office work and is even a prerequisite for administrative tasks at public authorities. It covers areas from transportation and telecommunication up to industrial control systems and the financial sector. In short, today's world is more cyber-dependent than ever. However, due to (i) its economic potential, but also because of (ii) the alleged and at least partially achievable anonymity, (iii) regularly occurring security vulnerabilities, and (iv) the lengthy international prosecution of cyber crimes, the Internet offers a considerable potential for abuse. To counter this abuse, various protection systems and programs have been published and established over the past decades. In parallel, the fundamentals of the Internet, such as standards and protocols, have also been improved or developed from scratch to reduce the risks involved with a broad usage of the Internet. Despite these efforts, the economic losses remain very high. In this regard, corresponding estimates are often problematic, due to (un)available data, the expected number of unreported cases and the complexity of indirect costs. A recent estimate by RAND, which was published as part of their "Cyber Risk Calculator", states that "the global cost of cyber crime has direct gross domestic product (GDP) costs of $275 billion to $6.6 trillion and total GDP costs (direct plus systemic) of $799 billion to $22.5 trillion (1.1 to 32.4 percent of GDP)" [1]. The authors emphasize the high sensitivity of the numbers regarding the input parameters. Nevertheless, even the "most favorable" case reveals the enormous loss that results from cyber incidents. According to McAfee [2], for instance, the global cost of cyber crime has now reached as much as $600 billion - about 0.8 percent of the global GDP. As already mentioned, such estimates are extremely difficult to perform, usually due to a lack of sufficient data. For many reasons, such as the fear of reputation loss, companies are often cautious whether to report cyber attacks or not. Even today, reporting obligations are limited to a few areas such as critical infrastructure. This also hampers the detection of cyber attacks. In various reports, cyber security companies have regularly warned that insufficient detection procedures can be expected in areas that report little or even no cyber attacks.

In practice, this fundamental problem becomes even more difficult once different groups and abilities of attackers, and the associated challenges of detection, are taken into account. In recent years, extensive measures for the preparation of the battlefield can be observed. In particular, the Snowden Leaks [3] and the Vault 7 [4] and 8 [5] files have revealed details about comprehensive programs for the manipulation of hardware, software and standards. Some of those attacks have gained broader public attention only by chance. This raises the fundamental question of whether some cyber activities may remain hidden even though they have a significant impact on our everyday lives, and how as yet unknown cyber involvements can be unmasked. In numerous cases, cyber attacks remained unrecognized for a long time, often to the surprise of the victim.

A. On Silent Battles and Their Relevance - A Brief Review

The element of surprise, that is the ability to conduct an attack without warning, is one of the central and therefore most discussed aspects of military theoreticians in general. According to the Prussian general and military theorist Carl von Clausewitz (one of the most well-known analysts of normative behavior and trends in military affairs and military history), the concept of war, therefore the act of violence to the opponent in order to the fulfill one's own will [6] requires first and foremost the pursuit of relative superiority [7]. For this in turn, the surprise of the enemy [8] is more or less always of utmost importance. Without surprise, superiority (the crucial point) is actually unthinkable [8]. Thus, surprise is the precondition for superiority, which in turn is the greatest precondition of victory [8]. Where both (surprise and superiority) succeeds, confusion, and broken courage of the opponent are the consequences [8]. These considerations are similar, for example, with those of Sun Tzu, another well-known military theorist. For Sun Tzu, in the fight, direct actions lead to confrontation, surprising actions lead to victory [9]. Although there are roughly 2,500 years between Sun Tsu and today, little has changed. Silence is the prerequisite for surprise, which in turn is a prerequisite for superiority. Superiority, after all, is the most general principle of victory.

B. Silent Battles - A Definition of Terms

The term Silent Battle therefore has several facets. Based on the aforementioned considerations, we define the term Silent Battle with regard to cyber activity as

- a hostile encounter or engagement between opposing parties (nations, organizations, military forces)
- characterized by an absence or near absence of "noise" or "sound".

This may in particular be due to the fact that the engagement (i) remains hidden (i.e., no actions have been discovered, no effects can be observed) or (ii) allows no attribution (thus no legally consistent link to the opposing party can be established), or (iii) is not relevant for the public (e.g., no media coverage) or (iv) is not disclosed (e.g., because a company prefers to keep an attack secret so as not to upset its customers).

C. Silent Battles in the Cyber Domain

At this point, in the context of hidden cyber attacks, one also has to raise the question, whether "noise" can also be used to distract attention. Concerning this matter, a look at the cyber incidents of recent years reveals that Distributed Denial-of-Service (DDoS) attacks seem to be used to mask the actual attack in order to divert the IT security department. For example, a comprehensive study of DDoS attacks published by Kaspersky in 2015 came to the conclusion, that "74% of attacks that lead to a noticeable disruption of service coincided with a different type of security incident, such as a malware attack, network intrusion or other type of attack" [10]. On the other hand, other companies disagree and present different results of the analysis of data available to them. With respect to DDoS attacks covering other breaches, Verizon made a humorous comparison to the government covering up evidence of alien visitation: it is often heard but not so easy to prove [11]. Based on their evaluation, "this year's data set only had one breach that involved a DoS, and in that one, the breach was a compromised asset used to help launch a DDoS, not the other way around" [11]. These essentially different results with respect to the same elementary attack vector, namely DoS, show the challenge of analyzing the cyber security environment. Therefore, Silent Battles in the cyber domain *may* be accompanied by noise like DDoS, but of course they do not have to be. Accordingly, aspects like this must also be taken into account when identifying opportunities for hidden cyber attacks, and other attack vectors must be considered as differentiated.

D. Structure of the Paper

To investigate the question whether some cyber activities may remain hidden even though they have a significant impact on our everyday lives or how yet unknown cyber involvements may be unmasked, the paper is structured as follows: an analysis of the evolution of the cyber security landscape is presented in Section 2, highlighting different aspects of what we know, and what we can expect. In order to develop new ways of detecting hidden attacks, Section 3 applies a 3-Layer Vulnerability Model to investigate potential attackers, their capabilities, and their characteristics. To clarify the particularities, some examples are discussed and a possible usage of the characteristics in order to identify hidden attacks is presented. Based on this theoretical foundation, Section 4 proposes a first model for the identification of hidden attacks. For this purpose, some Lemmata regarding observable respective useful properties are motivated and introduced before a three-dimensional extension of the current Cyber Kill Chain is proposed in order to improve the identification of hidden cyber attacks. Finally, Section 5 summarizes the key aspects of the paper and presents our next steps, including the evaluation of a prototypical implementation of our model by using suitable datasets.

2. EVOLUTION OF THE CYBERSECURITY LANDSCAPE

In order to identify further characteristics such as the importance of noise in the context of cyber attacks, the next step is to look at what we currently know about cyber attacks and the developments in this area of the last few years.

A. What We See

As there is nowadays a variety of reports on an annual, half-yearly or quarterly basis available, as well as occasion-related publications, only a few key findings and observations of selected recent reports, which are most important for the paper, are summarized here.

As companies like Verizon, Symantec, IBM or Kaspersky have huge "sensor networks" available like, for example, the evaluation of data generated by nodes which are equipped with endpoint protection, a good picture of different developments and incidents can be generated. However, it must not be forgotten that due to the complexity of the cyberspace and its systems, each and every system can only provide its viewpoint, which depends on many factors. Results of different systems can support each other, but also quite different results can be achieved, as in the aforementioned example of the DoS attacks.

Of course, the basic risk posed by cyberspace today is not only addressed in the reports of IT companies alone. Having a look at The Global Risks Report 2019 published by the World Economic Forum, data fraud or theft and cyber attacks are placed 4th and 5th on the Top 10 risks in terms of likelihood, with rising cyber dependency as one of the main risk-trends in 2019 [12]. Various reports underline the fact that a cyber incident is coming, and it is therefore essential for the companies to prepare themselves accordingly (e.g., see [11, 13]). As Verizon highlights, state institutions are in a particularly bad situation: "Depending on function, government entities may be targeted by state-affiliated groups, organized crime or employees" [11].

One phenomenon that has been observable for a long time is highlighted: the discrepancy between the perception of the threat of cyber attacks and the lack of *strategically* addressing the threat. While many companies are aware of the danger, it is rarely considered a strategic priority [14]; this also holds for the industry and the area of operational technology (OT), where for example "only 23% are compliant with minimal mandatory industry or government guidance and regulations" [15]. The risk increases all the more because of the increasing convergence of IT and OT and the growing use of Industrial Internet of Things (IIoT) devices [13].

