2017

9th International Conference on Cyber Conflict:

Defending the Core

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

CCDCOE
NATO Cooperative Cyber Defence
Centre of Excellence
Tallinn Estonia

IEEE
Advancing Technology
for Humanity

30 MAY - 02 JUNE 2017, TALLINN, ESTONIA

2017 9TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT: DEFENDING THE CORE

IEEE Catalog Number: CFP1726N-PRT
ISBN (print): 978-9949-9904-0-5
ISBN (pdf): 978-9949-9904-1-2

COPYRIGHT AND REPRINT PERMISSIONS

PRINTED COPIES OF THIS PUBLICATION ARE AVAILABLE FROM:

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 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 is a NATO-accredited knowledge hub, research institution, and training and exercise facility. The Tallinn-based international military organisation focuses on interdisciplinary applied research, consultations, trainings and exercises in the field of cyber security.

NATO CCD COE is the home of the Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. The Centre organises the world's largest and most complex international technical cyber defence exercise Locked Shields and the annual conference on cyber conflict, CyCon.

The Centre is a multinational and interdisciplinary hub of cyber defence expertise, uniting practitioners from 20 nations. The heart of the Centre is a diverse group of experts: researchers, analysts, trainers, educators. The mix of military, government and industry backgrounds means the NATO CCD COE provides a unique 360-degree approach to cyber defence. The organization supports its member nations and NATO with cyber defence expertise in the fields of technology, strategy, operations, and law.

The Centre is staffed and financed by its sponsoring nations and contributing participants. Belgium, the Czech Republic, Estonia, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Poland, Slovakia, Spain, Turkey, the United Kingdom and the United States are signed on as Sponsoring Nations of the NATO Cooperative Cyber Defence Centre of Excellence. Austria and Finland have become Contributing Participants, Sweden has applied for membership in the same format, a status eligible for non-NATO countries.

CYCON 2017 SPONSORS

TECHNICAL SPONSOR

IEEE

Advancing Technology
for Humanity

DIAMOND SPONSORS

Microsoft

guardtime

SPONSORS

Cyber Test
Systems

ROHDE&SCHWARZ
Cybersecurity

TABLE OF CONTENTS

INTRODUCTION

The 9th International Conference on Cyber Conflict (CyCon 2017), organised by the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE), once again brings together professionals from governments, the military, academia and the private sector with the aim of discussing and proposing solutions for issues related to cyber security and defence. CyCon continues to serve the cyber security community's technical experts, strategic thinkers, policymakers and lawyers as an interdisciplinary platform for networking and sharing knowledge.

The call for papers for CyCon 2017 invited submissions on the topic 'Defending the Core'. Digitalisation has transformed, often disrupted, our way of life, bringing innumerable social and economic benefits, while at the same time technological dependencies have altered security risks. No future conflict is likely to be fought without a cyber element. However, establishing effective defensive measures is difficult. The cross-border and ever-expanding nature of digital technologies complicates our understanding of critical cyber dependencies, threats and vulnerabilities. The call for papers was thus looking to address questions such as: What are the 'core' elements of cyber security? How do they relate to essential assets and principles in the technical, legal and political spheres? How can critical (information) infrastructure be protected? How can critical vulnerabilities be mitigated and the most detrimental threats countered? How can legal frameworks be established and applied in the cyber security realm? What technologies can help nations to counter emerging cyber threats? How can effective cyber security strategies be developed and implemented? What should be the role of the armed forces? How can cyber operations against core (national) assets be deterred?

These questions provided inspiration for nearly 200 abstract proposals submitted in October 2016. After many rounds of peer review, 14 articles were accepted for this proceedings book, which form the 'backbone' of the conference's sessions and workshops.

CyCon's interdisciplinary nature is reflected in this collection of articles, which can be broadly categorised into three sections: strategy, law and technology. The publication starts with a focus on strategic cyber security issues, as **Martin Libicki** provides his perspective on how states can establish effective international norms to limit cyber espionage. This is followed by **Max Smeets**, who analyses the possible benefits and risks of organisational integration of national offensive cyber capabilities. When discussing the challenges stemming from NATO's decision to recognise cyberspace as an operational domain, **Brad Bigelow** highlights the importance of mission assurance and advocates for a clear role for the NATO Command Structure. The strategy section ends with **Kenneth Geers**, who emphasises the importance and often underestimated value of traffic analysis in cyberspace.

Articles devoted to legal issues start with **Ido Sivan-Sevilla**'s study of the dynamics of United States federal law with regard to the privacy and security debate. Privacy is also addressed by **Eliza Watt**, who writes about the role of international human rights law in the protection of online privacy, focusing on the extraterritorial application of human rights treaties.

Jeffrey Biller's contribution then looks at a topical issue in international humanitarian law: the misuse of protected indicators in cyberspace. International humanitarian law is also represented by the following article by **Tassilo V. P. Singer**, who examines the possible extension of the period of direct participation in hostilities due to autonomous cyber weapons. Emerging international law is discussed in the last two legal articles. **Kubo Mačák** provides his view of how general international law is influenced by the development of the cyber law of war; and finally, **Peter Z. Stockburger** observes that there may be arising a new *lex specialis* governing state responsibility for third party cyber incidents: a 'control and capabilities' test.

The third section of the book covers technical cyber security matters. Focussing on the defence of core infrastructure, **Robert Koch** and **Teo Kühn** begin by introducing their concept of building an effective intrusion detection system, based on voltage levels and current drain, to protect unsecure industrial control systems. Continuing with the subject of defending cyber-physical systems, **Martin Strohmeier et al.** propose the establishment of a separate verification layer for sensitive wireless data, powered by crowdsourced sensors connected to the Internet. **Fabio Pierazzi et al.** then tackle the detection of advanced cyber attacks as they introduce a novel online approach for identifying intrusions, providing an alternative to existing frameworks. Last but not least, **Riccardo Longo et al.** look at the resilience of certification authorities in a scenario of a large-scale cyber attack and propose a solution by analysing the security of a blockchain-based Public Key Infrastructure protocol.

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

Academic Review Committee Members:

- Prof Gabriel Jakobson, CyberGem Consulting; Co-Chair of the Academic Review Committee
- Dr Rain Ottis, Tallinn University of Technology; Co-Chair of the Academic Review Committee
- Dr Iosif Androulidakis, Ioannina University
- Prof Giuseppe Bianchi, University of Roma
- Cdr Stefano Biondi, NATO CCD COE
- Bernhards Blumbergs, NATO CCD COE
- Maj Pascal Brangetto, Ministry of Defence, France
- Dr Aaron Brantly, The United States Military Academy
- Dr Russell Buchan, University of Sheffield
- Dr Steve Chan, MIT; Harvard University
- Prof Thomas Chen, Swansea University
- Prof Michele Colajanni, University of Modena and Reggio Emilia
- Dr Christian Czosseck, NATO CCD COE Ambassador; German Armed Forces CERT
- Prof Dorothy E. Denning, Naval Postgraduate School

- Gunnar Faith-Ell, NATO CCD COE
- Dr Kenneth Geers, NATO CCD COE Ambassador; Comodo
- Keir Giles, Conflict Studies Research Centre
- Rudi Gouweleeuw, Netherlands Organisation for Applied Scientific Research (TNO)
- Prof Michael Grimaila, Air Force Institute of Technology
- Prof Dimitris Gritzalis, University of Economics of Athens
- Dr Jonas Hallberg, Swedish Defence Research Agency (FOI)
- Jason Healey, Columbia University
- Dr Trey Herr, Harvard Kennedy School
- Prof David Hutchison, Lancaster University
- Margarita Levin Jaitner, Swedish Defence University
- Cpt Raik Jakschis, NATO CCD COE
- Urmet Jänes, Institute of Electrical and Electronics Engineers (IEEE)
- Maj Harry Kantola, NATO CCD COE
- Kadri Kaska, Estonian Information System Authority
- Prof Sokratis K. Katsikas, Norwegian University of Science & Technology
- Prof Jörg Keller, Hagen Open University
- Prof Panagiotis Kikiras, European Defence Agency
- LtCol Jens van Laak, Bundeswehr
- Clare Lain, NATO CCD COE
- Prof Konstantinos Lambrinoudakis, University of Piraeus
- LtCol Franz Lantenhammer, NATO CCD COE
- Dr Scott Lathrop, United States Military Academy
- Dr Sean Lawson, University of Utah
- Corrado Leita, Lastline Inc
- Prof Jarno Limnéll, Aalto University
- Dr Lauri Lindström, NATO CCD COE
- Dr Matti Mantere, Intel
- Prof Evangelos Markatos, University of Crete
- Cpt Markus Maybaum, NATO CCD COE Ambassador; Fraunhofer FKIE
- Prof Michael Meier, Bonn University
- Roy Mente, Netherlands Organisation for Applied Scientific Research (TNO)
- Tomáš Minárik, NATO CCD COE
- Dr Jose Nazario, Invincea Inc
- Dr Lars Nicander, Swedish National Defence College
- Dr Anna-Maria Osula, NATO CCD COE
- Nikolas Ott, Mercator Program Center for International Affairs (MPC)
- Liisa Past, Estonian Information System Authority
- Dr Patryk Pawlak, EU Institute for Security Studies
- Dr Roel Peeters, KU Leuven
- Raimo Peterson, NATO CCD COE
- Mauno Pihelgas, NATO CCD COE
- LtCol Nikolaos Pissanidis, NATO CCD COE
- Prof Gabi Dreo Rodosek, Bundeswehr University Munich

- Henry Rõigas, NATO CCD COE
- Prof Juha Röning, University of Oulu
- Prof Julie J.C.H. Ryan, George Washington University
- Prof Massimiliano Sala, University of Trento
- LtCol Jan Stinissen, The Netherlands Ministry of Defence
- Prof Bradley Thayer, Tallinn University
- Dr Lauri Tuovinen, University of Oulu
- Dr Jens Tölle, Fraunhofer FKIE
- Dr Enn Tõugu, Tallinn University of Technology
- Dr Risto Vaarandi, Tallinn University of Technology
- Teemu Uolevi Väisänen, VTT Technical Research Centre of Finland
- Ann Väljataga, NATO CCD COE
- Matthijs Veenendaal, NATO CCD COE
- Prof Ari Visa, Tampere University of Technology
- Dr Jozef Vyskoc, VaF Rovinka and Comenius University Bratislava
- Prof Bruce Watson, Stellenbosch University
- Prof Sean Watts, Creighton University
- Cdr Michael Widmann, NATO CCD COE
- Prof Stefano Zanero, Milan University

We also thank the Institute of Electrical and Electronic Engineers (IEEE), as the IEEE Estonia Section has been a technical co-sponsor for CyCon 2017 and this publication. In addition, we express our gratitude to the NATO CCD COE support staff for their excellent organising skills and assistance during the publication process. Finally, we thank all the authors of the papers collated in this publication for their thought-provoking submissions, their friendly cooperation and their commitment to advancing research on cyber security.

The CyCon 2017 Agenda Management Board

Cdr Stefano Biondi
Gunnar Faith-Ell
Cpt Raik Jakschis
Lauri Lindström
Tomáš Minárik
Mauno Pihelgas
LtCol Nikolaos Pissanidis
Henry Rõigas

NATO Cooperative Cyber Defence Centre of Excellence
Tallinn, Estonia, May 2017

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

The Coming of Cyber Espionage Norms

Martin Libicki
Distinguished Visiting Professor
US Naval Academy

Abstract: The proposition that cyber espionage is acceptable state behavior, even as cyber attack is unacceptable, is in question. The United States has raised objections to certain types of cyber espionage activity, notably: (1) Chinese economically-motivated cyber espionage; (2) the (feared) transfer of data taken from the US Office of Personnel Management (OPM) and provided to criminals; and (3) Russian doxing attacks, particularly against the Democratic National Committee (DNC). In effect, the United States has been edging towards advocating a new class of norms for cyber espionage – countries may carry it out, but not use the results for other than traditional intelligence purposes, that is for informing national security decision-making. Other forms of cyber espionage may come to be viewed as unacceptable, notably the uses of cyber espionage to enable cyber attacks on critical infrastructure.

Establishing a norm that holds some forms of cyber espionage to be acceptable and others not would raise issues. First, can the United States and its friends define such norms in ways that render unacceptable (many of) those practices it finds objectionable, but do not prevent its own practices from being deemed unacceptable? In particular, can there be norms expressed in ways that allow all targets and methods to be used but restrict only what can be done with the information collected? Second, can monitoring regimes be developed to distinguish acceptable from unacceptable cyber espionage and attribute such actions – not only correctly, but in ways that are accepted widely enough to dissuade further such activity?

Keywords: *cyberspace, espionage, norms, cyber espionage*

1. INTRODUCTION

For hundreds of years, physical attacks by countries have been treated as unacceptable, and hence as reasonable pretext for a forceful response. Extending this principle, the United Nations Group of Governmental Experts declared in 2013 that existing international law (originally developed for conventional combat) also applies in cyberspace (United Nations 2013).

Conversely, espionage by countries has been treated as acceptable state behavior, hence not a reasonable pretext. This understanding has been carried over into cyberspace. Responsible nations may carry out cyber espionage (violating a system's confidentiality), but they may not carry out cyber attacks (operations that violate a system's integrity or availability).

In recent years, the United States, together with like-minded countries, has indicated that it does not feel that all cyber espionage is acceptable state behavior. It has complained about Chinese industrial cyber espionage since early 2010, winning a presidential-level agreement from China to ban such a practice (White House 2015b). It was considering declaring unacceptable the transfer of personally identifiable information harvested from the US Office of Personnel Management (OPM) hack into cybercrime markets (the issue had never been raised because there was scant evidence of any such transfer). In late 2016, President Obama argued that using stolen information to maliciously publish discrediting or private information on individuals (doxing) and thereby interfere with electioneering deserved and would get a response (Detrow 2016). Soon thereafter, he levied sanctions on Russia. This was after a great deal of urging by Congressman Schiff and many others for a strong response (Williams 2016). The United States is not alone in such sentiments. A G-20 communiqué extended the Xi-Obama agreement to 18 other countries (Poplin 2015). In looking at the potential for similar incidents affecting its politics, the UK has shown that it is similarly worried about Russian "cyber war" (Haynes 2016).

Perhaps needless to add, the tussle over cyberspace norms reflects not only the newness of the medium, but recent geostrategic realities in which the United States and like-minded allies contend with a rising great power (China) that seeks to maximize its national advantage within a broader international community, a declining great power (Russia) whose leaders are increasingly defining their legitimacy by opposition to the West, and several outliers (Iran and North Korea).