This dangerous discrepancy can also be explained by the fact that cyber attacks are still a “mystery” for companies. For example, Accenture analyzed that for 71% of their respondents, cyber attacks are still a “bit of a black box; we do not quite know how or when they will affect our organization” [13]. On the other hand, if one looks at the professionalism of current attacks, this situation is particularly worrying. For example, the malware Triton (also called Trisis or Hatman [16]) was specifically targeting Safety Instrumented Systems (SIS), systems which enable the controlled shutdown of industrial processes when unsafe operating conditions are detected. While this malware was found in at least one critical infrastructure facility [13], the necessary knowledge and capacity to build such malicious programs is available to an increasing number of players [17, 18].

In this context, not only direct attacks are an increasing challenge, but especially attacks executed by exploiting the networks of third- or fourth-party supply chain partners. Here, a broad range of attack techniques is already in use. Accenture emphasizes, that they “have collected intelligence on recent campaigns that highlight the challenges of combating weaponized software updates, prepackaged devices, and supplier ecosystems as these all fall outside the control of victim organizations” [13]. Recent vulnerabilities with global impact like Meltdown and Spectre exacerbate the situation (see [19]), as periods of widespread vulnerability disclosure provide opportune times for actors to distribute malicious communications to users anticipating updates [13]. As a result of that, even the traditionally good advice to keep the patch level of the systems as up-to-date as possible is reaching its limits and requires practical precautions. For a better understanding of the threat, an evaluation of the attack path taken can be useful. By identifying the different steps and analyzing their characteristics, new detection opportunities may be discovered which later can be used to detect and possibly mitigate future cyber attacks [11].

From a more technical point of view, after a slight decline in 2017, significantly more malicious software was identified again in 2018 [20]. Due to the “success” for cyber criminals in 2017, an increasing number of ransomware campaigns could be observed in 2018 [20]. While these numbers are not surprising, the increasing proportion of encrypted cyberattacks is more interesting; the proportion of encrypted traffic in the Internet has been increasing for years. This follows the publication of the Snowden documents, and follows efforts from Google and projects such as Let’s Encrypt [21] (now reaching a level of almost 70 percent [20], and for Google services even more than 90 percent [22]). Cyber criminals are also increasingly using encryption to disguise malicious traffic [20], another example of the dual use challenge [23].

As a result, larger companies are increasingly using Secure Sockets Layer (SSL)/Transport Layer Security (TLS)-Scanning technologies. This in turn weakens the security of the encrypted link and introduces new attack vectors (see [24]). The fact that security systems can be the attack vector itself is shown by numerous examples from recent years. For example, Tavis Ormandy has repeatedly demonstrated how antivirus software could be exploited for attacks due to programming errors (see [25]), while the Snowden documents have revealed numerous examples of the deliberate weakening and incorporation of backdoors in firewalls [26].

B. What We Know

Thanks to some whistleblowers like Snowden, some light was shed into the shadow of the real cyber security situation, which goes well beyond what one can read from logs of systems and reports of companies and authorities. Of course, the use of such sources always requires a reality check and a certain dose of skepticism, because it is also conceivable that deliberately generated leaks may well have the goal of disseminating false information. From a scientific point of view, there was nothing really unexpected within the disclosures of Snowden. However, the whole dimension of surveillance, and thus the severe infiltration of security systems, hardware, firmware, software and even algorithms was somehow surprising and disturbing. The documents contained, for example: information about programs for firmware persistence implants with backdoor capabilities like JETPLOW [27]; BIOS persistence implants like SOUFFLETROUGH [27] for the installation in firewalls; or hardware implants like GODSURGE [28] which exploits the Joint Test Action Group (JTAG[1]) debugging interface of the server's processor.

Other interesting information was the disclosure known as the Vault 7 breach[2], containing information on the capabilities and hacking activities of the CIA [29]. Important details disclosed related to programs like MARBLE [30] which aim to obfuscate the source of a program or even try to motivate a false attribution, WEEPING ANGEL [31], a tool to exploit Smart TVs for the purpose of intelligence gathering, or programs which aim at the steering system of cars.

Another important area is the comprehensive analysis of incidents that initially did not necessarily have to be caused by cyber means. For example, having a look at the Ukraine's power outages in December 2015 and 2016, the suspicion of a cyber attack emerged quickly after the incidents, but only extensive investigations revealed the exact occurrence and the complex attack path [32]. Taking the many pieces of the jigsaw puzzle that results from the disclosures, leaks, and recent research suggests the approximate extent of the threats in cyberspace.

[1] IEEE 1149.1

[2] Disclosed by Joshua Adam Schulte.

C. What We Expect

When considering what must be expected, and what may be already applied in the real-world, news reports and stories, scientific work and the associated discussions must be taken into account, and evaluated holistically. A prominent example are hardware backdoors built into products like processors or server boards. While a lot is written about the possible endangerment, and research papers pertaining to reversing the x86 processor microcode and prototypically implementing microcoded Trojans into the AMD K8 & K10 processors [33] are available, actual real-world cases are rare. An interesting but controversial example was the discussion on a hardware backdoor in the Microsemi ProASIC3 processor. While the researchers found some processor commands onboard the chip which could be used as a backdoor [34], 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 can be true, but on the other hand, for a sensitive or classified application, it is a dangerous attack vector, regardless of what you call it.

In October 2018, there was a new and much more public discussion based on an article published by Bloomberg Businessweek called "The Big Hack: How China Used a Tiny Chip to Infiltrate U.S. Companies" [35]. Bloomberg claimed that China implemented tiny Trojan hardware into Supermicro servers at manufacturing time, and that government contractors and companies like Apple had been affected. Even after the immediate and vehement contradiction of the alluded companies and institutions, Bloomberg stood by their statement [36]. An analysis of the rare available technical details, completed by the Security Research Computer Laboratory of the University of Cambridge, concluded that an attack in the described manner is technically feasible [37].

Regardless of whether the case described by Bloomberg has taken place in this way, the threat of corresponding attacks is obvious, as they can be attractive for state actors because of their relatively simple feasibility, the complex and low detection options and, if properly carried out, the good opportunities for plausible deniability.

The identified attacks against the supply chain executed by nation-state threat groups like the Chinese cyber espionage group PIGFISH or the Russian BLACK GHOST KNIFEFISH group [13] and the introduction of malicious software and backdoors into industrial control systems and critical infrastructure, underlines the severe, real threat and demonstrates the further preparation of the battlefield [17].

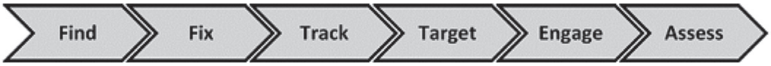
3. DETECTING HIDDEN CYBER THREATS

Due to the complexity and rapid development of cyberspace, an attacker has numerous attack vectors from various areas available that are difficult to detect.

A. Why Detection Fails

The use of singular detection techniques like antivirus or intrusion detection systems (IDS) is not enough for adequately detecting cyber threats nowadays. Even though heuristics and detection methods such as behavior-based detection are constantly being improved, attackers are able to avoid the protective mechanisms on a large scale again and again. Detection becomes particularly difficult if tools and malicious code are specifically developed or adapted within the scope of targeted attacks.
To identify attack vectors, weak links and also detection opportunities, a kill chain can be used for the analysis. A kill chain is a phased-based model to describe the stages of an attack (see [38]). By analyzing them, weaknesses can be identified, and subsequently can be hardened. The basic steps of a common kill chain are shown in Figure 1.

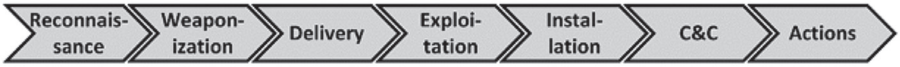
FIGURE 1. COMPONENTS OF THE KILL CHAIN



For a better application to cyber threats, Lockheed Martin proposed the so-called Cyber Kill Chain for identification and prevention of cyber intrusions activity by identifying what the adversaries must complete in order to achieve their objective [39].

The proposed Cyber Kill Chain contains seven steps, namely (1) Reconnaissance: harvesting email addresses, conference information, etc., (2) Weaponization: coupling an exploit with a backdoor into a deliverable payload, (3) Delivery of the weaponized bundle to the victim via email, web, etc., (4) Exploitation of the vulnerability to execute a code on the victim's system, (5) Installation of malware on the asset, (6) Command & Control channel installation for remote manipulation of the victim and finally (7) Actions on Objectives to accomplish the intruder's original goals [39]. Figure 2 summarizes the attack steps.

FIGURE 2. CYBER KILL CHAIN BY LOCKHEED MARTIN [39]



While this model and the different steps are stringent and well understandable, the current Cyber Kill Chain still does not seem to be sufficient for the detection of sophisticated cyber attacks. As mentioned above, the attack path and characteristics of an attack are often still not really known to the companies respectively defenders [13]. While a basic model of cyber attacks like the Cyber Kill Chain can be helpful there, such a simple, one-dimensional model is often not able to describe and eventually identify especially sophisticated cyber attacks for two basic reasons: the companies can overlook the respective indicators or they may not even be able to look for them; and/or cyber campaigns may inflict several different targets and specific attack steps may only be executed against selected ones. On the other hand, the composition, characteristics and transitions of the attack steps may not be exact enough or may even be faulty, depending on the adversary and their available attack techniques. For example, if an adversary is able to introduce the vulnerability they want to exploit by using a supply chain attack, at least steps 1 to 3 of the Cyber Kill Chain, depending on the implementation and the used trigger maybe even up to step 6, must *not* be executed.

Therefore, an extension which better reflects the attackers capabilities and the cyber security- respectively vulnerability-ecosystem is required to improve detection chances.

B. Vulnerability Model

A basically 3-Layered Model can be used to describe the different kinds of vulnerabilities and their specific characteristics. In their publication "Task Force Report: Resilient Military Systems and the Advanced Cyber Threat," Gosler et al. proposed a 6-Tier Cyber Threat Taxonomy to describe the capabilities of potential attackers [40]. The fundamental distinction of the attackers is based on the level of skills and breadth of available resources, building the different Tiers as follows (see [40]):

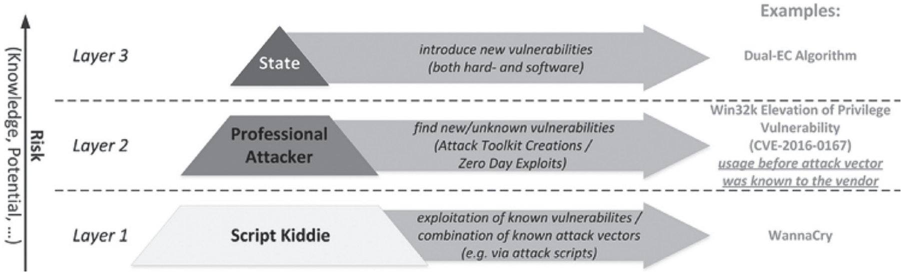
- Tiers I and II attackers primarily exploit known vulnerabilities;
- Tiers III and IV attackers are better funded and have a level of expertise and sophistication sufficient to discover new vulnerabilities in systems and to exploit them;
- Tiers V and VI attackers can invest large amounts of money (billions) and time (years) to actually create vulnerabilities in systems, including systems that are otherwise strongly protected.

The original model of the Task Force used the six Tiers presented, and grouped them into three layers. For the further considerations, the use of the respective three layers is sufficient:

- Layer 1 for the exploitation of known vulnerabilities;
- Layer 2 for finding new, yet (publicly) unknown vulnerabilities; and
- Layer 3 for deliberately introduced vulnerabilities.

The model is visualized in Figure 3.

FIGURE 3. 3-LAYER VULNERABILITY MODEL BASED ON [40]


Examples:
Risk
(Knowledge, Potential, ...)
Layer 3
State
introduce new vulnerabilities
(both hard- and software)
Dual-EC Algorithm
Layer 2
Professional Attacker
find new/unknown vulnerabilities
(Attack Toolkit Creations /
Zero Day Exploits)
Win32k Elevation of Privilege
Vulnerability
(CVE-2016-0167)
usage before attack vector
was known to the vendor
Layer 1
Script Kiddie
exploitation of known vulnerabilites /
combination of known attack vectors
(e.g. via attack scripts)
WannaCry

To gain a better understanding of the peculiarities of the different levels, some examples of corresponding vulnerabilities are described:

Layer 1 Vulnerabilities are the exploitation of publicly known shortcomings, which are already published, for example, by the Common Vulnerabilities and Exposures (CVE) database provided by the MITRE Corporation, including an identification number, a description, and at least one public reference [41].

Due to public knowledge of the vulnerabilities, a good defense against them *should* be possible. Typically, details of vulnerabilities will not be announced until several weeks after discovery so that developers of the affected product have time to generate and publish a patch. In practice, however, such known vulnerabilities can be exploited quite often. This can be due to a variety of reasons, for example, poor system maintenance if available patches are not installed in time. On the other hand, it may also happen that for detected and published vulnerabilities no more patches are provided., because the product is no longer supported by the responsible company ("end-of-life", EOL) or the company possibly no longer exists. In the area of operational technology (OT) such as industrial control systems (ICS) but also with devices of the so-called Internet of Things (IoT), it still happens again and again that discovered vulnerabilities cannot be closed because of insufficient system resources or other limiting factors. Certifications can also cause significant delays in deploying and installing patches, for example, in the medical area or in avionics, where any changes to the system, including patching, may require re-certification [42].

For these reasons, even the Layer 1 vulnerabilities can create significant trouble in everyday life. A prominent example is the ransomware WannaCry, which hit enterprises and institutions all over the world in May 2017 [43]. Its impacts included the taking offline of 61 National Health Service hospitals in the UK, production stops at car factories in France and Japan, and several further significant disruptions. While there was already a patch available for the exploited ETERNALBLUE[3] vulnerability [44], the ransomware particularly affected the older Windows XP/Server 2003 systems for which no patch had been published until the consequences of the worm run. Anyway, it was "just" the exploitation of a known vulnerability, but based on the aforementioned reasons, with very bad effects. At least, even if no patch is available, or in a case where it cannot be applied, the knowledge of a vulnerability can be used to prevent an exploitation by other means, for example mitigating the risk of an unpatched vulnerability by preparing respective firewall or IDS resp. intrusion prevention system (IPS) rules, etc.

Layer 2 Vulnerabilities are new, yet publicly unknown vulnerabilities which are found by techniques like code analysis, reverse engineering or fuzzing. Depending on the kind of vulnerability, it can have a quite different value, ranging from a few dollars up to 2 million dollars. A vulnerability of a less common system, or one which only generates a DoS-condition, is of course not so valuable like, for example, a remote code execution for Apple's macOS. While most companies have bug bounty programs nowadays where researchers are rewarded when submitting new identified vulnerabilities, it can be much more lucrative to sell them to companies like ZERODIUM which are working in a gray area, buying 0day vulnerabilities from researchers and selling them, to, e.g., governments. Based on the possible destructiveness of 0days, there is a debate in numerous countries whether or not governments should retain or disclose such vulnerabilities. Owning a corresponding arsenal is the prerequisite for being able to conduct cyber attacks reliably and at any time. Accordingly, they are of great importance for governments, but also in the context of organized crime and other areas, which promotes the corresponding market and trade. In this context, the RAND Corporation published the analysis of a data set of information about 0day vulnerabilities and exploits regarding the life status, longevity, and collision rates [45].

At this point, the increasingly important role of so-called 1days should be mentioned. 1days are vulnerabilities that have *just* been published. While in the optimal case, the corresponding patches are already provided and possibly even installed by automatic mechanisms, the reasons given above always result in a window of opportunity, where the corresponding patches have not yet been installed in a number of systems and therefore still can be exploited. For capable attackers, these are low hanging fruits; e.g., CrowdStrike published an evaluation that Russian hackers require only about *18*

[3] The vulnerability was stolen from the NSA by the Shadow Brokers in 2016 and published by them in April 2017.

minutes to infiltrate a computer network [46]. Furthermore, new companies and offers are emerging in this area, with very fast development and provision of 1day exploits right after the release of the vulnerabilities.