With that as background, this paper now intends to argue the following propositions:

First, the United States has been advocating what is, in effect, a new set of *peacetime* norms that limit what kind of cyber espionage countries can carry out. By norms, this paper means a set of understandings about what is or is not acceptable state behavior, whether or not such norms are codified by treaty.

Second, there may be other activities which may be deemed unacceptable, notably those that would help protect critical infrastructure.

Third, the United States (and its friends) should insist that such norms be shaped in terms of what countries do with the information they capture, rather than from which systems they are looking for such information, even if one is often a proxy for the other.

Fourth, while determining that state behavior violates norms is challenging if not impossible, several approaches are possible.

Although this essay is written from a US perspective (as per the author's experience), many of its basic assumptions, notably the distinctions between acceptable and unacceptable cyber espionage, are shared to a great extent among the other Five Eyes countries (Canada, the UK, Australia, New Zealand) and to a considerable extent by NATO members and other US allies.[1]

Consider, now, the three norms sketched below as exemplars of the shift in what can be considered acceptable state behavior.

2. NORM AGAINST ECONOMICALLY-MOTIVATED CYBER ESPIONAGE

The struggle between the United States and China over cyber espionage norms shows that creating them is no straightforward process. Starting in roughly 2009, US officials (as well those of allied countries) argued that while cyber espionage was acceptable state behavior if carried out to protect national security, it was not acceptable behavior if economically motivated – notably if the results were handed to corporations to give them an unfair trade advantage. As the Director of National Intelligence (DNI) said in 2013:

> What we do not do, as we have said many times, is use our foreign intelligence capabilities to steal the trade secrets of foreign companies on behalf of – or give intelligence we collect to – US companies to enhance their international competitiveness or increase their bottom line (Clapper 2013).

The absence of such a norm created the case for a new one.[2] To win Chinese accession to such a norm, or at least to modify China's behavior, the United States first tried cabinet-level admonitions and then moved onto presidential admonitions (the Sunnylands summit of June 2013), the May 2014 indictments of PLA officers (May 2014), and a March 2015 Executive Order that laid the groundwork for sanctions against companies (White House 2015a). By the summer of 2015 such sanctions had become a real possibility.

One tool the United States did not and could not use was to threaten to act the same way by arguing that if economically-motivated cyber espionage is acceptable, then we will go ahead and do it ourselves. This threat would not have been credible. Not only does the US *Government* lack the apparatus to identify intellectual property in private Chinese corporations that US *firms* might profitably use, it would have been legally problematic to provide the information to one US competitor without providing it to them all. Furthermore, even if the United States did so, the Chinese would still come out well ahead, because US companies, at this stage, have considerably more intellectual property at risk than do Chinese companies; consider, for example, the great disparity between technology licensing revenues accruing to firms headquartered in each country.

[1] There are differences between the US approach and those of many other NATO members, notably over mass surveillance and the right to privacy, but neither issue is discussed in this paper.

[2] The consensus was that the Agreement on Trade-Related Aspects of Intellectual Property Rights amendments to the World Trade Organization, of which China is a member, enjoins signatories against stealing intellectual property, but did not itself provide sufficient support for a norm against economically-motivated cyber espionage.

The norm against cyber espionage was initially framed as a norm against going after certain targets. The US position was essentially that it did not spy on commercial enterprises and therefore it was not hypocritical to expect others to refrain. But the Snowden revelations made it difficult to deny that the United States did spy on commercial companies. Such cyber espionage was undertaken: (1) to look for weaknesses in commercial products, the knowledge of which would facilitate compromising the systems of their customers (who could well be legitimate national-security targets) (Sanger and Perlroth 2014); (2) to help track terrorists who used commercial systems (notably telecommunications) supplied by these companies (Schneier 2014); or (3) to aid the US negotiating position vis-à-vis foreign countries (Romero 2013). So, the US argument switched to: we do not spy on commercial companies *for the purposes of helping US companies compete*.[3] In other words, the US formulation was reframed from enjoining certain targets of cyber espionage to limiting what can be done with the results.

As it was, China never officially argued that economically-motivated cyber espionage was no worse than national-security cyber espionage. Instead it argued, with increasing implausibility, that the United States (or other accusers) had no proof that the Chinese had carried it out (Economist 2013). The February 2013 release of the Mandiant report, however, made that argument difficult to sustain; this was followed by an avalanche of similar revelations by other US cyber security companies. By May 2015, few in Beijing were trying to pretend that the Chinese did not carry out economically-motivated cyber espionage.[4]

Ultimately, the United States succeeded. In September 2015, President Xi Jinping promised that China would neither conduct nor tolerate such cyber espionage. Evidence to date (for instance, from Dilanian 2016) suggests that China has largely stuck by its promise (Nakashima 2016). Indeed, FireEye, the company that bought Mandiant, has reported that its monthly investigations into Chinese cyber espionage for corporate clients had fallen from 35 per month before the agreement to 3-to-10 per month afterwards (Marks 2016). Although the Chinese did react to the agreement by avoiding detection better, tradecraft simply does not improve fast enough to account for such a fall-off.

This introduces the first norm that the United States (successfully) fought for: cyber espionage is acceptable *unless* the results are used to help a country's firms compete.[5]

3. THE CYBERCRIME MARKETS NORM

As irritating as the hack on the US OPM was, in and of itself, it did not violate established cyber espionage norms. The then-Director of National Intelligence and a former head of the

[3] Although foreign officials do not always believe that the United States actually follows such rules, proof that it has not has been difficult to come by.

[4] As my colleague, Scott Harold, and I found when interviewing Chinese academics and government officials for the RAND Publication, *Getting to Yes with China in Cyberspace* (Harold, Libicki and Cevallos 2016).

[5] China privately criticized the US position against intellectual property theft for its hypocrisy; after all, stealing technologies gave a push to US industrial development in the late 18th and early 19th century. Now that the United States is the global leader in licensing technology, it had more to lose than gain from such theft – hence its insistence on such norms. But perhaps the Chinese, themselves, were beginning to recognize that without respecting intellectual property – many thefts of which are internal – their own companies would not develop their own technology only to see it stolen.

Central Intelligence Agency (CIA) both indicated that they would have done the same had the opportunity presented itself.[6]

Nevertheless, the hack raised questions about what the Chinese would do with the information. They could use it themselves, perhaps to help focus their recruitment of US citizens as spies, or they could trade such information with Russia for similar purposes (see, for instance, Gallagher 2015). Both were acceptable acts of state. However, there were also fears that the Chinese would sell such data into the black market allowing criminals to use it for identity theft and other scams. Indeed, OPM reacted to the theft by offering 22 million individuals free access to credit monitoring services (Abel 2015).

To date, there is scant indication that the Chinese have handed such information over to cyber criminals, and no such information has been reported as found being traded within cybercrime markets. But whereas the Chinese government and Chinese cyber criminals are distinct, in other countries the division between national security and crime is gauzier. There is a constant exchange of information and perhaps other resources between the Russia's intelligence agencies and their cyber *mafiya* (see, for instance, Kramer 2017). And if reports are true that the theft of $81 million from the Bangladesh Bank was carried out by the government of the Democratic People's Republic of Korea, then at least one government is no stranger to cybercrime as such (Lyngaas 2016).

Russia may be the test case for such a norm, because of the difficulty that outsiders have in distinguishing the actions of official state intelligence organizations (the FSB and the GRU) from those of cyber criminals. Reportedly, the latter have avoided prosecution and extradition, so long as their victims are not Russian (Schwartz 2016), in part because they have been willing to lend their services to the state (Schwirtz 2017). The indictment of Russian FSB officials for the criminal hack of Yahoo is further indication of such entanglement (Nakashima 2017). In the end, there may be little valid distinction between state-paid cyber espionage in the service of crime and state-condoned hackers carrying out cyber espionage for criminal purposes.

The whole affair introduces a second norm on which US officials may have been prepared to act were their OPM fears founded: cyber espionage is acceptable *unless* the results are converted to criminal purposes.

4. THE NO-POLITICAL-DOXING NORM

The Russian hack of the Democratic National Committee (DNC) coupled with the delivery of such files to WikiLeaks for public posting embroiled the US presidential election and its aftermath. No serious individual in US public life has found this act to be acceptable, even if some argue that the content of the posted material deserved more attention than the question

[6] "Don't blame the Chinese for the OPM hack," former NSA and CIA Director Michael Hayden said, arguing that he "would not have thought twice" about seizing similar information from China if he had the chance (Ferraro 2015). Director of National Intelligence James Clapper echoed the sentiment, saying at a recent conference: "You have to kind of salute the Chinese for what they did. [...] If we had the opportunity to do that [to them], I don't think we'd hesitate for a minute" (Sciutto 2015).

of how it was brought to attention. Many voices (even on the dovish half of the spectrum[7]) have argued in the subsequent months that the US failure to respond would only encourage the Russians to continue, referring to the several key European elections in 2017 subject to similar influence.

Distinctions have been drawn over how the take from the DNC hack can and cannot legitimately be used. John Brennan, the Director of Central Intelligence, has said that while spying on each other's political institutions is fair game, making data public – in true or altered form – to influence an election was a new level of malicious activity, far different from ordinary spy vs. spy maneuvers (Sanger 2016).

So, what exactly is it that is objectionable – and something that the United States, itself, would be willing to forego in the course of persuading others to do likewise?

To be clear, the DNC hack was *not* vote-tampering, even if it raised the odds that Russia, as a rogue player in the international system armed with first-rate hackers, might try its hand at the game. This was a real fear, given the pervasive use of potentially vulnerable electronic voting systems (Barrett 2016). President Obama warned Russia against tampering with voting systems in a private meeting with President Putin in September 2016 (Landler and Sanger 2016). That noted, vote-tampering is a cyber attack (in the sense that it could disrupt the voting practice or corrupt the results), while the DNC hack was an example of cyber espionage. No new norms were needed to warn Russia against vote-tampering. But new norms would be needed to properly condemn what happened to the DNC.

But what norm, exactly, would have been violated by the DNC hack and the DNC's subsequent doxing? Was it that countries should abjure from influencing elections in other countries? This precept is probably not the place to make a stand.[8] Perhaps prudence dictates that the United States not influence an electorate against voting for someone who then goes on to win anyway. Yet if intervention can make a difference, it may be a chance worth taking. President Obama did speak against both Brexit and Scottish independence. Other foreign leaders have expressed opinions about the US Presidential Election of 2016. One might propose a norm based on not breaking the laws of the country holding the election, but laws vary greatly between countries. Some of these laws (e.g. those restricting the freedom of expression) may strike the United States as illegitimate; breaking them is a good not a bad thing. A problem with limiting such a norm to political processes carried out by elections is a pronounced lack of appeal to countries that do not have elections, or those who maintain elections only for the sake of appearances.

[7] "Any response from the Obama administration or the FBI will be viewed through this partisan lens, especially because the president is a Democrat. We need to rise above that. These threats are real and they affect us all, regardless of political affiliation. That this particular attack targeted the DNC is no indication of who the next attack might target. We need to make it clear to the world that we will not accept interference in our political process, whether by foreign countries or lone hackers" (Schneier 2016). "This is not just about the United States, it is not just about Trump or Clinton, or just about American democracy," said Thomas Rid, a professor of security studies at King's College London. "If they consider this a success, they may conclude that, 'Of course, we can do this elsewhere. We can do this again. We can probably also find things, *kompromat*, on the next president'" (Taub 2016).

[8] "On the other hand, the United States has frequently and unapologetically intervened in other countries' elections. In Latin America, the Middle East, or Eastern Europe, this intervention has been open or covert, ranging from funding pro-democracy organizations and providing training to political leaders to handpicking candidates to install in power. In these interventions, the United States certainly uses products of the intelligence agencies" (Gessen 2016).

Perhaps the relevant norm is to not tamper with political processes in general. But the United States and its friends not only hold that elections have a legitimacy other processes may lack, but that their outcomes can be very sensitive to otherwise minor influences. Furthermore, such processes are subject to a myriad of influences. The revelation of the Panama papers – which Russians blame on the United States – shows how events in one country can affect the political processes of many distant countries (e.g., Iceland, Ukraine, Pakistan) (Rutinsky and Arkhipov 2016).[9]

Hence, if one would write a norm that makes the Russian DNC hack unacceptable, it cannot easily rest on a general prohibition against political interference, but against the yoking of a currently accepted practice (cyber espionage) to a problematic practice (unwarranted influence in another country's political processes). In other words, to be on safe ground, such a norm could be about the misuse of cyber espionage. Given the US discomfort with Russian operations on the Soros Foundation (Hattem 2016) or the World Anti-Doping Agency (WADA) (Goodin 2017) that resulted in the public release of thousands of documents, perhaps the norm could be generalized: it is unacceptable for states to acquire materials by cyber espionage and release them to the public for doxing.

But such a norm comes with costs to other US values. What differentiates doxing from whistle-blowing? The Russians could easily argue that activities of the Soros Foundation or WADA[10] which they considered anti-Russian were those that *should* have been brought to light (although inserting fake documents into the mix of released files hardly helps their case). The Snowden revelations – and more recently the Vault7 disclosures from WikiLeaks – are variously put in either basket. Furthermore, is it necessarily in the interests of the United States – which is largely an open society where certain types of disclosure are not only encouraged but mandated (see its Freedom of Information Act) – to press for norms that protect the leaders of closed societies from disclosures, particularly those with high levels of corruption that *ought* to be exposed?[11] The penetration of the *New York Times'* network by the Chinese in late 2012 was carried out because the newspaper had revealed the $2 billion dollar fortune of China's premier (Barboza 2012). In a slightly different universe where it was US government-conducted cyber espionage rather than journalistic "shoe-leather" that brought the information to the *New York Times*, which government would have been the more irresponsible: the United States for carrying out cyber espionage to gather the information, or the Chinese for carrying out cyber espionage to figure out where the information came from? Could a norm in which governments collectively pledge to keep everyone else's secrets be viewed as an international "conspiracy" to permit government corruption? So, while a norm against using cyber espionage to support doxing may nevertheless be worthwhile, it does have to be written carefully.

Many countries, especially those with limited cyber espionage capabilities, may well sign up to such a norm. Agreement may be possible even from China, whose willingness to curb economic cyber espionage suggests that they see valid limits to stealing information. Their assent is more likely if China gets to help write the norm – one of China's objections to the Budapest Convention is that it had no say in its drafting.

[9] Perhaps the DNC hack was Putin's revenge for doing so (see Hamburger and Nakashima 2016; Golodryga 2016).

[10] But the Russian tune has changed (see Ruiz 2016).