Layer 3 Vulnerabilities are the culmination of the opportunities available to attackers as they offer assured access combined with a very low detection risk. This is achieved by intentionally introducing vulnerabilities into products, often without the provider of the product learning about them. The most dangerous manipulations at this level are involving algorithms and even standards. An example in this context is the Dual-EC algorithm, which was provided by the NSA with a kleptographic backdoor. This example also highlights the small number of players who are able to perform this kind of high-level attacks. In addition to the required mathematical knowledge [47], information is also needed on the corresponding influence, in this case this was on a standardization body. Other Layer 3 attacks can involve the manipulation of hardware, for example, by adding backdoors to chips or adding malicious components to a system like that highlighted in the discussion earlier in this paper on the article from Bloomberg Businessweek [35]. Recent attacks on the supply chain, which are increasing rapidly, underline the corresponding risk and may open up opportunities for sophisticated cyber attacks [48]. The complexity of today's supply chains makes it easier to attack them. At the same time, proving a manipulation can be difficult, even after a detection, as the discussion on the Microsemi ProASIC3 has shown. Thus, at least in certain cases, by appropriate reasoning, even in the case of discovery, a malicious intent may be denied, which may be another incentive to perform such manipulation.

Based on this 3-Layer Vulnerability Model, the respective attackers and their capabilities can also be described, and opportunities of detection and defense can be discussed.

C. Detection Opportunities

Taking the characteristics of the vulnerability model into account and combining them with the Cyber Kill Chain approach, new detection opportunities arise which may be useful to build new and more powerful and effective detection systems.

The fundamental detection challenge of sophisticated, and therefore quite often hidden cyber attacks, is as follows: due to the large and ever increasing amounts of data, as well as the complexity of the systems and the speed in cyberspace, a high degree of automation of the evaluation is required. On the other hand, attack vectors which are to be expected for sophisticated attacks, are often not recognizable by today's systems, which merely evaluate the data traffic by using different techniques. This includes, for example, signatures and heuristics for data classification, the evaluation of the process flow or user behavior to identify malign programs, threats and activities. While this may be sufficient for identifying and preventing Layer 1 attacks, already protection against Layer 2 attacks is only possible to a limited extent when using these techniques, and regularly ineligible for Layer 3 attacks. One of the main problems is that important elements of the attack path of sophisticated attacks cannot be identified "by cable," for example, social engineering attacks (in the sense of the original social engineering with direct interaction [49], not in the sense of indirect vectors like spear phishing emails where there is no direct social interaction between the involved parties). Regardless, even in the case of Layer 3 attacks, interaction with systems and networks is required sooner or later, otherwise, it would not be a cyber attack.

The related actions of sophisticated attacks often stay under the radar of current detection technology, as they are specifically adapted or even designed for the respective target. However, adding knowledge about the attacker and their capabilities, as well as adding additional sources for the evaluation, means that different measures can be taken to retrospectively identify evidence of a sophisticated, yet hidden cyber attack.

FIGURE 4. ENRICHMENT OF THE DATA TO BE EXAMINED DEPENDS ON THE LEVEL OF THE ATTACKER.

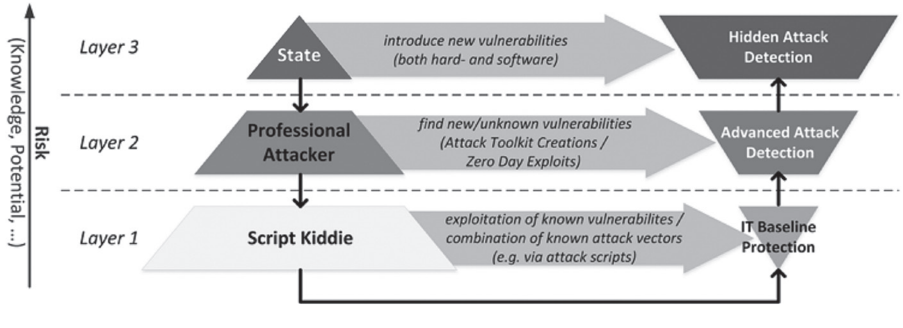

Risk
(Knowledge, Potential, ...)
Layer 3
Layer 2
Layer 1
State
Professional Attacker
Script Kiddie
introduce new vulnerabilities
(both hard- and software)
find new/unknown vulnerabilities
(Attack Toolkit Creations /
Zero Day Exploits)
exploitation of known vulnerabilites /
combination of known attack vectors
(e.g. via attack scripts)
Hidden Attack Detection
Advanced Attack Detection
IT Baseline Protection

Figure 4 highlights the basic idea of the hidden cyber attack detection: depending on the Layer of the attacker, additional resources are included into the evaluation, indicated by the opposing surfaces; the more hidden the cyber attack, the broader the information base must be. For example, if there are indications that an attacker has Layer 2 capabilities, additional sources of information can be evaluated and existing data can be re-evaluated. For example, the detection threshold of a system can be adapted and possible anomalies can be recalculated, or external sources of information like news about leaks and vulnerabilities can be consulted for the evaluation. In order to model this, some expectations regarding the opponent have to be defined.

4. TOWARDS UNMASKING HIDDEN CYBER ATTACKS

To enable a retrospective identification of hidden cyber attacks, we propose a new evaluation scheme based on the combination of Cyber Kill Chains and the 3-Layer Vulnerability Model, therefore resulting in a 3d-detection model consisting of the respective Cyber Kill Chains on each Layer and linked with a corresponding timeline on each Layer.

A. Behavioral Rule

In order to implement a corresponding evaluation, it is necessary to create a basis of how the cyber attacker may move. The following Lemmata are proposed; note, that an adversary may try to use this knowledge when choosing her means to again reduce the probability of detection of a cyber attack. Nevertheless, this again may affect other detectable traces by non-controllable side-effects, including changes on the 0day-market, or resulting in an increased operational risk.

1. Vulnerabilities of higher levels are normally only used if there are no vulnerabilities at a lower level available with the same probability of success and the same detection risk.
2. The attacker is more willing to deploy a 0day the lower the risk of detection and the higher the need for operational protection.
3. In times of increased tensions, the direct use of higher-level vulnerabilities is more likely.
4. The attacker prefers the use of unpublished vulnerabilities discovered by others to the exploitation of their own ones, as long as the operational protection requirements allow this.
5. The attacker is more willing to deploy 0days of Layer 2 the older they are, taking their limited life time, decreasing value and higher probability of detection into account.
6. One-shot Layer 3 Vulnerabilities which are exposed with their use, are only deployed in an emergency.
7. Layer 3 Vulnerabilities are all the more likely and more regularly used, the lower the probability of detection of the overall deployment process (including communication channels) and the higher the plausible deniability is.

For the implementation of the respective decisions, the cyber risk must be calculated. Therefore, a quantitative approach is required; currently, we are assessing different quantitative security risk analysis models as well as calculations and experiences in the field of cyber insurance, and we examine how the characteristics of our model, respectively the Lemmata, can be integrated.

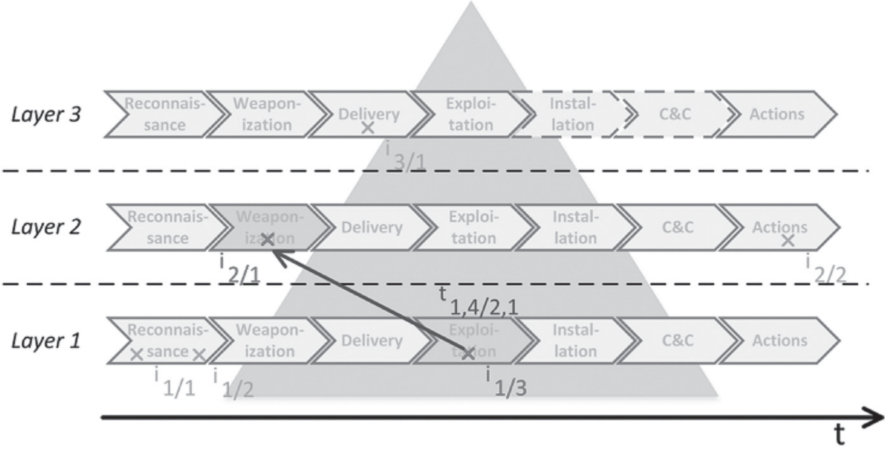
Furthermore, we implemented a matrix that reflects the custom Cyber Kill Chains of each layer. For each layer and each step, characteristics are defined which trigger a further evaluation of another Cyber Kill Chain step and Layer, typically going back in time on the new Layer. Here, sensitivities and threshold values can be adapted for the next evaluation step. There are also characteristics which trigger further evaluations of another step of the current Cyber Kill Chain, or even within the current steps. As the full set of definitions in the transition matrix is a centerpiece of the prototype currently being implemented, and due to the limited space, we are not yet fully presenting the transition matrix at this point. This will be part of our further work. However, for a better understanding of the task and functionality of the transition matrix, two examples of corresponding transitions, respectively actions, are provided:

- Adaptation of thresholds for a re-evaluation of IDS logs in order to detect very slow scans of a network, which remain normally below the detection threshold. For example, one can think about actions like a scan with "paranoid timing" by the network scanner nmap[4]: nmap `-T0`, moving the search window from a Delivery Step back to a Reconnaissance Step. Note, that the effects in the evaluation are based on the change of the associated conditional probabilities and not resulting from the mere change in the sensitivity of the analysis, therefore they are also *not directly visible* in this example.
- Moving the search window from the Exploitation Step on Layer 1 to the Weaponization Step on Layer 2 based on unexpected system behavior or program crashes.

Such a transition is visualized by the arrow going from $i_{1/3}$ to $i_{2/1}$ in Figure 5, and which denotes the selection of $t_{1,4/2,1}$, moving the window of the search from Exploitation Step on Layer 1 to the Weaponization Step on Layer 2 because of identified, abnormal and suspicious system behavior.

[4] Note, that this example is for illustration – as the parameterization of nmap is well-known, changing the search pattern wrt. the timing options of nmap is not enough to improve the detection quality significantly.

FIGURE 5. EXEMPLARY VISUALIZATION OF THE 3D-DETECTION SCHEME FOR THE RETROACTIVE IDENTIFICATION OF HIDDEN CYBER ATTACKS. I REPRESENTS INDICATIONS ON THE RESPECTIVE LAYER AND COMBINED WITH THE EVENT NUMBER, AND T REPRESENTS TRANSITIONS, GOING FROM ONE LAYER TO ANOTHER AND COMBINED WITH THE STEPS OF THE RESPECTIVE CYBER KILL CHAINS. NOTE, THAT THE TIMELINE BETWEEN THE DIFFERENT LAYERS IS *NOT* SYNCHRONIZED.



B. System Composition

Based on the presented Behavioral Rules, a new detection system is proposed as follows: adapted to the respective target network and the systems, the regular information sources such as logs and IDS messages are evaluated in order to recognize steps such as reconnaissance and delivery. Raw data, flows, log entries as well as already processed data, e.g., from an integrated security information and event management (SIEM) system, can be used.

Second and of particular importance, a basic set of “non-wire” and indirect data sources is continuously evaluated for every single Layer, searching for step-specific indications and information of attack. This data is filtered and ranked based on the system environment and stored into a database. These consulted data sets involve, for example, surveillance of pastebin websites and respective forums for the appearance of new leaks, information on vulnerabilities or disclosures, and monitoring of the 0day market and its price development. Information about for instance, operating systems and vulnerabilities of applications which are not used in the system environment are dropped. This database is crucial for the 3d-detection scheme, as a holistic overview is required to identify possible clues related to cyber attacks.

Figure 5 outlines the basic searching process. Based on indicators in the different levels and layers, the probability of transitions are calculated on the basis of the respective cyber risk and according to the defined rules. By the identification of possible transitions, the elements of the potential, multilayer attack paths are dumped for the further, manual evaluation.

5. CONCLUSION AND FURTHER WORK

The entire world is becoming increasingly networked and dependent on the cyberspace. Because of its properties such as a certain degree of anonymity, the cyber arena is more and more interesting for a variety of actors, from script kiddies to nation states. Therefore, a lot of attacks may be seen in cyberspace. Or not. Some of the known attacks have gained broader public attention only by chance. This raises the fundamental question whether some cyber activities may remain hidden even though they have a significant impact on our everyday lives - how can yet unknown cyber involvements get unmasked? If you look at the data and compare it with the possibilities of various attackers, the assumption is reinforced that more incidents may have a (yet undiscovered) cyber background.

The reason detection of sophisticated cyber attacks fails is caused by corresponding steps which are not executed “over the wire”, at least not over the wire of the attacked company, and which are therefore not detectable for conventional systems. Using a 3-Layer Vulnerability Model, the attackers can be characterized based on their capabilities and available attack vectors. By evaluating methods for the analysis of the attack path, it became clear that they are not sufficient to investigate complex attacks, and thus are not suitable for discovering them. To improve the detection of sophisticated cyber attacks and to move towards the identification of yet unknown, hidden cyber attacks, we propose a three-dimensional model based on the combination of the 3-Layer Vulnerability Model and Cyber Kill Chains.

Currently, we are completing a prototypical implementation of our model by using the KNIME Analytics Platform. The next step is building up the required databases, before different data sets can be evaluated. For this purpose, first the databases are filled with the identified information types and sources for the respective Layers and attack steps, and then the logs and system data of selected networks in which cyber attacks were discovered after a long time will be imported. Based on that, the search and evaluation process of the proposed three-dimensional detection scheme will be analyzed to identify necessary adjustments of the cyber risk and transition calculations, as well as the algorithms for adapting the sensitivity of the sensory. As this process requires real-world data of complex networks, we invite companies interested in cooperation and evaluation of their networks and systems to contact us, as having a broad dataset is crucial for enabling the detection scheme. Of course, the used data will be anonymized appropriately.

While the first prototypical implementation will only be able to retroactively identify indications of hidden cyber attacks, the ultimate goal is to minimize the necessary time window required for the process, and to investigate which indicators can also be used to detect an ongoing campaign or campaign under preparation. For that purpose, machine learning techniques will also be applied; a prerequisite, however, is the access to sufficient data sets and their evaluation and marking.

REFERENCES

[1] P. Dreyer, K. Jones, Theresecand Klima, J. Oberholtzer, A. Strong, J. W. Welburn, and Z. Winkelman. *Estimating the global cost of cyber risk: methodology and examples*. RAND Corporation. Technical Report; 2018.

[2] J. Lewis *Economic impact of cybercrime - no slowing down*. McAfee; 2018. Available: https://csis-prod.s3.amazonaws.com/s3fs-public/publication/economic-impact-cybercrime.pdf.

[3] D. P. Fidler. *The Snowden Reader*. Indiana: Indiana University Press; 2015.

[4] WikiLeaks. *Vault 7: Cia hacking tools revealed*. 2017. Available: https://wikileaks.org/vault8/.

[5] WikiLeaks. *Vault 8*. 2017. Available: https://wikileaks.org/vault8/.

[6] Carl Philipp Gottlieb von Clausewitz, *Vom Kriege, Book 1, Chapter 1*, Bassford, Christopher, 1832. Available: https://www.clausewitz.com/readings/VomKriege1832/_VKwholetext.htm .

[7] Carl Philipp Gottlieb von Clausewitz, V*om Kriege, Book 3, Chapter 8*, Bassford, Christopher, 1832. Available: https://www.clausewitz.com/readings/VomKriege1832/_VKwholetext.htm.

[8] Carl Philipp Gottlieb von Clausewitz, *Vom Kriege, Book 3, Chapter 9*, Bassford, Christopher, 1832. Available: https://www.clausewitz.com/readings/VomKriege1832/_VKwholetext.htm.

[9] S. Tzu, The art of war. In: Mahnken TG, Maiolo JA. (eds). *Strategic Studies, A Reader*. 2nd edn. New York: 2008. p28.

[10] Kaspersky Lab. *Denial of Service: how businesses evaluate the threat of DDoS attacks*. Kaspersky Lab. Technical Report; 2015.

[11] Verizon. *Data breach investigation report*. Verizon. Technical Report; 2018.

[12] W. E. Forum *The global risks report*. 14th edition. Switzerland: World Economic Forum. Cologny/Geneva, Switzerland, Technical Report; 2019.

[13] J. Ray, H. Marshall, R. Coderre, E. Cody, and J. Jean *Cyber threatscape report 2018 - midyear cybersecurity risk review*. Accenture Security. Technical Report; 2018.

[14] Ponemon Institute. *2018 Study on global megatrends in cybersecurity*. Ponemon Institute LLC. Technical Report; 2018.

[15] W. Schwab and M. Poujol, *The state of industrial cybersecurity 2018*. Kaspersky Lab. Technical Report; 2018.

[16] M. Dudek. *TRISIS / TRITON / HatMan Malware Repository*. Available: https://github.com/MDudek-ICS/TRISIS-TRITON-HATMAN.

[17] R. Koch and M. Golling, The cyber decade: cyber defence at a x-ing point. in *2018 10th International Conference on Cyber Conflict (CyCon)*. IEEE, 2018, p. 159–186.