[11] See, for instance, Thomas Friedman's fantasy of the CIA exposing Putin's corruptly-acquired billions (Friedman 2016).

Getting the assent of Russia, whose recent behavior is what has spurred such consideration, will be a major hurdle, unless its leadership signs up in the blithe belief that it can still do what it wants as long as it can deny having done so. Or Russia may realize it has more to fear from an aggressive use of cyber espionage-plus-doxing than it has to gain by doing it to others.[12] Russia, after all, is a country in which corruption and censorship are rife; for instance, its 2014 blogger law (Birmbaum 2014) is increasingly relied upon to forestall stories that the state does not want to see. Cyber espionage can reveal the former, and other cyber tricks can be used to move information into the flow of news accessible to Russians. Russia also has a long history of using *kompromat* to discredit (political) enemies. Because the legitimacy of Russia's government rests on popular approval of its leaders rather than popular approval of the process by which leaders are selected (a role, for instance, played by the US Constitution), it is far more open to question.

In any case, the US reaction to the DNC hack introduces a third norm: cyber espionage is acceptable *unless* the results are used publicly for political influence operations.

Put all three norms together, and the United States is edging towards a norms regime that allows countries to carry out cyber espionage *as long as* the results are used in a "professional" manner: that is, to foster a country's national security by influencing the decisions that the governments which collected the information make and facilitating their ability to carry them out. The results have to be kept in-house (or shared with allies to keep in-house), as US intelligence agencies do. They cannot be provided to commercial enterprises, criminals, or to the public.

But would the effort to sanction certain uses of cyber espionage be limited to those three categories? Perhaps not.

5. PEACETIME CYBER ESPIONAGE AGAINST CRITICAL INFRASTRUCTURE

Michael Hayden, formerly the NSA's director, said, "ideas have been raised about forming the cyber equivalent of demilitarized zones for sensitive networks, such as the power grid and financial networks, that would be off-limits to attack from nation states" (Zetter 2010). There are indications from the Chinese that they would be receptive to such a deal using the non-aggression pact in cyberspace that Russia and China inked in 2015 as precedent (see Ostroukh and Lyngaas 2015). Indeed, in late 2015, the UN Group of Governmental Experts (to which China belongs) agreed: "a State should not conduct or knowingly support ICT activity that intentionally damages or otherwise impairs the use and operation of critical infrastructure" (United Nations 2015; see also Grigsby 2015). Unfortunately, as with any agreement not to engage in certain activities characteristic of war (as a serious attack on infrastructure could be considered), enforcement by punishment is unlikely to take place while war or warlike activities

[12] And from the US perspective, Russia has made enough enemies to lift the burden of cyber espionage from its own shoulders. "Ukrainian hackers behind recent Kremlin email leaks said on Thursday they planned to release more information taken from accounts linked to senior Russian officials, including to President Vladimir Putin's chief spokesman. A network of Ukrainian hacking groups, called the Cyber Alliance, has been releasing emails they say were sent to one of Putin's top advisers - a bid to disprove Russia's denial it has stoked separatism in eastern Ukraine and played a direct role in the 2-1/2-year-old conflict there" (Prentice and Chornokondratenko 2016).

ensue. At that point, treaties would be less compelling and considerations of escalation control (if we do X, they might do X or even Y) would likely be the more relevant influence on each combatant's actions. Thus, the rules provide little real inhibition to bad behavior.

Yet, because cyber espionage is almost always a pre-requisite for cyber attack, particularly if implants are used, making it difficult to carry out *peacetime* cyber espionage can retard or inhibit wartime or warlike cyber attacks, at least in the first few weeks and months of conflict (that is, until new entryways into the target systems are found or developed). Even better, countries that abjured cyber espionage would have a hard time coercing others by threatening immediate cyber attacks on critical infrastructure without at the same time admitting that their word not to carry out (prefatory) cyber espionage on critical infrastructure was worthless. Thus, a serious norm (in the sense of a norm whose violation brings serious repercussions) prohibiting attacks on critical infrastructure or threats thereof could require banning cyber espionage on each other's infrastructure. This linkage is understood in both Washington DC and Beijing.

Because the aim is to ban activity *before* its consequences are manifest, this norm agreement requires some mechanism to find violations: to attribute as well as detect system intrusions, as well as to distinguish deliberate infection from random malware drift. Any such monitoring mechanism must pass three tests. One is getting attribution right. Two, more importantly, is to build a good case for what happened so that it can convince skeptics. Three, most importantly, is to present the case so that the accused accepts the results as fairly derived and not arbitrary (the process of attribution need not be resolved in every case as long as it resolves often enough to inhibit cheating[13]). Similarly, the process cannot be so stringent that victims of cyber espionage, who may have access to information that they will not or cannot pass forward, conclude that they need other ways to press the point with the accused.

Part of the political problem of enforcement from the perspective of China (and perhaps also Russia) is that the United States catches a much greater share of others' spying than the others catch of US government spying. Consider that all of the Snowden revelations about overseas cyber espionage were new, in that there was no example that *confirmed* a discovery of a specific cyber espionage case linked to the United States.[14] Putatively, any process that produces similar results may be viewed as deeply biased *even if accurate*. China's ability to detect and attribute cyber espionage from the United States, for instance, is far lower than the US ability to detect and attribute cyber espionage from China; there has simply been no equivalent of the 2013 Mandiant report. This arises from three differences. First, China's operational security lags behind US operational security, making Chinese spying easier to detect.[15] Second, China's ability to detect intrusions (especially from the US Government) lags behind the US ability

[13] Not all attribution evidence is publicly releasable (see Sanger and Fackler 2015). This suggests an unbridgeable difference between the confidence that US officials place in attribution and the confidence felt by a fair-minded individual working from open sources, but unwilling to take the word of US sources at face value.

[14] Since Snowden, one cyber security firm, Kaspersky, has uncovered what it called 'Regin' malware possibly linked to the United States (Kaspersky 2014) and the work of the Equation Group, probably linked to the United States (Kaspersky 2015).

[15] From Segal (2016, p. 113): "Security analysts consider the Chinese particularly noisy in networks, especially compared to the Russians". Part of how one can tell good operational security is to see which intrusion sets attributed to each country have gone undetected for more than, say, five, years before being eventually detected.

to detect intrusions. Third, China's ability to attribute detected intrusions lags behind the US ability[16] to attribute detected intrusions.

As long as all three are true, will the Chinese (or, correspondingly, the Russians or Iranians) accept that compliance verification would be even-handed? Until the Chinese and others gain confidence in their own attribution capabilities, they may not even believe that US attribution capabilities are good enough; those caught spying may believe that they have been fairly caught but unless they tell others, China's policy-making community may retain their skepticism.

How might an attribution process merit trust? One could start with a multi-national body which, like the US National Transportation Safety Board (NTSB), focuses on characterizing a cyber espionage hack rather than assigning guilt for it. At the very least, such a body may permit examining evidence that a system's penetration was accidental (e.g., malware found in one system could have drifted over from another) or, if purposeful, may not have had prohibited intent. The International Atomic Energy Agency (IAEA) or the International Civil Aviation Organization (ICAO) may provide other models that can be used to build multi-lateral and highly technical examinations of norms violation. Microsoft has suggested that such a group be characterized by strong technical expertise, diverse geographic representation and peer review and the group should only undertake analyses for significant events (Charney et al. 2016). Since most cyberspace forensic authorities spend much of their career with their respective governments, it may be a challenge for professionals of one side to trust their foreign counterparts. Those from the West may consider Chinese or Russian representatives beholden to their governments and thus unlikely to be given enough latitude to offer an independent perspective. The Russians and the Chinese may, in turn, argue that US experts are comparably beholden themselves *and* would also have so much more background at making attribution as to reduce their own experts to spectators. Yet there are grounds for believing that working together for long enough can alleviate much of the unwarranted mistrust.

Conversely, might the Chinese and others be more forthcoming if they understood the mechanisms of attribution better?[17] If so, would the prospect of winning an agreement from the Chinese and others on norms be sufficiently attractive to justify the US teaching (or sending third parties to teach) others *some* of its forensic attribution techniques? Would the Chinese and others then be willing to credit such techniques as evidence of verification? An ancillary benefit is that stronger Chinese attribution capabilities could reduce the chances of a catalytic conflict in which China is attacked by someone masquerading as a US source. At first glance, this notion appears untenable: countries do not teach others such technology. Yet the United States encourages other countries to adopt permissive action links (PALs) for their nuclear weapons so that such weapons are not used accidentally or at the instigation of rogue nuclear warriors. There are also offsetting benefits when foes build enough surveillance capability to let them

[16] In the United States, a large share of detection and intrusions are carried out by private companies (many of whom employ former NSA employees). China is only starting to develop its own cyber security companies. That noted, it can buy cyber security expertise: even if some US companies might refuse Chinese business, cyber security companies from beyond the United States (e.g., Israel) are for hire.

[17] Richard Bejtlich (2015) has testified, "When either one, or both, opponents possess low attribution capabilities, it is a less stable situation. This could be a problem with the agreement between China and the United States. Private and public teams in the United States can perform high levels of attribution on Chinese activity. Private and public teams in China do not share the same capabilities at present. China could therefore suspect that the United States is behind certain hacks, although such activity could be caused by Russia or other actors. This is one reason to welcome the rise of private or nongovernment security companies in China, who may improve the country's attribution capabilities."

believe that the United States is adhering to arms deals in much the same way that aerial and later space surveillance technology assured the United States that it had little need to engage in a missile race with Russia (circa 1960). Furthermore, helping bring Chinese attribution capabilities closer to those available in the United States does not mean that the United States should be expected to teach others how to *detect* cyber espionage intrusions or how to keep their own penetrations from being *detected* by the United States.

6. PROHIBITING UNWELCOME USES OF STOLEN INFORMATION MAY MAKE CERTAIN TARGETS OFF-LIMITS

There is no clear line between prohibiting certain *targets* of cyber espionage and prohibiting specific *uses* of cyber espionage. Take the norm against industrial cyber espionage. Notionally, it would allow cyber espionage against companies for, say, purposes such as evaluating another country's capacity for developing dual-use technology (a national security rationale). In practice, it puts certain targets off-limits, because it creates a rebuttable presumption that the only good reason to spy on, say, automobile companies is to help someone compete in the automobile market. By contrast, US irritation at the threat that OPM data might find itself in criminal markets or the transfer of DNC data to WikiLeaks is *only* about the use of such data.

For some of the several potential norms discussed above, non-intelligence uses of captured data are strongly implied by the choice of targets: e.g., private corporations, or some critical infrastructure. Here, cyber espionage is indistinguishable from pre-cyber attack (and pre-coercion) preparations. As for critical infrastructure, while there seems little good reason to spy on electric grids, compromising communications nodes is almost *sine qua non* for wide-scale surveillance, while compromising financial systems helps in tracking financial crimes carried on with the connivance of unscrupulous bankers.

Nevertheless, from the perspective of the United States and its friends, the best strategy may be to insist that all prohibitions relate to post-cyber espionage uses of data rather than written as target-specific. This would foster the argument that such a form of professionalism differentiates cyber espionage carried out by the United States and its allies from the sort that should be prohibited. More importantly, if the collection of norms turns out to be a package deal, such a stance could inhibit this package from becoming a collection of unrelated items. Again, it would be understood, even if not stated outright, that evidence of cyber espionage within certain systems is a *prima facie* case of cyber espionage for the wrong reasons. So, in practice, there would essentially be a prohibition on spying on certain classes of target. But policy coherence helps in making the case for a broad set of prohibitions.

7. MAKING NORMS HAPPEN

Norms-setting is a deliberative process, but not necessarily multilateral. The United States, after all, could declare a set of red lines, call them norms if it pleases, and then warn others against violating them lest they face punishment. Or, it could take the trouble to work out norms with other countries, which would be accepted as norms by others, not only because the United States says so but because other countries agree they merit approval and they have had a hand in the process by which norms were generated and agreed to.

Working from red lines and unilaterally calling them norms means that the United States recognizes the norms it wants and only those norms; no concessions are needed. Working with others offers two advantages, though. First, it forces the advocates of norms to be explicit about what behavior is proscribed, and why. Second, any result is more likely to get buy-in from other countries; the same benefit applies to getting buy-in for a US response to a violation of such norms. Conversely, the negotiations path is slower to finish than the unilateral route (which, in turn, is slower than the after-the-fact reaction route). And even negotiations exclusively with US friends may force the United States to make compromises over what is unacceptable behavior (European countries, for instance, tend to favor stronger privacy protections than those of US law) and perhaps even what criteria are used to determine that the behavior of a given country is unacceptable.

The process may not necessarily generate norms all of which the United States would feel totally comfortable with – although the package, on the whole, very well might (otherwise they will not gain US assent). The United States is likely to fight back against any package that goes beyond cyber espionage norms to include demands from other countries (e.g., on Internet governance or legitimizing censorship to "protect" cyberspace). The prospect that such a package is focused on cyber espionage alone rests on a reasonable presumption that some countries see value in constraining US cyber espionage operations, in large part because they fear that the United States is very good at carrying them out.

Finally, as far as norms without monitoring (let along consequences) are but sweet sentiments, the case for cyberspace norms has always been fraught (Roth 2016), although attribution has got better (Panetta 2012; Rid 2014 has a nice explanation of some of the techniques used). The Xi-Obama agreements appear to have worked so far without any formal compliance mechanism,[18] although as a rule multilateral agreements may require more compliance mechanisms than bilateral ones, in part because one side can quit the agreement if it deems it being violated by the other side, but in a multilateral agreement the party that leaves may be hurting itself more than it hurts the cheater. Yet foreign governments, which generally have weaker attribution capabilities than the United States does, may fear signing up to a cyber espionage norm that the United States, but only the United States, can police. In some cases, the *use* of the information gained from cyber espionage – e.g., a threat made against critical infrastructure – can provide sufficient evidence. Conversely, if attribution for cyber espionage is adequate, determining that data has been misused is often straightforward. For instance, if Chinese culpability for the OPM hack is certain, then finding this data on the black market means that China must have put it there, or managed it in ways that allowed it to get there. Similarly, if one knows for certain that Russia took the DNC's e-mails, then similar conclusions can be reached by observing that these e-mails ended up on WikiLeaks *even if* a transmission chain cannot be proven.

[18] This is the consensus of many observers (see, for instance, Lynch 2016).

By focusing on what uses can properly be made of information acquired by cyber espionage, the United States and its friends can make progress on the challenging issues of developing a favorable package of norms and finding ways to monitor compliance.

8. CONCLUSIONS

The premise that cyber espionage (like physical espionage) is acceptable state activity has become increasingly untenable. The United States has won support for norms against economically-motivated cyber espionage, pushed back against the use of cyber espionage for political doxing, and was prepared to condemn the transfer of information from state-sponsored cyber espionage into cybercrime markets. These three examples suggest that countries look seriously at what sorts of cyber espionage should and should not be deemed acceptable. Further norms are possible, not least of which are those against cyber espionage for the purpose of implanting attacks into critical infrastructure, although the challenge of determining the *purpose* of an implant only adds to the challenges of determining who put them there.

This paper has suggested a normative framework in which cyber espionage is *unacceptable* unless the results are used *only* to inform national-security decision-making. Left open with this formulation is whether cyber espionage can be used to open the door for cyber attacks (e.g., via implants) – particularly cyber attacks on critical infrastructure. Further research can be used to refine the definitions of norms, explicate the interests of the various stakeholders in the norms formation process, and develop monitoring methods that can meet the twin tests of being accurate and accepted.

REFERENCES

Abel, Jennifer, “OPM hack fallout: feds pay $133 million for (largely useless) ID theft monitoring services,” *Consumer Affairs*, September 9, 2015. https://www.consumeraffairs.com/news/opm-hack-fallout-feds-pay-133-million-for-largely-useless-id-theft-monitoring-services-090915.html.

Barboza, David, “Billions in Hidden Riches for Family of Chinese Leader,” *New York Times*, October 25, 2012. http://www.nytimes.com/2012/10/26/busainess/global/family-of-wen-jiabao-holds-a-hidden-fortune-in-china.html.

Barrett, Brian, “America’s Electronic Voting Machines are Scarily Easy Targets,” *Wired*, August 2, 2016. https://www.wired.com/2016/08/americas-voting-machines-arent-ready-election/.

Bejtlich, Richard, “Outside perspectives on the Department of Defense cyber strategy,” Testimony before the US House of Representatives Committee on Armed Services on September 29, 2015.

Birmbaum, Michael, “Russian blogger law puts new restrictions on Internet freedoms,” *Washington Post*, July 31, 2014. https://www.washingtonpost.com/world/russian-blogger-law-puts-new-restrictions-on-internet-freedoms/2014/07/31/42a05924-a931-459f-acd2-6d08598c375b_story.html?utm_term=.c2501c618038.

Charney, Scott, Erin English, Aaron Kleiner, Angela McKay, Nemenja Malisevic, Jan Neutze, and Paul Nicholas, 2017. *From Articulation to Implementation: Enabling Progress on Cybersecurity Norms*. https://mscorpmedia.azureedge.net/mscorpmedia/2016/06/Microsoft-Cybersecurity-Norms_vFinal.pdf.

Clapper, James R., Statement by Director of National Intelligence James R. Clapper on Allegations of Economic Espionage, September 8, 2013. https://www.dni.gov/index.php/newsroom/press-releases/191-press-releases-2013/926-statement-by-director-of-national-intelligence-james-r-clapper-on-allegations-of-economicespionage.

Detrow, Scott, "Obama on Russian Hacking: 'We Need to Take Action. And We Will'," *NPR News*. December 15, 2016. http://www.npr.org/2016/12/15/505775550/obama-on-russian-hacking-we-need-to-take-action-and-we-will.

Dilanian, Ken, "Russia May Be Hacking Us More, But China Is Hacking Us Much Less," *NBC News*, October 12, 2016. http://www.nbcnews.com/news/us-news/russia-may-be-hacking-us-more-china-hacking-us-much-n664836.

Economist, "Admit nothing and deny everything," June 8, 2013. http://www.economist.com/news/china/21579044-barack-obama-says-he-ready-talk-xi-jinping-about-chinese-cyberattacks-makes-one.

Ferraro, Matthew F., "On the OPM Hack, Don't Let China Off the Hook," *The Diplomat*, July 14, 2015. http://thediplomat.com/2015/07/on-the-opm-hack-dont-let-china-off-the-hook/.

Friedman, Thomas, "Let's Get Putin's Attention," *New York Times*, October 5, 2016. http://www.nytimes.com/2016/10/05/opinion/lets-get-putins-attention.html.

Gallagher, Sean, "China and Russia cross-referencing OPM data, other hacks to out US spies," *Ars Technica*, August 31, 2015. http://arstechnica.com/security/2015/08/china-and-russia-cross-referencing-opm-data-other-hacks-to-out-us-spies/.

Gessen, Masha, "Arguing the Truth with Trump and Putin," *New York Times*, December 17, 2016. http://www.nytimes.com/2016/12/17/opinion/sunday/arguingthetruthwithtrumpandputin.html.

Goodin, Dan, "US athletes' doping tests published by Russian hackers, agency says," *Ars Technica*, September 13, 2016. http://arstechnica.com/security/2016/09/anti-doping-agency-pins-leak-of-us-gold-medalists-data-on-russian-hackers/.

Golodryga, Bianna. "3 Major Implications of the Panama Papers Leak," *Huffington Post*, April 21, 2016. http://www.huffingtonpost.com/bianna-golodryga/3-major-implications-of-t_b_9748512.html.

Grigsby, Alex, "The 2015 GGE Report: Breaking New Ground, Ever So Slowly," *Council on Foreign Relations Guest Blog*, September 8, 2015. http://blogs.cfr.org/cyber /2015/09/08/the-2015-gge-report-breaking-new-ground-ever-so-slowly/.

Hamburger, Tom and Ellen Nakashima, "Clinton campaign — and some cyber experts — say Russia is behind email release," *Washington Post*, July 24, 2016. https://www.washingtonpost.com/politics/clinton-campaign--and-some-cyber-experts--say-russia-is-behind-email-release/2016/07/24/5b5428e6-51a8-11e6-bbf5-957ad17b4385_story.html.

Harold, Scott, Martin Libicki, and Astrid Cevallos, 2016, *Getting to Yes with China in Cyberspace*, Santa Monica CA (RAND).

Hattem, Julian, "Thousands of Soros docs released by alleged Russian-backed hackers," *The Hill*, August 15, 2016. http://thehill.com/policy/national-security/291486-thousands-of-soros-docs-released-by-alleged-russia-backed-hackers.

Haynes, Deborah, "Russia waging cyberwar against Britain," *The Times*, December 17, 2016. http://www.thetimes.co.uk/edition/news/russiathreattobritaingpd98bz83.

Kaspersky Lab, "Regin: a malicious platform capable of spying on GSM networks," November 24, 2014. http://www.kaspersky.com/about/news/virus/2014/Regin-a-malicious-platform-capable-of-spying-on-GSM-networks.

Kaspersky Lab, “Equation Group: The Crown Creator of Cyber-Espionage,” February 16, 2015. http://www.kaspersky.com/about/news/virus/2015/equation-group-the-crown-creator-of-cyber espionage.

Kramer, Andrew, “Top Russian Cybercrimes Agent Arrested on Charges of Treason,” *New York Times*, January 25, 2017. https://www.nytimes.com/2017/01/25/world/europe/sergei-mikhailov-russian-cybercrimes-agent-arrested.html.

Landler, Mark, and David Sanger, “Obama Says He Told Putin: ‘Cut It Out’ on Hacking,” *New York Times*, December 26, 2016. https://www.nytimes.com/2016/12/16/us/politics/obama-putin-hacking-news-conference.html.

Lynch, David, and Geoff Dyer, “Chinese Hacking of US Companies Declines,” *Financial Times*, April 13, 2016. http://www.ft.com/cms/s/0/d81e30de-00e4-11e6-99cb-83242733f755.html.

Lyngaas, Sean, “Debating the Sino-Russian cyber pact,” *Federal Computer Week*, May 12, 2015. http://fcw.com/articles/2015/05/12/russian-chinese-cyber.aspx.

Lyngaas, Sean, “Symantec traces Swift banking hacks to North Korea,” *Federal Computer Week*, May 31, 2016. https://fcw.com/articles/2016/05/31/swift-hack-dprk.aspx.

Mandiant, APT1, *Exposing One of China’s Cyber Espionage Units*, March 2013. sintelreport.mandiant.com/Mandiant_APT1_Report.pdf.

Marks, Joseph, “US-China Cyber Dialogue to Continue Under Trump,” *Nextgov*, December 9, 2016. http://www.nextgov.com/cybersecurity/2016/12/us-china-cyber-dialogue-continue-under-trump/133782/.

Nakashima, Ellen, “Treasury and Justice officials pushed for economic sanctions on China over commercial cybertheft,” *Washington Post*, December 27, 2016. https://www.washingtonpost.com/world/national-security/2016/12/27/fc93ae12-c925-11e6-8bee-54e800ef2a63_story.html.

Nakashima, Ellen, “Justice Department charging Russian spies and criminal hackers in Yahoo intrusion,” *Washington Post*, March 15, 2017. https://www.washingtonpost.com/world/national-security/justice-department-charging-russian-spies-and-criminal-hackers-for-yahoo-intrusion/2017/03/15/64b98e32-0911-11e7-93dc-00f9bdd74ed1_story.html.

Ostroukh, Andrey. “Russia, China Forge Closer Ties With New Economic, Financing Accords: Moscow turns to Asian investors to reduce reliance on Europe and the U.S. amid standoff over Ukraine,” *Wall Street Journal*, May 8, 2015. http://www.wsj.com/articles/russia-china-forge-closer-ties-with-new-economic-financing-accords-1431099095.

Panetta, Leon E., “Remarks by Secretary Panetta on Cybersecurity to the Business Executives for National Security, New York City,” October 11, 2012; http://archive.defense.gov/transcripts/transcript.aspx?transcriptid=5136.

Poplin, Cody, “Cyber Sections of the Latest G20 Leaders’ Communiqué,” *Lawfare Blog*, November 17, 2015. https://www.lawfareblog.com/cyber -sections-latest-g20-leaders-communiqué.

Prentice, Alessandra and Margaryta Chornokondratenko, “Ukrainian Hackers Promise Leaks on Putin Spokesman,” *Reuters*, November 4, 2016. http://in.reuters.com/article/ukraine-crisis-cyber -russia-idINKBN12Y2P5.

Rid, Tomas and Ben Buchanan, “Attributing Cyber Attacks,” *Journal of Strategic Studies*, 2014.

Romero, Simon, “N.S.A. Spied on Brazilian Oil Company, Report Says,” *New York Times*, September 9, 2013. http://www.nytimes.com/2013/09/09/world/americas/nsa-spied-on-brazilian-oil-company-report-says.html.

Roth, Andrew, “How the Kremlin is sure to keep its fingerprints off any cyberattack,” *Washington Post*, August 2, 2016. https://www.washingtonpost.com/world/europe/how-the-kremlin-is-sure-to-keep-its-fingerprints-off-any-cyberattack/2016/08/02/26144a76-5829-11e6-8b48-0cb344221131_story.html.

Ruiz, Rebecca, “Russians No Longer Dispute Olympic Doping Operation,” *New York Times*, December 27, 2016. http://www.nytimes.com/2016/12/27/sports/olympics/russia-doping.html.

Rudnitsky, Jake and Ilya Arkhipov, “Putin Sees US, Goldman Sachs Behind Leak of Panama Papers,” *Bloomberg.com*, April 14, 2016. https://www.bloomberg.com/news/articles/2016-04-14/putin-sees-u-s-goldman-sachs-behind-leak-of-panama-papers.

Sanger, David, “U.S. wrestles with how to fight back against cyberattacks,” *New York Times*, July 31, 2016. http://www.nytimes.com/2016/07/31/us/politics/us-wrestles-with-how-to-fight-back-against-cyber attacks.html.

Sanger, David and Martin Fackler, “N.S.A. Breached North Korean Networks Before Sony Attack, Officials Say,” *New York Times*, January 19, 2015. http://www.nytimes.com/2015/01/19/world/asia/nsa-tapped-into-north-korean-networks-before-sony-attack-officials-say.html.

Sanger, David and Nicole Perlroth, “N.S.A. Breached Chinese Servers Seen as Security Threat,” *New York Times*, March 22, 2014. http://www.nytimes.com/2014/03/23/world/asia/nsa-breached-chinese-servers-seen-as-spy-peril.html.

Schneier, Bruce, “NSA Hacking of Cell Phone Networks,” *Lawfare Blog*, December 8, 2014. https://www.lawfareblog.com/nsa-hacking-cell-phone-networks.

Schneier, Bruce, “Hacking the Vote,” *Schneier on Security*, August 1, 2016. https://www.schneier.com/blog/archives/2016/08/hacking_the_vot.html.

Schwartz, Matthew, “Russia: 7-Year Sentence for Blackhole Mastermind (Jail Time for Russian Cybercriminals is Rare),” Bankinfosecurity.com, April 15, 2016. http://www.bankinfosecurity.com/notorious-blackhole-exploit-kit-author-sentenced-a-9048.

Schwirtz, Michael and Joseph Goldstein, “Russian Espionage Piggybacks on a Cybercriminal’s Hacking,” *New York Times*, March 12, 2017. https://www.nytimes.com/2017/03/12/world/europe/russia-hacker-evgeniy-bogachev.html.

Sciutto, Jim, “Director of National Intelligence blames China for OPM hack,” *CNN*, June 25, 2015. http://www.cnn.com/2015/06/25/politics/james-clapper-china-opm-hacking/.

Segal, Adam, *The Hacked World Order*, 2016. New York NY (Public Affairs).

Taub, Amanda, “D.N.C. Hack Raises a Frightening Question: What’s Next?” *New York Times*, July 29, 2016. http://www.nytimes.com/2016/07/30/world/europe/dnc-hack-russia.html.

United Nations, General Assembly, A/68/98, *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, http://www.un.org/ga/search/view_doc.asp?symbol=A/68/98.

United Nations, General Assembly, A/70/174, *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, July 22 2015 p. 2. http://www.un.org/ga/search/view_doc.asp?symbol=A/70/174.

White House (2015a), Executive order, April 2, 2015. https://www.whitehouse.gov/the-press-office/2015/04/01/executive-order-blocking-property-certain-persons-engaging-significant-m.

White House (2015b), “Fact Sheet: President Xi Jinping’s State Visit to the United States,” September 25, 2015. https://obamawhitehouse.archives.gov/the-press-office/2015/09/25/fact-sheet-president-xi-jinpings-state-visit-united-states.

Williams, Katie Bo, “Dems urge Obama to release info on Russian links to DNC hack,” *The Hill*, July 27 2016. http://thehill.com/policy/national-security/289485-intel-dems-urge-obama-to-release-info-on-russian-involvement-in-dnc.