[18] M. Giles. *Triton is the world's most murderous malware, and it's spreading*. Available: https://www.technologyreview.com/s/613054/cybersecurity-critical-infrastructure-triton-malware/.

[19] C. Canella, J. Van Bulck, M. Schwarz, M. Lipp, B. von Berg, P. Ortner, F. Piessens, D. Evtyushkin, and D. Gruss, A systematic evaluation of transient execution attacks and defenses. *arXiv preprint* arXiv:1811.05441; 2018.

[20] SonicWall. *2018 Sonicwall cyber threat report mid-year update*. SonicWall Inc. Technical Report; 2018.

[21] Internet Security Research Group. *Let's Encrypt Stats*. Available: https://letsencrypt.org/stats/.

[22] Google. *HTTPS encryption on the web*. Available: https://transparencyreport.google.com/https/overview?hl=en.

[23] Cisco Systems. *Cisco 2018 Annual cybersecurity report*. Cisco Systems Inc. Technical Report; 2018.

[24] US-CERT. *Alert (TA17-075A) https interception weakens TLS security*. Available: https://www.us-cert.gov/ncas/alerts/TA17-075A.

[25] T. Ormandy Sophail: *A critical analysis of Sophos antivirus*. Proc. of Black Hat USA; 2011.

[26] Canadian Journalists For Free Expression. *Snowden archive*. Available: https://www.cjfe.org/snowden.

[27] Electronic Frontier Foundation. *NSA ANT Catalogue*. Available: https://www.eff.org/files/2014/01/06/20131230-appelbaum-nsa\s\do5(a)nt\s\do5(c)atalog.pdf.

[28] Infosec Institute. *A close look at the NSA monitor catalog - server hacking*. Available: https://resources.infosecinstitute.com/close-look-nsa-monitor-catalog-server-hacking/.

[29] J. Assange. *Vault 7: CIA hacking tools revealed*. Available: https://wikileaks.org/ciav7p1/.

[30] WikiLeaks. *Marble Framework*. Available: https://wikileaks.org/vault7/#Marble.

[31] WikiLeaks. *Weeping Angel*. Available: https://wikileaks.org/vault7/#Weeping.

[32] R. M. Lee, M. J. Assante, and T. Conway, *Analysis of the cyber attack on the Ukrainian power grid*, Electricity Information Sharing and Analysis Centre. Technical Report; 2016.

[33] P. Koppe, B. Kollenda, M. Fyrbiak, C. Kison, R. Gawlik, C. Paar, and T. Holz, Reverse engineering x86 processor microcode. *Proceedings of the 26th USENIX Security Symposium*. USENIX Association; 2017. p.1163–1180.

[34] S. Skorobogatov and C. Woods, Breakthrough silicon scanning discovers backdoor in military chip. *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer. 2012.p. 23–40.

[35] J. Robertson and M. Riley. *The Big Hack: How China used a tiny chip to infiltrate U.S. companies*. Available: https://www.bloomberg.com/news/features/2018-10-04/the-big-hack-how-china-used-a-tiny-chip-to-infiltrate-america-s-top-companies.

[36] J. Naughton. *The tech giants, the US and the Chinese spy chips that never were... or were they?* Available: https://www.theguardian.com/commentisfree/2018/oct/13/tech-giants-us-chinese-spy-chips-bloomberg-supermicro-amazon-apple.

[37] T. Markettos. *Making sense of the Supermicro motherboard attack*. Available: https://www.lightbluetouchpaper.org/2018/10/05/making-sense-of-the-supermicro-motherboard-attack/.

[38] J. A. Tirpak. *Find, Fix, Track, Target, Engage, Assess. Air Force Magazine. 2000*;83 (7): 24–29. Available: http://www.airforcemag.com/MagazineArchive/Documents/2000/July.

[39] LM Corporation. *The Cyber Kill Chain*. Available: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

[40] J. R. Gosler and L. von Thaer, *Task force report: resilient military systems and the advanced cyber threat*. Washington, DC: Department of Defense, Defense Science Board; 41, 2013.

[41] The MITRE Corporation. *Common vulnerabilities and exposures (CVE)*. Available: https://cve.mitre.org/.

[42] R. Koch and T. Kühn, Defending the grid: backfitting non-expandable control systems. *2017 9th International Conference on Cyber Conflict (CyCon)*. IEEE, 2017. p. 1–17.

[43] Q. Chen and R. A. Bridges, Automated behavioral analysis of malware a case study of WannaCry ransomware. *arXiv preprint* arXiv:1709.08753; 2017.

[44] N. Grossman. *Eternalblue - everything there is to know*. Available: https://research.checkpoint.com/eternalblue-everything-know/.

[45] L. Ablon and A. Bogart, *Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits*. Rand Corporation; 2017.

[46] CrowdStrike. *2019 Crowdstrike global threat report - adversary tradecraft and the importance of speed*. CrowdStrike. Technical Report; 2019.

[47] D. J. Bernstein, T. Chou, C. Chuengsatiansup, A. Hülsing, T. Lange, R. Niederhagen, and C. van Vredendaal, *How to manipulate curve standards: a White Paper for the Black Hat*. Cryptology ePrint Archive. Report 2014/571; 2014.

[48] Accenture. *Cyber threatscape report 2018 - midyear cybersecurity risk review*. Accenture. Technical Report; 2018.

[49] K. D. Mitnick and W. L. Simon. *The Art of Deception: Controlling the Human Element of Security*. New York: Wiley & Sons; 2011.

BIOGRAPHIES

This section includes the biographies of the editors and co-editors and of those authors who presented their research at the conference.

Editors

Siim Alatalu joined the NATO Cooperative Cyber Defence Centre of Excellence in January 2015 as Head of International Relations, to lead the development of the Centre's relations with its growing network of partners from government, military, academia and industry. In 2018, he joined the Centre's Strategy Branch to be in charge of cyber strategy and policy research and training related to NATO and the European Union. His prior professional career includes several advisory and managerial positions at the Estonian Ministry of Defence from 2001, including a diplomatic assignment at the Estonian Delegation to NATO in Brussels. During the Estonian Presidency of the Council of the European Union in 2017, he also co-led the development of the EU's cyber policy and strategy as the Vice Chair of the Council's Horizontal Working Party for Cyber Issues. Siim is a graduate of the Maxwell School of Syracuse University (Master of Arts in International Relations, 2006, as a Fulbright Fellow), the Baltic Defence College (Higher Command Studies Course, 2011) and the University of Tartu (BA in History, 2001). He is currently pursuing his PhD at the Tallinn University of Technology.

Commander **Stefano Biondi** is an Italian Navy IT officer with more than 20 years of professional experience in the fields of IT and Intelligence. During his tour of duty at the NATO Cooperative Cyber Defence Centre of Excellence as a Cyber Intelligence researcher in the Operations Branch, Stefano developed the Operational Cyber Threat Intelligence Course aimed to fill the gap between the operational and tactical levels of military intelligence operations. Prior to this, he worked in Military Intelligence in management and technical positions focusing on information technology, cybersecurity and communication systems to develop and enhance information gathering capabilities. He holds a level II Master's degree in Information Technology from the Department of Electronic Engineering of Rome 2 - Tor Vergata University, University of Rome.

Tomáš Minárik is a researcher in the NATO Cooperative Cyber Defence Centre of Excellence's Law Branch. His current research focuses on the legal aspects of cyberspace operations (international cyber law and the interactive toolkit project), the activities of international organisations in cyberspace (the INCYDER project), the right to privacy and anonymity networks. He also helps to prepare and execute the legal aspect of the Locked Shields and Crossed Swords exercises and is responsible for the legal track of the CyCon Conference. He has worked as a legal adviser at the National Cyber Security Centre of the Czech Republic, and before that at the International Law Department of the Czech Ministry of Defence. He holds a law degree from Charles University in Prague.

Massimiliano Signoretti is a Lieutenant Colonel in the Italian Air Force and a researcher in the Law Branch of the NATO Cooperative Cyber Defence Centre of Excellence. His research area is public international law, international humanitarian law and the law of armed conflict. He graduated in law at the University of Rome (La Sapienza) and was admitted to the Bar. He also studied at the University of Stockholm, faculty of law, and has worked at the Italian Defence General Staff – Office of International Legal Affairs. His career also includes four years' service at the NATO Partnerships Division, ACO SHAPE, Belgium. His academic achievements include a Master's degree in Strategic International Military Studies and a level II Master's degree in International Humanitarian Law and the Law of Armed Conflict.