Zetter, Kim. “Former NSA Director: Countries Spewing Cyberattacks Should Be Held Responsible,” *Wired*, July 29, 2010. http://www.wired.com/2010/07/hayden-at-blackhat/.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Organisational Integration of Offensive Cyber Capabilities: A Primer on the Benefits and Risks

Max Smeets
Department of Politics and International Relations
University of Oxford
Oxford, UK
max.smeets@politics.ox.ac.uk

Abstract: Organisational Integration has become a key agenda point for policy-makers as governments continue to change and create new organisations to address the cyber threat. Passing references on this topic, however, far outnumber systematic treatments. The aim of this paper is to investigate the potential effects of organisational integration of offensive cyber capabilities (OIOCC). I argue that OIOCC may lead to three key benefits: enhanced interaction efficiency, greater knowledge transfer and improved resource allocation. There are, however, several negative effects of integration, which have so far received little attention. OIOCC may lead to an intensification of the cyber security dilemma, increase costs overall, and impel 'cyber mission creep'. Though the benefits seem to outweigh the risks, I note that ignoring the potential negative effects may be dangerous, as activity is more likely to go beyond the foreign-policy goals of governments and intrusions are more likely to trigger a disproportionate response by the defender.

Keywords: *organisational integration, offensive cyber capabilities, cyber weapons*

1. INTRODUCTION

The principle of organisation integration (OI) is commonly examined in relation to firms' performance.[1] OI is perceived to be a form of organisational innovation, with its potential value on a par with technological innovation.[2] It is considered to be an essential means for firms to remain competitive in the market. Governments often seek OI too, as a way to reduce costs or provide services more effectively. Extending Kenneth Waltz's famous analogy between the market economy and the international state system, one might even say that integration helps states to enhance their relative power in the international system and ensure survival.[3]

OI is also a key agenda point for senior policy-makers seeking to find effective ways to address the cyber threat.[4] The institutional landscape has been shaken up by the new cyber security challenges that countries face. South Korea, for example, has a National Cybersecurity Centre which leads investigations into cyber security incidents. It also has a National Cyber Threat Joint Response Team, comprised of actors from the military, civilian and private sectors, which provides assistance during a crisis. South Korea's CERT manages cyber incidents (there is also a private CERT called CONCERT). And within the Ministry of National Defence it has established a Cyber Command.[5] Similarly, the Netherlands has established various organisations like the National Cyber Security Centre, the Cyber Security Council, the Defence Cyber Education and Training Centre, and the Defence Cyber Command.[6] It has also expanded the missions of several organisations including the National Coordinator for Counter Terrorism and Security, and the Ministry of Security and Justice.

In most countries, the creation and reorientation of institutions dealing with cyber security is ongoing and occurs in parallel to a range of other initiatives such as strategy formulation, regulation, and the creation of informal partnerships. To 'defend the core' effectively requires

[1] It occupies a central place in several bodies of literature, including organisational theory, management, information systems, and organisational strategy. Jay Barney, 'Firm resources and sustained competitive advantage', *Journal of Management*, 17:1 (1991), 99-120; Ricardo Chalmeta, Christina Campos, Reyes Grangel, 'Reference architectures for enterprise integration', *The Journal of Systems Software*, 57 (2001)175-191; John Ettlie and Ernesto M. Reza, "Organizational Integration and Process Innovation', *Academy of Management Journal*, 35:4 (2001), 795-827; Gregory E. Truman, 'Integration in electronic exchange environments', *Journal of Management Information Systems*, 17:1 (2000), 209-244.

[2] Ettlie and Reza, 'Organizational Integration and Process Innovation'.

[3] Kenneth Waltz, *A Theory of International Politics*, (New York: McGraw-Hill: 1979).

[4] See, for example, the most recent national cyber security strategy of the United Kingdom: 'UK Government, National Cyber Security Strategy 2016-2021', (2016). Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf; or Australia: Chris Brookes, 'Cyber Security: Time for an integrated whole-of-nation approach in Australia', Centre for Defense and Strategic Studies, (March 2015). Retrieved from: http://www.defence.gov.au/ADC/Publications/IndoPac/150327%20Brookes%20IPS%20paper%20-%20cyber%20(PDF%20final).pdf.

[5] This is not an inclusive list of institutions. For an overview, see: International Telecommunication Union, 'Cyberwellness Profile Republic of Korea', (December 1, 2014). Retrieved from: https://www.itu.int/en/ITU-D/Cybersecurity/Documents/Country_Profiles/Korea.pdf; James Andrew Lewis, 'Advanced Experiences in Cybersecurity Policies and Practices: An Overview of Estonia, Israel, South Korea, and the United States', Inter-American Development Bank, Discussion Paper IDB-DP-457, (2016, July). Retrieved from: https://publications.iadb.org/bitstream/handle/11319/7759/Advanced-Experiences-in-Cybersecurity-Policies-and-Practices-An-Overview-of-Estonia-Israel-South-Korea-and-the-United%20States.pdf?sequence=.

[6] National Cyber Security Centrum, 'Cybersecuritybeeld Nederland CSBN 2016', (2016); The Netherlands Ministry of Justice, 'De Nationale Cybersecurity Strategie 2: Van bewust naar bekwaam', (2013, October); The Netherlands Ministry of Defence, 'Cyber Command'. Retrieved from: https://www.defensie.nl/english/topics/cyber-security/contents/cyber-command.

awareness about how these organisations and activities can be balanced, coordinated and combined. In other words, it requires an understanding of OI in relation to cyber security.

However, passing references far outnumber systematic treatments of this issue. We know that organisational design and internal politics play a central role in the use of military capabilities, yet organisational analysis is a comparatively underdeveloped aspect of the study of the use of cyber capabilities. Scholars who set out to explain the use of cyber capabilities normally arrive at arguments that focus on the 'nature' or 'meaning' of cyberspace. Yet we cannot fully understand the use of cyber capabilities without studying the organisational structure in which its use of these capabilities is embedded. Through this analysis, we acknowledge that organisational structure can potentially shape the use of offensive cyber capabilities.

I therefore seek to address the following question: *what are the potential effects of organisational integration of offensive cyber capabilities (OIOCC)?* I argue that OIOCC may lead to three key benefits: enhanced interaction efficiency, greater knowledge transfer, and improved resource allocation. There are however several negative effects of integration, which have so far received little attention. OIOCC may lead to an intensification of the cyber security dilemma,[7] increase costs overall, and impel 'cyber mission creep'. Though the benefits seem to outweigh the risks, I note that ignoring the potential negative effects may be dangerous, as actors are more inclined to go beyond the foreign-policy goals of governments and intrusions are more likely to trigger a disproportionate response by the defender.

OI in relation to cyber security occurs at different levels and for different purposes. Table 1 provides a basic overview of the different types of OI. First, the table distinguishes between defensive and offensive organisational activities. Second, it distinguishes between three levels of integration: the highest level refers to the integration of cyber activities between the government and other entities such as the private sector; mid-level OI refers to integration between government organisations of which some do not (initially) focus on cyber activities; and the lowest level considers organisational integration between organisations which focus on cyber activities.[8]

The aim of this paper is to investigate the lowest level of OI in relation to the development of offensive cyber capabilities. According to NATO's Deputy Assistant Secretary General for Emerging Security Challenges, Jamie Shea, 'about 100 countries in the world – 100 countries, which is the majority – are actively developing offensive, not defensive, but offensive – cyber capabilities'.[9] Similar estimates are provided in a report from James Lewis written for the United Nations, and others provide more conservative estimates.[10]

[7] See section 4 for an explanation of this dilemma.

[8] Notice that the categories found in the table concern 'ideal types' of OI which serve to clarify the scope of this paper. In practice, these forms of OI probably overlap.

[9] Jamie Shea, 'Lecture 6 - Cyber attacks: hype or an increasing headache for open societies?', (29 February, 2012). Retrieved from: http://www.nato.int/cps/en/natolive/opinions_84768.htm; For a similar statement see: INFOSEC, 'The Rise of Cyber Weapons and Relative Impact on Cyberspace', (October 5, 2012). Retrieved from: http://resources.infosecinstitute.com/the-rise-of-cyber-weapons-and-relative-impact-on-cyberspace/.

[10] James Lewis, 'The Cyber Index: International Security Trends and Realities', United Nations Institute for Disarmament Research, 2013. Retrieved from: http://www.unidir.org/files/publications/pdfs/cyber-index-2013-en-463.pdf; Kim Zetter, 'We are at Cyberwar: A global guide to nation-state digital attacks', *Wired*, 2015. Retrieved from: https://www.wired.com/2015/09/cyberwar-global-guide-nation-state-digital-attacks/.

TABLE 1. TYPES OF OI IN RELATION TO CYBER SECURITY

		Defence	Offence
High-Level OI	Government organisation – Proxy, Private, or other State	Type A[11]	Type D[12]
Mid-Level OI	Government organisation with no cyber activities – Government organisation with cyber activities	Type B[13]	Type E[14]
Low-Level OI	Government organisation with cyber activities – Government organisation with cyber activities	Type C[15]	Type F[16]

I focus on this level because it is the one which has received the least amount of rigorous analysis but where the stakes are potentially the highest. Unlike discussions about initiatives promoting defensive measures, offensive cyber capability development has remained shrouded in secrecy, perhaps even more so than conventional security issues. Yet organisational mismanagement of offensive cyber activity can lead to unnecessary cycles of provocation, with potentially disastrous consequences.[17]

OIOCC is not only relevant for those states which seek to establish initial operability to conduct sophisticated cyber attacks. In fact, OIOCC also addresses some of the core concerns that

[11] Ralf Bendrath, 'The Cyberwar Debate: Perception and Politics in US Critical Infrastructure Protection', *Information & Security*, 7, (2001), 80-103; Myriam Dunn Cavelty, *Cyber-Security and Threat Politics US Efforts to Secure in the Information Age* (Routledge: 2008); Robert K. Knake, 'Internet Governance in an Age of Cyber Insecurity', Council on Foreign Relations, Special Report No.56, (2010); Jerry Brito and Tate Watkins, 'The Cybersecurity-Industrial Complex', *Reason*, 43,4 (2011); Myriam Dunn-Cavelty, and Manuel Suter, 'Public-Private Partnerships are no silver bullet: An expanded governance model for Critical Infrastructure Protection', *International Journal for Infrastructure Protection*, 2(2009), 179- 187; Shmuel Even, 'The Strategy for Integrating the Private Sector in National Cyber Defense in Israel', *Military and Strategic Affairs*, 7:2 (2015).

[12] See Tim Maurer, '"Proxies" and Cyberspace', *Journal of Conflict and Security Law*, 21:3 (2016) 383-403; Tim Maurer, 'Cyber Proxies and the Crisis in Ukraine', in Kenneth Geers (ed.), *Cyber War in Perspective: Russian Aggression against Ukraine*, (NATO CCD COE Publications: Tallinn: 2015); Richard Clarke, 'War from Cyberspace', *National Interest*, 104 (2009), 31-36.

[13] Rachel Yould, 'Beyond the American Fortress: Understanding Homeland Security in the Information Age'. In *Bombs and Bandwidth: The Emerging Relationship Between Information Technology and Security*, ed. Robert Latham. (The New Press: 2003); John Tritak, 'Protecting America's Critical Infrastructures: How Secure Are Government Computer Systems?', Hearing before the Committee on Energy and Commerce, (5 April 2001).

[14] Fred Kaplan, *Dark Territory: The Secret History of Cyber War*, (Simon & Schuster: New York: 2016); The US Department of Defense, 'The DoD Cyber Strategy', (April 2015); Sorin Ducaru, 'The Cyber Dimension of Modern Hybrid Warfare and its relevance for NATO', *Europolity*, Continuity and Change in European Governance, 10:1 (2016). Retrieved from: http://europolity.eu/wp-content/uploads/2016/07/Vol.-10.-No.-1.-2016-editat.7-23.pdf; Frank Hoffman, 'Hybrid Warfare and Challenges', *Joint Force Quarterly*, 52 (2009), 34-39.

[15] Cheryl Pellerin, 'New Threat Center to Integrate Cyber Intelligence', US Department of Defense, (February 11, 2015). Retrieved from: https://www.defense.gov/News/Article/Article/604093; Richard Bejtich, 'What are the Prospects for the Cyber Threat Intelligence Integration Center?' Brookings Institution, (February 19, 2015). Retrieved from: https://www.brookings.edu/blog/techtank/2015/02/19/what-are-the-prospects-for-the-cyber-threat-intelligence-integration-center/.

[16] Mark Pomerleau, 'Services integrating cyber and traditional military forces', (September 30, 2016). Retrieved from: http://www.c4isrnet.com/articles/services-integrating-cyber-and-traditional-military-forces.

[17] For example, it can lead to the use of offensive capabilities not in line with legal conduct or increase the chances of 'cyber accidents'.

highly advanced cyber powers are currently grappling with. Adam Segal penned five takeaways from the annual 'National Thought Leaders' visit to the National Security Agency (NSA) in December 2016: i) the reasoning for 'loud' cyber weapons;[18] ii) the splitting of NSA and Cyber Command;[19] iii) the need for a new workforce model; iv) private sector outreach; and v) the creation of a new cyber force.[20] Through this OI analysis, in which I also analyse how the development of offensive cyber capabilities is distinctive compared to other processes, we gain a better understanding of how some of these critical challenges can be resolved.

The remainder of this paper is structured as follows. First, it deals with the first-order question of defining OI and OIOCC. It also discusses the optimal product design of offensive cyber capabilities and current forms of OIOCC. In the second part, I develop several propositions about the general benefits of OIOCC. The third part develops three propositions about the potential negative effects of integration, and the final part concludes and draws out lessons on which form of OI would be suitable for the development of offensive cyber capabilities.

2. THE NATURE OF OI AND OIOCC

There are dozens of definitions of organisational integration. For the purposes of this discussion, I follow Lawrence and Lorsch's definition of OI as 'the process of achieving unity of effort among the various subsystems in the accomplishment of the organisation's tasks'.[21] In more poetic terms, integration is about achieving and committing to a harmonious marriage of activities. Like any good marriage, there are many ways in which OI can be successfully achieved.[22]

As a primary step to understanding the nature and requirements of OIOCC, a discussion of the desired 'output' is required. The expectation is that OIOCC of states is directed towards the development of *sophisticated* capabilities.[23] Sophistication in this context refers to the complexity of techniques put into the development of a capability to enable it to gain

[18] As Segal writes 'A senior official confirmed that sometimes Cyber Command wants an adversary to know it has conducted an operation and so in some instances it embeds the equivalent of ['from US Cyber Command' in the code'. Adam Segal, 'Takeaways From a Trip to the National Security Agency', *Council on Foreign Relations*, (December 21, 2016). Retrieved from: http://blogs.cfr.org/cyber/2016/12/21/takeaways-from-a-trip-to-the-national-security-agency/.