Lieutenant **Ihsan Tolga** is a researcher at the NATO Cooperative Cyber Defence Centre of Excellence. Prior to taking up his post, he worked for three years in the Turkish Navy Research Centre Command and before that as an officer in the Turkish Navy Submarine Fleet Command. He is a graduate of the Turkish Naval Academy (BSc in Computer Engineering) and the University of Southern California (MSc in Computer Science).

Major **Gábor Visky** is a researcher in the Technology Branch of the NATO Cooperative Cyber Defence Centre of Excellence, where his main field of expertise is industrial control systems. Gábor´s prior assignments include 15 years designing hardware and software for embedded control systems, and researching their vulnerabilities by reverse engineering. Gábor holds an MSc degree in Information Engineering in the specialty of industrial measurement, and a BSc in the field of telecommunication.

Authors

Gil Baram is Head of Research at the Yuval Ne'eman Workshop for Science, Technology and Security, and a research fellow at the Blavatnik Interdisciplinary Cyber Research Centre (ICRC), Tel Aviv University. She is a PhD candidate at Tel Aviv University School of Politics and International Relations, specialising in cyber conflicts and national security. Currently she is an adjunct fellow at the Centre of Excellence for National Security, Nanyang Technological University, Singapore. Her current research focuses on national strategies during cyber conflicts. She explores the reasons why, although cyber attacks are covert in nature and therefore can be hidden or denied, in many cases countries choose to reveal the attack and 'go public' about it. Gil is developing a new conceptual-theoretical framework and is examining it by combining qualitative and quantitative research methods. Gil Baram holds an MA in Security Studies from Tel-Aviv University (Magna Cum Laude).

Alicia Bargar works as a research engineer at the Johns Hopkins Applied Physics Laboratory, where she draws on her background in machine learning and social network analysis to research social phenomena in the online domain. Projects include the investigation of wildlife trafficking on social media, multilinguals' role in online message propagation, and the study of online protest dynamics. Ms. Bargar earned an MSc in Computer Science – Machine Learning, from the Georgia Institute of Technology under a Presidential Fellowship. She is a member of the Largescale Data Analytics Systems team and leads the Network Analysis interest group.

Brad Bigelow has over 30 years' experience in military communications, information security, space operations and project and programme management, including 25 years as a US Air Force officer. 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 **Joe Burton** is Senior Lecturer at the New Zealand Institute for Security and Crime Science, University of Waikato. His research focuses on cybersecurity, 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* (SUNY Press: New York). He has recently been a visiting researcher at the NATO Cooperative Cyber Defence Centre of Excellence and is a recipient of the Taiwan Fellowship and US State Department's Study of the US Institutes (SUSI) fellowship.

Professor **Michele Colajanni** has been Professor of Security Engineering at the University of Modena, Italy, since 2000. His research interests include system and network security, scalable and reliable architectures and security analytics. He directs the Research Centre on Security and Safety, the Cyber Academy for ethical hackers, and Master's degree in Cyber Defence governance for the Armed Forces and the Bologna Business School.

Pierre Dumont has been a cybersecurity engineer at Kudelski Security (Switzerland) since June 2018 as part of the DevOps team. He delivers software solutions to process security alerts from customers that are analysed by Kudelski's Security Operations Centre. He earned a Master's degree in Electrical Engineering and Information

Technology from ETH Zürich, by developing a processing pipeline of internet scans and analysing the collected data at Kudelski's R&D team. He participated at the NATO Cooperative Cyber Defence Centre of Excellence in creating a visualisation system for their cyber defence exercise during his internship at armasuisse, the R&D centre of the Swiss Department of Defence.

Dr **Roman Graf** is a research engineer at the Centre for Digital Safety and Security in the Austrian Institute of Technology GmbH. He works on cybersecurity and data analytics topics, contributing to the development of several European research projects like Ecossian, Planets, Assets and SCAPE. He has published widely in the area of cybersecurity and risk management in digital preservation, being an active member of the Open Preservation Foundation (OPF). Dr Graf supported the development of cyber threat intelligence solution CAESAIR, serving as one of the key developers, and contributed a module to the Open Source Threat Intelligence Platform (MISP).

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 Hartmann 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 cybersecurity proposals within Horizon 2020 and related areas.

Jason Healey is a Senior Research Scholar at Columbia University's School for International and Public Affairs, specialising in cyber risk and conflict. Prior to this, he was the founding director of the Cyber Statecraft Initiative of the Atlantic Council where he remains a Senior Fellow. He is the editor of the first history of conflict in cyberspace, *A Fierce Domain: Cyber Conflict, 1986 to 2012*. A frequent speaker on these issues, he is a 'top-rated' speaker for the RSA Conference and won the inaugural 'Best of Briefing Award' at Black Hat. He helped the world's first cyber command in 1998, the Joint Task Force for Computer Network Defense, where he was one of the early pioneers of cyber threat intelligence. During his time in the White House, he was a director for cyber policy, coordinating efforts to secure US cyberspace and critical infrastructure. He created Goldman Sachs' first cyber incident response team and later oversaw the bank's crisis management and business continuity in Asia. He served as the vice chair of the Financial Services Information Sharing and Analysis Center (FS-ISAC). He is on the review board of the DEF CON hacker conference and served on the Defense Science Board task force on cyber deterrence. He started his career as a US Air Force intelligence officer with jobs at the Pentagon and National Security Agency and is President of the Cyber Conflict Studies Association.

Daniel Kapellmann Zafra works as a senior cyber threat intelligence analyst for FireEye's cyber-physical team. A former Fulbright scholar, he holds a Master's degree in Information Management from the University of Washington, with a specialisation in cybersecurity. His multidisciplinary background includes work experience ranging from consulting to IT planning and architecture in the energy sector. Among other achievements, he was awarded first place at Kaspersky Academy Talent Lab in 2017 for designing an application to address security beyond antivirus. In his free time, he is also a journalist, writing for organisations such as Bertelsmann Stiftung, Siemens Stiftung, OECD and Fair Observer.

Dr **Joonsoo Kim** is a senior researcher in National Security Research Institute (NSR), South Korea. His current research focus is on how to design and execute effective national cybersecurity exercises that use a multidisciplinary approach to simulate realistic cyber threats to national critical digital assets. Since 2018, he has led the NSR team to participate in Locked Shields exercises and also worked as a visiting scholar at TalTech, Estonia. Before joining NSR, he gained diverse industrial experience in Intel, Qualcomm and IBM Research Lab. He is a graduate of Seoul National University (BSc in Electrical Engineering) and the University of Texas at Austin (MSc and PhD in Electrical and Computer Engineering).

Commander Dr **Robert Koch** is a General Staff Officer of the German Federal Armed Forces. Robert received his Diploma in Computer Science in 2002. After that, he had a 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 visualisation and the application of artificial intelligence. Currently, he is building up the new penetration testing capability at the Cyber-Security Centre of the Federal Armed Forces.

First Lieutenant **Jiyoung Kong** is an ROK Air Force officer. She has been working in the Agency for Defence Development as a cybersecurity researcher for three years. Kong has a Bachelor's degree from Korea University, Department of Cyber Defence, and is currently taking the combined Master's and PhD course in Information Security at Korea University.

Professor **Jeff Kosseff** is an assistant professor of cybersecurity law in the United States Naval Academy's Cyber Science Department. His latest book, *The Twenty-Six Words That Created the Internet: A History of Section 230 of the Communications Decency Act*, was published in Spring 2019 by Cornell University Press. He also is the author of *Cybersecurity Law*, a textbook and treatise published by Wiley in 2017, with a second edition forthcoming in October 2019. Jeff practised cybersecurity, privacy, and First Amendment law at Covington & Burling, and clerked for Judge Milan D. Smith, Jr. of the United States Court of Appeals for the Ninth Circuit and Judge Leonie M. Brinkema of the United States District Court for the Eastern District of Virginia. Before becoming a lawyer, he was a technology and political journalist for *The Oregonian* and was a finalist for the Pulitzer Prize for national reporting and a recipient of the George Polk Award for national reporting. He received a JD from Georgetown University Law Centre, and a BA and MPP from the University of Michigan.

Kenneth Kraszewski is a doctoral candidate in international law at the University of Helsinki. He divides his time between researching international law regulation of operations in cyberspace, particularly those falling beneath the threshold of use of force, and the private practice of law at a prominent business law firm in Helsinki, Finland.