[19] It was officially announced in December 2016 that the US will change the dual hat arrangement at the NSA and Cyber Command. See: Ellen Nakashima, 'Obama moves to split cyberwarfare command from the NSA', *Washington Post*, (December 23, 2016). Retrieved from: https://www.washingtonpost.com/world/national-security/obama-moves-to-split-cyberwarfare-command-from-the-nsa/2016/12/23/a7707fc4-c95b-11e6-8bee-54e800ef2a63_story.html?utm_term=.8dba21add7e9; US Department of Defense, 'Joint Concept on Cyberspace – US Department of Defense', (2011).

[20] Note that, according to Segal, an official of the NSA was against the creation of a new cyber force. Segal, 'Takeaways from a Trip to the National Security Agency'.

[21] Paul R. Lawrence and Jay W. Lorsch, 'Differentiation and Integration in Complex Organisations', *Administrative Science Quarterly*, 12:1 (1967), 1-47; also see Henri Barki and Alain Pinsonneault, 'A model of Organizational Integration, Implementation Effort, and Performance', *Organization Science*, 16:2 (2005)165-179; As the definition indicates, I focus on the cross-function orientation reflecting linkages *within* government, i.e. internal integration.

[22] For a contrasting view, see Tolstoy's Anna Karenina Principle.

[23] Note that a sophisticated actor does not always have to use sophisticated capabilities. For a more extensive discussion on this topic see: Ben Buchanan, 'The Legend of Sophistication in Cyber Operations', Harvard Kennedy School Belfer Center, Working Paper Series, (January, 2017), 1-27.

its objective.[24] As David Aitel notes, cyber operations can differ in the sourcing and use of capabilities, networking, testing, persistence and operational security.[25] This means there are no strict necessary and sufficient conditions when a capability can be considered 'sophisticated', but they share 'family resemblances' such as: i) the exploitation of zero-day vulnerabilities; ii) the implementation of various obfuscation techniques; iii) the ability to deliver payload to difficult-to-reach targets; and iv) the implementation of customised malware of firmware.[26]

Sophisticated cyber attacks generally take place across multiple stages,[27] creating a sequential interdependence between the activities.[28] We can distinguish between four general stages.[29] The first stage is *reconnaissance*. This includes the attacker 'sniffing' (to eavesdrop on existing data traffic), 'footprinting' (to gain knowledge of the network or security posture), and 'enumeration' (to identify user accounts which can potentially be exploited). The second stage concerns intrusion. We commonly distinguish between user intrusion (with non-administrative user privileges) and root intrusion (with administrative user privileges). The third stage of a cyber attack is *privilege escalation*.[30] At this stage a vulnerability is exploited in an operating system (OS) or specific service or software package. Finally, there is the stage which gives away the *goal* of the attacker. This could be denial of service (DoS), installation of a backdoor, exfiltrating data, espionage or corruption (with as its aim to cause harm or damage). The order of activities conducted by Advanced Persistent Threat (APTs) is often highly complex, and different tasks may be executed – even outside cyberspace – with long periods of time in between. For example, a backdoor (i.e. a way to bypass normal authentication) may be initially installed at t=0, and check-ups might take place in later time periods t=1, t=2, and t=3 to see if it still exists, but a future attack using that backdoor might occur only in t=x.

The optimal product design of a sophisticated offensive cyber capability – which OIOCC aims to help achieve – meets four criteria. First, it should be *effective* in achieving its desired goal. In the case of a cyber weapon, this means it can reliably provide unauthorised access to a computer system to cause harm or damage to a living being or system.[31] Second, it needs to comply with

[24] See Max Smeets, 'What it Takes to Develop a Cyber Weapon', Columbia University SIPA: Tech & Policy Initiative, Working Paper Series 1 (2016), 49-67.

[25] David Aitel, 'Useful Fundamental Metrics for Cyber Power', *CyberSecPolitics*, 2016. Retrieved from: https://cybersecpolitics.blogspot.com/2016/06/useful-fundamental-metrics-for-cyber.html.

[26] Ibid. I differ from Aitel, as his framework does not distinguish between type of vulnerability exploited (i.e. zero-day versus non-zero-day exploits).

[27] The term 'multi-stage attack' is often used in the literature. However, it has two different meanings. For Landau and Clarke this occurs when computer A penetrates computer B, which is then used to penetrate computer C, and so on. For others (see for example Rid and Buchanan), multi-stage is when a cyber attack occurs through steps that can be temporarily be distinguished. I refer to the latter meaning in this article. See David D. Clarke and Susan Landau, 'The problem isn't attribution; it's multi-stage attacks', *ACM ReArch*, (November 30, 2010); Thomas Rid and Ben Buchanan, 'Attributing Cyber Attacks', *Journal of Strategic Studies*, 38:1-2 (2015), 4-37.

[28] Other common forms of interdependence concern 'pooled interdependence' and 'reciprocal interdepended'. James D. Thompson, *Organizations in Action: Social Science Bases of Administrative Theory*, (McGraw-Hill: 1967).

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

[30] This is usually for buffer overflow attacks, in which the program overwrites memory adjacent to a buffer that should not have been modified.

[31] Smeets, 'What it Takes to Develop a Cyber Weapon'.

(international) legal standards. A responsible actor seeks more from a weapon than merely the ability to cause harm or damage. Principally, the weapon needs to be discriminate, in that it can be used in accordance with the principles of distinction and proportionality.[32] Third, the development should be *cost efficient* in that it produces the desired result for the least amount of resources poured into it. Finally, it needs to be at an actor's disposal on a permanent basis. For industries, a key factor leading to OI is the demand for steady volumes of output or the assurance that a service is available at all times.[33] In the context of offensive cyber capability development, ideally, governments aim to organise their operations in such a manner that the 'cyber option' is always available as a potential (strategic) asset to use.

One should also be aware of what OIOCC does *not* aim to achieve. A key goal of OI is normally to standardise the output of a process, especially in the manufacturing industry. The aim of a manufacturer is to ensure that every final product has the same properties. For example, customers drinking a certain soda brand expect that their drink will have the same taste each time they buy it. Yet, as Lindsay and Gartzke observe, the essential element of cyber weapons' success is deception, with its basic tactics of dissimulation (hiding what is there) and simulation (showing what it is not).[34] As the attacker constantly needs to find innovative ways to mislead the enemy to ensure the attack is successful, uniform products are *not* desired.[35]

As we can observe from the activities of various governments, OIOCC can come in many shapes and forms: through appointing the same director for the intelligence services as for the command conducting military activities; through establishing a significant intelligence constituency within the military command; through offering the same training programme to those conducting espionage operations and those conducting offensive (destructive) cyber operations; and through people moving from an intelligence gathering unit to a military unit and vice versa. There is no ideal configuration of integration, as it depends on size and capital (also human capital) of the cyber operations. There are however a number of potential benefits and risks of OIOCC which are discussed below.

3. THE BENEFITS OF OIOCC

I develop three propositions about the positive effects of OIOCC in helping to achieve 'optimal' output: i) interaction efficiency, ii) knowledge transfer and organisational learning and iii) mission overlap.

A. Proposition 1: OIOCC Leads to Interaction Efficiency of Intelligence and Military Activities

The obstacles that actors have to overcome to conduct a sophisticated cyber attack are often

[32] Legal compliance increases the costs of development due to the additional need for testing, grading costs, and the losses of rejected capabilities.

[33] J.A. Seagraves and C.E. Bishop, 'Impacts of Vertical Integration on Output and Industry Structure', *Journal of Farm Economics*, 40 (1968), 1814-1827.

[34] Lindsay and Gartzke consider deception to be a distinct strategy, similar to deterrence in the nuclear era. Erik Gartzke and Jon R. Lindsay, 'Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace', *Security Studies*, 24:2, (2015), 316-348.

[35] As Martin Libicki states, there is no 'forced entry' when it comes to offensive cyber operations. Martin Libicki, *Conquest in Cyberspace: National Security and Information Warfare* (New York: Cambridge University Press, 2007), 31-36.

underestimated.[36] Smeets distinguishes between three different types of obstacles for different types of capabilities: first, there are the knowledge/intelligence barriers to overcome in developing a capability; second, there are economic and material obstacles; and third, there are organisational obstacles as various actors have to work together to develop a certain capability.[37] With respect to the latter, the interdependence between the intelligence collection activity and the military activity is essential as a great amount of preparation is required to understand the targeted systems.[38] OI facilitates this interaction, making it the most obvious benefit.

The importance of this close link between intelligence and military has been well-documented in the case of Stuxnet.[39] As Jon Lindsay observes:

> Stuxnet infections have been traced to five different industrial companies within Iran, all of which dealt in [Industrial Control System] equipment […]. These domains were infected on multiple occasions with an average of nineteen days between malware compilation and the date of infection.[40]

To match Stuxnet's payload with the specific type of programmable logic controller required highly specific intelligence, and to engineer a 'perfect' code, thorough testing was required. A mock-up plant was therefore created using Qaddafi's P-1 centrifuges.[41] Over time, the tests grew in size and sophistication. According to David Sanger, at some point the United States and Israel were 'even testing the malware against mock-ups of the next generation of centrifuges the Iranians were expected to deploy, called IR-2s, and successor models, including some the Iranians still are struggling to construct'.[42] Overall, as Lindsay concludes, 'the Stuxnet operation required substantial time and institutional infrastructure'.[43] Like the small elements of a high-quality Swiss clockwork beautifully working together to show the time with complete precision, to conduct a sophisticated cyber operation like Stuxnet, a close connection between numerous actors and processes is necessary.[44]

[36] Smeets, 'What it Takes to Develop a Cyber Weapon'; Jon R. Lindsay, 'Stuxnet and the Limits of Cyber Warfare', *Security Studies*, 22:3 (2013), 365-404.

[37] Smeets, 'What it Takes to Develop a Cyber Weapon'.

[38] This means a serial relationship exists as the output from the reconnaissance operation becomes the input for the cyber operations with as aim to cause harm or damage.

[39] Stuxnet was allegedly developed and deployed by the United States and Israel. The origins of the worm go back as early as 2006, midway through George W. Bush's second term, as the administration aimed to diversify its options against Iran. David Sanger, *Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power* (Random House: New York: 2012).

[40] Jon R. Lindsay, 'Stuxnet and the Limits of Cyber Warfare', *Security Studies*, 22:3 (2013), 365-404.

[41] Ibid.

[42] Sanger, 'Confront and Conceal', 198.

[43] Lindsay, 'Stuxnet and the Limits of Cyber Warfare', 387.

[44] According to Rob Morgus, there was potentially a similar close link in Russia's cyber operations against Ukraine. Also Kim Zetter writes that 'skilled and stealthy strategists [launched] a synchronized assault in a well-choreographed dance.' Yet, considering that there were clear delineations between the various phases of the operations suggests that this was a form of collaboration (between potentially criminal and nation-state actors, as Zetter writes) rather than integration. Robert Morgus, 'Whodunnit? Russia and Coercion through Cyberspace', 2016. Retrieved from: https://warontherocks.com/2016/10/whodunnit-russia-and-coercion-through-cyberspace/; Kim Zetter, 'Inside the Cunning Unprecedented Hack of Ukraine's Power Grid', *Wired*, 2016. Retrieved from https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/.

B. Proposition 2: OIOCC Increases the Opportunity for Knowledge Transfer

OIOCC also stimulates the transfer of knowledge. Knowledge transfer refers to the process through which one individual, group, or unit is affected by the experience of another.[45] Empirical evidence suggests that interconnected organisations such as franchises and alliances hold comparative advantages over their more autonomous counterparts due to the ability to transfer knowledge across their constituent parts. For example, a restaurant may put a new dish on the menu that was successfully served at its sister restaurant located in a different part of town.[46]

The required knowledge component for developing offensive cyber capabilities comes in two forms.[47] First, there is the part which is explicit and can be transferred in a systematic manner. For example, it concerns knowledge about how different elements of an industrial control system work. Another example of this form of knowledge is the ability of a person to write code in a certain programming language. The most significant knowledge component is often tacit and difficult to articulate.[48] For example, this might concern knowledge embedded in a hacker's experience or a cyber command's implicit operational processes.[49] The tacit knowledge required to conduct offensive cyber activities cannot be formally communicated. Forms of OIOCC, however, allow for the opportunity to also transfer this type of knowledge.

C. Proposition 3: OIOCC Minimises Mission Overlap and Improves Resource Allocation

Finally, OIOCC allows for a more efficient allocation of resources, because the same processes or tasks are not replicated unnecessarily. In any organisation, some overlap in tasks is inevitable; yet OIOCC leads to enhancement resource allocation on three levels.

First, it can put people to better use. In a number of countries, the growth of offensive cyber capabilities in militaries already allowed for greater specialisation in cyber weapon production. The US Cyber Command now has 133 teams in operation, making it easier to dedicate specialised units to specific types of cyber operations.[50] OI frees up further resources for specialisation. This means that task complexity can be increased, as tasks can be divided on the basis of who is most proficient in the process. Second, 'capital' can be used more efficiently. Two basic forms of 'capital' of offensive cyber capabilities concern software and infrastructure. On the former, software can often be divided into smaller modules (called modularity); OIOCC makes it easier to reuse parts of code from a different operation in order to save time and resources. On the latter, a type of infrastructure frequently used by attackers is command and control (C&C) servers, which help to maintain communications with compromised machines. Also, to launch

[45] Linda Argote, Paul Ingram, John M. Levine, and Richard L. Moreland, 'Introduction: Knowledge Transfer in Organizations: Learning from the Experience of Others', *Organizational Behavior and Human Decision Processes*, 82 (1) (2000), 1-8; 3.

[46] Ibid, 4.

[47] Smeets, 'What it Takes to Develop a Cyber Weapon'.

[48] Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy*, (University of Chicago Press, Chicago: 1958).

[49] See also: Thomas Rid, *Cyber War Will Not Take Place*, (C. Hurst & Co: London: 2013), 83-84; Martin Davies, 'Knowledge – Explicit, implicit and tacit: Philosophical aspects', in *International Encyclopaedia of Social and Behavioral Sciences*, James D. Wright (ed.) (Elsevier: 2015).

[50] According to Rob Morgus, this was also the case for Russia's cyber operations against Ukraine. Robert Morgus, 'Whodunnit? Russia and Coercion through Cyberspace', *War on the Rocks*, 2016. Retrieved from: https://warontherocks.com/2016/10/whodunnit-russia-and-coercion-through-cyberspace/.

distributed denial-of-service (DDoS) attacks, a botnet of 'zombies' (compromised computers) are often used flood the bandwidth of a targeted system.[51] Again, integration makes it easier to repurpose this type of infrastructure for a different cyber operation.[52] Third, and related, OIOCC makes it easier to leverage previous activities and accomplishments. For example, if a backdoor in a computer system is already established by one actor, it can conveniently be used by another.

4. THE RISKS OF OIOCC

Organisational integration, however, does not only have positive consequences. I consider three propositions about the negative effects created by OIOCC.

A. Proposition 1: OIOCC Intensifies the Cyber Security Dilemma

A prominent theoretical idea in International Relations is the security dilemma. The situation occurs in an anarchical system when states cannot be certain about each other's intentions. Mutual fear leads states to resort to an accumulation of capabilities to defend themselves, which in turn leads to (unintended) spirals of worsening relationships with a potentially tragic outcome.[53] Buchanan indicates that also a cyber security dilemma exists, as states which aim to assure their own (cyber) security have an incentive to intrude into strategically important networks of other states and will thus threaten – often unintentionally – the security of those other states, risking escalation.[54]

The first potential downside of OIOCC is that it intensifies this cyber security dilemma. The knowledge transfer incentive, mentioned above as a positive effect of OIOCC, also has a negative externality for the development of offensive capabilities. Argote and Ingram note that '[k]nowledge transfer in organisations manifests itself through changes in the knowledge or performance of the recipient units'.[55] It leads to a loss of distinct organisational cultures with their practices and 'playbooks'. The commonality in behaviour of OIOCC further blurs the line between when a cyber espionage operation ends, and a destructive operation starts. From the

[51] See Rid and Buchanan, 'Attributing Cyber Attacks', 17.

[52] FireEye offers an example in their report on attribution of advanced cyber attacks: 'four separate attacks that use different exploits, different lures, and different first-stage malware implants. But they all target religious activities. And […] they are all sent from the same server […]. This evidence points to multiple actors on the same team, using the same infrastructure'. See: FireEye, 'Digital Bread Crumbs: Seven Clues to Identifying Who's Behind Advanced Cyber Attacks', (2013, June). Retrieved from: https://www.fireeye.com/content/dam/fireeye-www/global/en/current-threats/pdfs/rpt-digital-bread-crumbs.pdf.

[53] For original formulations, see: Herbert Butterfield, *History and Human Relations*, (London: Collins, 1951); John Herz, *Political Realism and Political Idealism: A Study in Theories and Realities*, (Chicago: University of Chicago Press, 1951); Robert Jervis, *Perception and Misperception in International Politics*, (Princeton, NJ: Princeton University Press, 1976), Chap. 3; and Jervis, 'Cooperation under the Security Dilemma', *World Politics* 30: 2 (1978), 167–214. For a more recent review see: Shiping Tang, 'The Security Dilemma: A Conceptual Analysis', *Security Studies*, 18:3 (2009), 587-623.

[54] For a more detailed discussion see: Ben Buchanan, *The Cyber Security Dilemma*, (Oxford: Oxford University Press: 2017).

[55] Linda Argote and Paul Ingram, 'Knowledge Transfer: A Basis for Competitive Advantage in Firms' *Organizational Behavior and Human Decision Processes*, 82:1 (2000), 150-169.

defenders' viewpoint, it is more difficult to discern the cyber attackers' intention and accurately respond.[56]

The matter can be put into perspective considering the following (ever more relevant) question: what do we do if we find out about cyber espionage activities on our computer systems? 'Respond proportionately' would be the diplomatic answer. Indeed, in the last press conference of 2015, President Obama said the US would take 'proportionate' action in response to a cyber attack on Sony Pictures by North Korea. Proportionality was also the cornerstone of President Obama's response to the Democratic National Committee (DNC) hack. For a long time, it was unclear what the US government does against Russia's aggression, other than 'respond in a proportionate manner'.

But what does 'proportionate' mean in this context? The response is likely not just tailored towards the *actual* 'observed' activity (i.e. an espionage operation) but also the perceived *intent* of the attacker (i.e. an assessment of whether a destructive cyber attack will follow or not). The underlying rationale of the attacker is, however, difficult to interpret, especially if an espionage operation and a destructive operation are conducted in the same operational fashion. This means that we are unable to discern if our *perceived* proportionate response is *actually* proportionate or not.

B. Proposition 2: OIOCC Leads to Cost Ineffectiveness in the Long Run

Offensive cyber capability integration may seem to be economical in the short-term, but will most likely increase costs down the road. Offensive cyber capabilities are said to be transitory in nature, meaning they are relatively short lived in their ability to cause harm or damage.[57] Cyber capabilities are not equally transitory, and one of the factors which explains this differentiation concerns the type of payload. All other things being equal, it can be said that destructive cyber capabilities causing visible damage are more likely to be discovered than espionage capabilities, and hence lead to the patching of a vulnerability.[58]

However, due to the integration of activities, the deployment of a destructive cyber weapon also increases the risk that espionage capabilities – exploiting the same vulnerabilities and same coding procedures – are exposed as well. The issue is particularly pertinent now that threat intelligence firms (and possibly states) are establishing special detection tools in attempt to uncover clusters of capabilities. This means that integration makes an attacker's offensive cyber 'arsenal' more volatile and costly, as multi-year cyber programs are more susceptible to a low return on investment after capabilities with a destructive payload are used.

A more general risk stems from changes in information access due to OIOCC which may also ultimately increase costs. Integration aims to simulate knowledge transfer within the

[56] The US Cyber Command has recently proposed the development of 'loud' cyber weapons for when you want to be attributed; though this raises a great number of operational questions; Chris Bing, 'US Cyber Command director: We want 'loud', offensive cyber tools', *Fedscoop*, (August 30, 2016). Retrieved from: http://fedscoop.com/us-cyber-command-offensive-cybersecurity-nsa-august-2016; Herb Lin, Developing 'Loud' Cyber Weapons (September 1, 2016). Retrieved from: https://www.lawfareblog.com/developing-loud-cyber-weapons; Herb Lin, 'Still More on Loud Cyber Weapons', *Lawfare*, (October 19, 2016) retrieved from: https://www.lawfareblog.com/still-more-loud-cyber-weapons.

[57] Notice that certain cyber capabilities are more transitory than others. For a more extensive discussion on this point, see: Max Smeets, 'A Matter of Time: On the Transitory Nature of Cyber Weapons', *Journal of Strategic Studies*, (2017), 1-28.

[58] Ibid.

organization. Yet, it is inherently more demanding to manage the risks of information security (normally shortened to InfoSec) when more people are able to view classified documents and are briefed about certain operations.[59]

C. Proposition 3: OIOCC Leads to Cyber Mission Creep

Mission creep refers to the gradual expansion of a project or mission beyond its original goals. This usually occurs following the initial successes of an operation. As Caraccilo writes, 'mission creep is usually considered undesirable due to the dangerous path of each success breeding more ambitious attempts, only stopping when a final, often catastrophic, failure occurs'.[60]

The ongoing transfer of civilian tasks to military agencies has been well-documented in the case of the global war on terrorism. In 2008, US Secretary of Defense Robert Gates warned:

> [o]verall, even outside Iraq and Afghanistan, the United States military has become more involved in a range of activities that in the past were perceived to be the exclusive province of civilian agencies and organisations. This has led to concern among many organisations about what's seen as a creeping 'militarisation' of some aspects of America's foreign policy. This is not entirely unreasonable sentiment.[61]

Mission creep is also observed in domestic and international organisations.[62] Monahan, for example, critiques the emergence of data-sharing 'fusion centres' intended to reduce crime and prevent terrorism. His main critique is that:

> these centres […] mutate into 'all-crimes' and 'all-hazards' organisations […] in order to ensure their continued existence in the face of future budget cuts.[63]

It can be said that OIOCC enhances the risk of what I label cyber mission creep: the risk that an organisation responsible for offensive cyber activities goes beyond their core mission in the use of these capabilities. Cyber mission creep can come in two forms. First, there is the risk that initial intelligence activities through cyberspace go beyond their core mission and lead to the use of cyber weapons.[64] This is especially likely to occur when intelligence services in a country are perceived to be more capable than the military. The second form of cyber mission creep occurs when the military is gradually conducting more cyber activities, even when it

[59] InfoSec normally has three components: i) confidentiality, ii) integrity, and iii) availability.

[60] Dominic Joseph Caraccilo, *Beyond Guns and Steel: A War Termination Strategy*, (Santa Barbara: Praeger Security International), 178.

[61] Robert M. Gates, 'Secretary of Defense Speech', US Global Leadership Campaign, July, 8 2008, Washington, DC; Gordon Adams and Shoon Murray, *Mission Creep: The militarization of US Foreign Policy*, (Georgetown University Press: Washington, DC: 2014).

[62] Jessica Einhorn, 'The World Bank's Mission Creep', *Foreign Affairs*, 80:5 (2001), 22–35; Ngaire Woods, *The Globalizers: The IMF, the World Bank and their Borrowers* (Ithaca, NY: Cornell University Press, 2006).

[63] Torin Monahan, 'The murky world of Fusion Centres', *Criminal Justice Matters*, 75:1 (2009), 20-21.

[64] Although I refer to this as a new form of mission creep, I do not argue that the same dynamics which cause military mission creep cause cyber mission creep.

concerns minor (domestic) disorders rather than war or another emergency situation.[65] Indeed, there is the risk that intelligence agencies mutate into 'all-cyber' organisations.

5. CONCLUSION

In response to cyber security challenges, governments have established a wide range of informal and formal institutions. The focus of this paper was to enhance our understanding of the organisational integration of offensive cyber capabilities. In doing so, it will help states which already have developed, as well as those which are developing, offensive cyber capabilities to make better decisions about their organisational structures.

OIOCC can come in many shapes and forms, and the ideal configuration depends on a state's set up. This paper has developed in total six general propositions about the potential consequences of OIOCC. Following the propositions of the benefits of OIOCC, it was argued that integration reduces mission overlap, increases cost effectiveness, and improves communication. I also call attention to the risks of OIOCC, which are still not well understood. OIOCC deepens the cyber security dilemma and leads to potential cyber mission creep. Also, the potential cost effectiveness created through OIOCC in the short run might turn into cost ineffectiveness overall, as 'clusters' of capabilities are more likely to be discovered. Although the benefits seem to outweigh the risks, a misunderstanding of the risks can have significant consequences for conflict escalation.

This paper also presents some limitations and opportunities for future research. First, this paper has been primarily theoretical and should still be subject to empirical scrutiny in order to test the validity of the OIOCC propositions. Ideally, a comparative analysis would be conducted between countries with different levels of OIOCC to verify whether the propositions presented in this paper hold in practice. Second, it was noted that the nature of cyber capabilities places greater emphasis on innovation than on forced control. This means that in the case of OIOCC, an increase in organisational efficiency should not come at the cost of organisational and individual flexibility. More research should be conducted on the mechanisms to achieve these

[65] Another form of creepism can be observed for offensive cyber capabilities, which is called in product design terms 'feature creep' (and for computer programs 'software bloat'). This refers to ongoing addition of new features in products, resulting in over-complication. We have this also with offensive cyber capabilities. Stuxnet is again a good example. Ralph Langner indicates that Stuxnet actually refers to two weapons, instead of one. While the early version focused on loosening the isolation valves of the Natanz uranium enrichment facility, the later version sought to change the speeds of the rotors in the centrifuges. Langner notes in his report that 'The attack routines for the overpressure attack [of the second version] were still contained in the payload, but no longer executed – a fact that must be viewed as deficient OPSEC. It provided us by far the best forensic evidence for identifying Stuxnet's target, and without the new, easy-to-spot variant the earlier predecessor may never have been discovered. That also means that the most aggressive cyber-physical attack tactics would still be unknown to the public – unavailable for use in copycat attacks, and unusable as a deterrent display of cyber power'. We however have not information to conclude whether this type of feature creep – leading to a less stealthy (a potentially less effective) capability – was the result of an *inter-agency* problem or an *organisational integration* problem. See: Ralph Langner, 'Stuxnet's Secret Twin', Foreign Policy, (19 November 2013). Retrieved from: http://foreignpolicy.com/2013/11/19/stuxnets-secret-twin/; Ralph Langner, 'To Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve', *The Langner Group*, (November 2013). Retrieved from: http://www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf, 5; Geoff McDonald, Liam O Murchu, Stephen Doherty, Eric Chien, 'Stuxnet 0.5: The Missing Link' *Symantec*, (26 February, 2013). Retrieved from: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/stuxnet_0_5_the_missing_link.p.

goals concurrently. Third, I have limited the scope of my research to government actors; so-called 'supplier integration' of other actors has not been the focus of my research. An obvious extension would be to look at the interaction between private actors and government actors in developing offensive cyber capabilities. After all, when the demand for offensive capabilities increases (both in complexity and size), states can expand their activities by persuading legislators to increase budgets and size of departments or by taking on 'private suppliers'.

ACKNOWLEDGMENTS

For written comments on early drafts, I am indebted to Florian Egloff, Richard Harknett, Lennart Maschmeyer, Nikolas Ott, Henry Rõigas, James Shires, and anonymous reviewers. An earlier version of this paper was presented at the Oxford Cyber Studies Working Group.

REFERENCES

Adams, Gordon and Shoon Murray, Mission Creep: *The militarization of US Foreign Policy*, (Georgetown University Press: Washington, DC, 2014).

Aitel, David, 'Useful Fundamental Metrics for Cyber Power,' *CyberSecPolitics*, 2016. Retrieved from: https://cybersecpolitics.blogspot.com/2016/06/useful-fundamental-metrics-for-cyber.html.

Argote, Linda and Paul Ingram, 'Knowledge Transfer: A Basis for Competitive Advantage in Firms,' *Organizational Behavior and Human Decision Processes*, 82:1 (2000), 150-169.

Argote, Linda, Paul Ingram, John M. Levine, and Richard L. Moreland, 'Introduction: Knowledge Transfer in Organizations: Learning from the Experience of Others,' *Organizational Behavior and Human Decision Processes*, 82 (1) (2000), 1-8.

Barki, Henri and Alain Pinsonneault, 'A model of Organizational Integration, Implementation Effort, and Performance,' *Organization Science*, 16:2 (2005), 165-179.

Barney, Jay, 'Firm resources and sustained competitive advantage,' *Journal of Management*, 17:1 (1991), 99-120.