Artūrs Lavrenovs is a researcher at NATO Cooperative Cyber Defence Centre of Excellence, focusing on the web and network technologies while teaching security courses, performing applied and academical research, and contributing to cyber exercises. Artūrs has taught web technology and IT security courses at the University of Latvia, where he is currently a PhD student doing research in the cybersecurity domain.

Dr **Vincent Lenders** is the head of the newly created Swiss Cyber-Defence Campus and of the C4I group at armasuisse Science and Technology. He is also the co-founder and chairman of the executive boards of the OpenSky Network and Electrosense associations. He graduated with a MSc and PhD in Electrical Engineering and Information Technology from ETH Zurich and was a postdoctoral researcher at Princeton University. He was also Industrial Director of the Zurich Information Security and Privacy Centre (ZISC) at ETH Zurich from 2012 to 2016 and has served for almost ten years as Research Director for Cyber and Information for the Swiss DoD. His work has appeared in more than 100 publications for prestigious peer-reviewed international conferences and journals and has received various best paper awards.

Professor **Martin C. Libicki** (PhD, UC Berkeley 1978) holds the Keyser Chair of Cybersecurity Studies at the US 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 of *Conquest in Cyberspace: National Security and Information Warfare* and various related RAND monographs. Prior employment includes 12 years at the National Defense University, three years on the Navy Staff (logistics) and three years for the US GAO.

Bilyana Lilly is a Pardee Fellow at the RAND Corporation in Los Angeles, California. She conducts research and leads projects on Russian cyber threat actors, information warfare, election cybersecurity, machine learning, and cyber indications and warning. She regularly attends Russia's military technical forum, 'Army'. Prior to joining Pardee RAND, Ms. Lilly was a research associate at the Brookings Institution where she focused on US security strategy and NATO's policy toward emerging powers. She also worked at the Conference on Disarmament at the United Nations in Geneva, Switzerland. She is the author of the book *Russian Foreign Policy Toward Missile Defence*, which includes interviews that Ms. Lilly conducted with Russia's Deputy Defence Minister and Deputy Minister of Foreign Affairs.

Roland Meier is a third-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 Meier received his Master's degree in Electrical Engineering and Information Technology from ETH Zürich in 2015.

Senior Captain **Erwin Orye** joined the NATO Cooperative Cyber Defence Centre of Excellence in January 2017 as a Researcher in the Strategy Branch. He is currently pursuing a PhD in aviation cybersecurity in TalTech. After finishing the Royal Military Academy in Brussels, Belgium in 2000 as a civil engineer in telecommunications, he worked on technical aspects of tactical communications. His career began as deputy commander and progressed to commander of the Belgian military technical support team for tactical communications and information systems. Afterwards, he became a material resources manager responsible for tactical radios. He next served as the staff officer for plans and training in the Belgian 6th Signal Battalion. In 2013, he became team lead for material resources managers responsible for secure and tactical networks. In his final position before moving to the NATO CCD COE in Estonia, he worked at the Belgian military Cyber Security Operations Centre (CSOC).

Dr **Anna-Maria Osula** is a senior policy officer at Guardtime, a software security company that offers solutions for data governance and real-time detection and mitigation of cyberattacks based on blockchain technologies. She also serves as senior researcher and lecturer at Tallinn University of Technology. Previously, she worked as a legal researcher at the NATO CCD COE, where her research areas included national cybersecurity strategies, international organisations, international criminal cooperation and law. During 2013-2015, she was the lead for the NATO CCD COE 'Cyber Norms' project and the co-editor of the book *International Cyber Norms: Legal, Policy and Industry Perspectives*. In addition to a PhD in law from the University of Tartu, she also holds an LLM in IT law from Stockholm University.

Nikolas Ott is a Project Manager in the cyber/ICT security team within the Transnational Threats Department of the Organisation for Security and Co-operation in Europe (OSCE). In this capacity, he manages the confidence-building measure implementation efforts and co-ordinates international outreach. Previously, he has worked at the NATO CCD COE, the NATO Cyber Defence Section and the Hertie School of Governance. He is an alumnus of the Mercator Fellowship of International Affairs and the German Academic Scholarship Foundation. He holds an MA in Law and Diplomacy from The Fletcher School of Law and Diplomacy (Tufts University) and a BA in Political Science from the Freie Universität Berlin.

Captain **Barış Egemen Özkan** is a Turkish Navy officer. Following warfare officer tasks at unit level, he has gained more than 15 years' experience in development and programme management of C4ISR systems for naval systems at the Turkish Navy Research Centre Command. He was assigned to the Operations and Exercise Branch as Head of Cyber Division at SHAPE in 2016, and has been intimately involved in the operationalisation of cyberspace as a new domain of operations within NATO. Since the establishment of the Cyberspace Operations Centre at SHAPE in August 2018, he has been responsible for planning cyberspace operations at SHAPE as Plans Branch Head.

James Pavur is a second-year DPhil student at Oxford University. He was awarded a Rhodes Scholarship in 2017 to study at Oxford's Cybersecurity Centre for Doctoral Training, where his research focuses on the security of satellites and space-based systems. His professional experience includes work in computer forensics, embedded reverse engineering and vulnerability research. He also holds a Bachelor's degree from Georgetown University's Walsh School of Foreign Service where he graduated as valedictorian with a major in Science, Technology and International Affairs.

Dr **Przemysław Roguski** is a lecturer in law at the Jagiellonian University in Kraków, Poland, and an expert on cybersecurity and international law at the Kościuszko Institute.

His research focuses on the law of peacetime cyber operations and different aspects of international law relating to cybersecurity, ICT and internet governance. Previously, Przemysław has worked in private practice and as a lecturer for the German Academic Exchange Service (DAAD). He holds law degrees from the University of Mainz and Trinity College Dublin and a PhD in international law from Jagiellonian University.

Dr **Barrie Sander** is a Postdoctoral Fellow at Fundação Getulio Vargas (FGV) in Brazil. His research interests include global cybersecurity norms, human rights and technology, and international criminal law. Currently, Barrie is examining the paradigms of international law that apply to State-sponsored cyber influence operations on elections, the governance of online speech by social media platforms, and the use of new information and communication technologies in mass atrocity contexts. Barrie holds a PhD in International Law from the Graduate Institute of International and Development Studies (IHEID), an LLM in Public International Law from the University of Leiden, and a BA in Law from Jesus College, Cambridge.

Dr **Max Smeets** is a Cybersecurity Fellow at Stanford University Center for International Security and Cooperation (CISAC). He is also a non-resident Cyber Security Policy Fellow at New America and a Research Associate at the Centre for Technology & Global Affairs, University of Oxford. Max was awarded the annual 2018 Amos Perlmutter Prize by the Journal of Strategic Studies for the most outstanding manuscript submitted for publication by a junior faculty member. In 2015, he received the Young Writer Award of the German Marshall Fund for an article written jointly with George Bogden. Max was previously a College Lecturer at Keble College, Oxford. He has also held research positions at the Oxford Cyber Studies Programme, Columbia University SIPA, Sciences Po CERI, and NATO CCD COE. He holds an undergraduate degree (interdepartmental major with minor in statistics) from University College Roosevelt, Utrecht University, and an MPhil (Brasenose College) and DPhil (St. John's College) in International Relations from the University of Oxford.

Dr **Simona R. Soare** (PhD 2011) is Security and Defence Advisor to the Vice-President of the European Parliament and Associate Researcher with the Institut d'Études Européennes (IEE) at Université Saint Louis in Brussels. Previously she served as a researcher with the Romanian MoD and was a US DoS fellow. Simona's research focuses on transatlantic defence cooperation and capability development. She is particularly interested in exploring the impact of artificial intelligence, advanced robotics and automation on military capabilities, operations and interoperability in a transatlantic context. She recently published on European strategic autonomy and the impact of European defence exports on European security.

Professor **Armando Tacchella** obtained his PhD in Electrical and Computer Engineering from the University of Genoa in 2001 and his ‘Laurea’ (MSc equivalent) in Computer Engineering in 1997. His research interests are in the field of AI, with a focus on the safety and security of intelligent systems. On this and other topics, he has published about 100 papers in international conferences and journals. In 2007 the Italian Association of Artificial Intelligence (AI*IA) awarded him the Marco Somalvico Prize for the best young Italian researcher in AI.