Bejtich, Richard, 'What are the Prospects for the Cyber Threat Intelligence Integration Center', Brookings Institution, (February 19, 2015). Retrieved from: https://www.brookings.edu/blog/techtank/2015/02/19/what-are-the-prospects-for-the-cyber-threat-intelligence-integration-center/.

Bendrath, Ralf, 'The Cyberwar Debate: Perception and Politics in US Critical Infrastructure Protection,' *Information & Security*, 7, (2001), 80-103.

Bing, Chris, 'US Cyber Command director: We want "loud," offensive tools,' *FedScoop* (August 30, 2016). Retrieved from: http://fedscoop.com/us-cyber-command-offensive-cybersecurity-nsa-august-2016.

Brito, Jerry and Tate Watkins, 'The Cybersecurity-Industrial Complex,' *Reason*, 43, 4 (2011).

Brookes, Chris, 'Cyber Security: Time for an integrated whole-of-nation approach in Australia,' Centre for Defence and Strategic Studies, (March 2015). Retrieved from: http://www.defence.gov.au/ADC/Publications/IndoPac/150327%20Brookes%20IPS%20paper%20-%20cyber%20(PDF%20final).pdf

Buchanan, Ben, *The Cyber Security Dilemma*, (Oxford: Oxford University Press: 2017).

Buchanan, Ben, ‘The Legend of Sophistication in Cyber Operations,’ Harvard Kennedy School Belfer Center, Working Paper Series, (January, 2017), 1-27.

Butterfield, Herbert, *History and Human Relations*, (London: Collins, 1951).

Chalmeta, Ricardo, Christina Campos, Reyes Grangel, ‘Reference architectures for enterprise integration,’ *The Journal of Systems Software*, 57 (2001), 175-191.

Clarke, David D. and Susan Landau, ‘The problem isn’t attribution; it’s multi-stage attacks,’ *ACM ReArch*, (November 30, 2010).

Clarke, Richard, ‘War from Cyberspace,’ *National Interest*, 104 (2009), 31-36.

Davies, Martin, ‘Knowledge – Explicit, implicit and tacit: Philosophical aspects,’ in *International Encyclopaedia of Social and Behavioral Sciences*, James D. Wright (ed.) (Elsevier, 2015).

Ducaru, Sorin, ‘The Cyber Dimension of Modern Hybrid Warfare and its relevance for NATO,’ *Europolity*, Continuity and Change in European Governance, 10:1 (2016). Retrieved from: http://europolity.eu/wp-content/uploads/2016/07/Vol.-10.-No.-1.-2016-editat.7-23.pdf.

Dunn Cavelty, Myriam and Manuel Suter, ‘Public–Private Partnerships are no silver bullet: An expanded governance model for Critical Infrastructure Protection,’ *International Journal for Infrastructure Protection,* 2 (2009), 179-187.

Dunn Cavelty, Myriam, *Cyber-Security and Threat Politics US Efforts to Secure in the Information Age,* (Routledge: 2008).

Dominic Joseph Caraccilo, *Beyond Guns and Steel: A War Termination Strategy*, (Santa Barbara: Praeger Security International).

Ettlie, John and Ernesto M. Reza, ‘Organizational Integration and Process Innovation,’ *Academy of Management Journal*, 35:4 (2001), 795-827.

Einhorn, Jessica, ‘The World Bank’s Mission Creep,’ *Foreign Affairs*, 80:5 (2001), 22-35.

Even, Shmuel, ‘The Strategy for Integrating the Private Sector in National Cyber Defense in Israel,’ *Military and Strategic Affairs*, 7:2 (2015).

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

Gartzke, Erik and Jon R. Lindsay, ‘Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace,’ *Security Studies*, 24:2, (2015), 316-348.

Gates, Robert M., ‘Secretary of Defense Speech,’ US Global Leadership Campaign, Washington, DC (July 8, 2008).

Herz, John, P*olitical Realism and Political Idealism: A Study in Theories and Realities*, (Chicago: University of Chicago Press, 1951).

Hoffman, Frank, ‘Hybrid Warfare and Challenges,’ *Joint Force Quarterly*, 52 (2009), 34-39.

INFOSEC, ‘The Rise of Cyber Weapons and Relative Impact on Cyberspace,’ (October 5, 2012). Retrieved from: http://resources.infosecinstitute.com/the-rise-of-cyber-weapons-and-relative-impact-on-cyberspace/.

International Telecommunication Union, ‘Cyberwellness Profile Republic of Korea,’ (December 1, 2014). Retrieved from: https://www.itu.int/en/ITU-D/Cybersecurity/Documents/Country_Profiles/Korea.pdf.

Jervis, Robert, ‘Cooperation under the Security Dilemma,’ *World Politics* 30: 2 (1978), 167-214.

Jervis, Robert, *Perception and Misperception in International Politics*, (Princeton, NJ: Princeton University Press, 1976).

Kaplan, Fred, *Dark Territory: The Secret History of Cyber War*, (Simon & Schuster: New York: 2016).

Knake, Robert K., 'Internet Governance in an Age of Cyber Insecurity,' *Council on Foreign Relations*, Special Report No.56, (2010).

Langner, Ralph, 'Stuxnet's Secret Twin,' *Foreign Policy*, (November 19 2013). Retrieved from: http://foreignpolicy.com/2013/11/19/stuxnets-secret-twin/.

Langner, Ralph, 'To Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve,' *The Langner Group*, (November 2013). Retrieved from: http://www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf.

Lawrence, Paul R. and Jay W. Lorsch, 'Differentiation and Integration in Complex Organizations,' *Administrative Science Quarterly*, 12:1 (1967), 1-47.

Lewis, James Andrew, 'Advanced Experiences in Cybersecurity Policies and Practices: An Overview of Estonia, Israel, South Korea, and the United States,' Inter-American Development Bank, Discussion Paper IDB-DP-457, (2016, July). Retrieved from: https://publications.iadb.org/bitstream/handle/11319/7759/Advanced-Experiences-in-Cybersecurity-Policies-and-Practices-An-Overview-of-Estonia-Israel-South-Korea-and-the-United%20States.pdf?sequence=.

Lewis, James Andrew, 'The Cyber Index: International Security Trends and Realities,' United Nations Institute for Disarmament Research, (2013). Retrieved from: http://www.unidir.org/files/publications/pdfs/cyber-index-2013-en-463.pdf.

Libicki, Martin, *Conquest in Cyberspace: National Security and Information Warfare,* (New York: Cambridge University Press, 2007).

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

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

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

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

Maurer, Tim, '"Proxies" and Cyberspace', *Journal of Conflict and Security Law*, 21:3 (2016), 383-403.

Maurer, Tim, 'Cyber Proxies and the Crisis in Ukraine,' in Kenneth Geers (ed.), *Cyber War in Perspective: Russian Aggression against Ukraine*, (NATO CCD COE Publications: Tallinn, 2015).

McDonald, Geoff Liam O Murchu, Stephen Doherty, Eric Chien, 'Stuxnet 0.5: The Missing Link,' *Symantec*, (February 26, 2013). Retrieved from: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/stuxnet_0_5_the_missing_link.p.

Monahan, Torin, 'The murky world of 'Fusion Centres',' *Criminal Justice Matters*, 75:1 (2009), 20-21.

Morgus, Robert 'Whodunnit? Russia and Coercion Through Cyberspace,' *War on the Rocks*, (October 19, 2016). Retrieved from: http://warontherocks.com/2016/10/whodunnit-russia-and-coercion-through-cyberspace/.

Nakashima, Ellen ‘Obama moves to split cyberwarfare command from the NSA,’ *Washington Post*, (December 23, 2016). Retrieved from: https://www.washingtonpost.com/world/national-security/obama-moves-to-split-cyberwarfare-command-from-the-nsa/2016/12/23/a7707fc4-c95b-11e6-8bee-54e800ef2a63_story.html?utm_term=.8dba21add7e9.

National Cyber Security Centrum, ‘Cybersecuritybeeld Nederland CSBN 2016,’ (2016). Retrieved from: https://www.ncsc.nl/actueel/Cybersecuritybeeld+Nederland/cybersecuritybeeld-nederland-2016.html.

Pellerin, Cheryl, ‘New Threat Center to Integrate Cyber Intelligence,’ US Department of Defense, (February 11, 2015). Retrieved from: https://www.defense.gov/News/Article/Article/604093.

Polanyi, Michael, *Personal Knowledge: Towards a Post-Critical Philosophy*, (University of Chicago Press, Chicago: 1958).

Pomerleau, Mark ‘Services integrating cyber and traditional military forces,’ (September 30, 2016). Retrieved from: http://www.c4isrnet.com/articles/services-integrating-cyber-and-traditional-military-forces.

Rid, Thomas, *Cyber War Will Not Take Place*, (C. Hurst & Co: London: 2013), 83-84.

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

Sanger, David, *Confront and Conceal: Obama’s Secret Wars and Surprising Use of American Power* (Random House: New York: 2012).

Seagraves, J.A. and C.E. Bishop, ‘Impacts of Vertical Integration on Output and Industry Structure,’ *Journal of Farm Economics*, 40 (1968), 1814-1827.

Segal, Adam ‘Takeaways from a Trip to the National Security Agency,’ *Council on Foreign Relations*, (December 21, 2016). Retrieved from: http://blogs.cfr.org/cyber/2016/12/21/takeaways-from-a-trip-to-the-national-security-agency/.

Shea, Jamie, ‘Lecture 6 - Cyber attacks: hype or an increasing headache for open societies?’ (February 29, 2012). Retrieved from: http://www.nato.int/cps/en/natolive/opinions_84768.htm.

Smeets, Max, ‘What it Takes to Develop a Cyber Weapon,’ Columbia University. SIPA: Tech & Policy Initiative, Working Paper Series 1 (2016), 49-67.

Smeets, Max, ‘A Matter of Time: On the Transitory Nature of Cyber Weapons,’ *Journal of Strategic Studies*, (2017), 1-28.

Tang, Shiping, ‘The Security Dilemma: A Conceptual Analysis,’ *Security Studies*, 18:3 (2009), 587-623.

The Netherlands Ministry of Justice, ‘De Nationale Cybersecurity Strategie 2: Van bewust naar bekwaam,’ (2013, October).

The Netherlands Ministry of Defense, ‘Cyber Command’. Retrieved from: https://www.defensie.nl/english/topics/cyber-security/contents/cyber-command.

Thompson, James D., *Organizations in Action: Social Science Bases of Administrative Theory*, (McGraw-Hill: 1967).

Tritak, John, ‘Protecting America’s Critical Infrastructures: How Secure Are Government Computer Systems?’ hearing before the committee on Energy and Commerce, (April 5, 2001).

Truman, Gregory E., ‘Integration in electronic exchange environments,’ *Journal of Management Information Systems*, 17:1 (2000), 209-244.

UK Government, National Cyber Security Strategy 2016-2021, (2016). Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf.

US Department of Defense, 'Joint Concept on Cyberspace - US Department of Defense,' (2011).

US Department of Defense, 'The DoD Cyber Strategy,' (April 2015).

Waltz, Kenneth, *A Theory of International Politics*, (New York: McGraw-Hill, 1979).

Woods, Ngaire, *The Globalizers: The IMF, the World Bank and their Borrowers,* (Ithaca, NY: Cornell University Press, 2006).

Yould, Rachel, 'Beyond the American Fortress: Understanding Homeland Security in the Information Age'. In *Bombs and Bandwidth: The Emerging Relationship Between Information Technology and Security*, ed. Robert Latham. (The New Press: 2003).

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

Zetter, Kim, 'We are at Cyberwar: A global guide to nation-state digital attacks,' *Wired*, (2015). Retrieved from: https://www.wired.com/2015/09/cyberwar-global-guide-nation-state-digital-attacks/.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Mission Assurance: Shifting the Focus of Cyber Defence*

Brad Bigelow
Principal Technical Advisor
SHAPE DCOS CIS and Cyber Defence
Mons, Belgium
brad.bigelow@shape.nato.int

Abstract: With the decision by the North Atlantic Council to recognize cyberspace as an operational domain, the NATO Command Structure is now taking on the task of implementing the doctrine, organization and capabilities to incorporate operations in cyberspace into the overall framework of joint operations. This paper outlines some of the challenges implicit in the Council's decision, which was both long-expected due to growing awareness of cyber security challenges within the Alliance and bold in its willingness to recognize what is still an immature and evolving discipline. It addresses two key challenges facing those involved in implementing cyberspace as a domain: understanding the complex composition of cyberspace and accurately identifying the consequences of the asymmetric nature of cyberspace threats. The paper then addresses two key aspects for cyberspace as a domain: mission assurance and collective defense. In the context of implementing cyberspace as an operational domain in traditional military operations and missions, cyberspace operators need to focus on mission assurance, which recognizes the reality of a contested cyberspace, and not simply on cyber security concerns. Although the military role in collective cyber defense is still a somewhat politically-charged issue, the author argues that the best way to enable effective mission assurance in cyberspace is to recognize the need for a clear role for the NATO Command Structure to act as an enabler for the open exchange of cyber defense information with military, civil and commercial organizations.

Keywords: *cyberspace, cyber defence, mission assurance, NATO*

* The views and opinions expressed in this article are those of the author alone and do not necessarily reflect those of NATO.

1. INTRODUCTION

The implications of cyberspace as a new domain for national and collective security have increasingly consumed the time and attention of political and military leaders. The rise of the Internet (and now the Internet of Things), of ubiquitous connectivity, of electronic commerce, of networks with scales several orders of magnitude larger than anything seen even a decade ago, have made cyberspace an essential element in all aspects of public life. At the same time, media coverage highlighting the growing frequency, sophistication and impact of cyber attacks has led to a number of policy and organizational decisions. For the North Atlantic Treaty Organisation (NATO), the most recent and significant of these was the decision, taken at the Warsaw Summit in July 2016, to recognize cyberspace as a domain of operations:

> in which NATO must defend itself as effectively as it does in the air, on land, and at sea. This will improve NATO's ability to protect and conduct operations across these domains and maintain our freedom of action and decision, in all circumstances. It will support NATO's broader deterrence and defence: cyber defence will continue to be integrated into operational planning and Alliance operations and missions, and we will work together to contribute to their success (NATO 2016).

Since 2002, cyber defense topics have been included in the deliberations of the North Atlantic Council and Defense Ministers. Attacks on public and private networks in Estonia in 2007 spurred NATO to issue its first policy on cyber defense in 2008. This was updated in 2011 and again in the enhanced policy issued at the Wales Summit in 2014. At the Warsaw Summit in July 2016, Allies pledged to be capable of defending themselves in cyberspace as in the air, on land and at sea. The decision to recognize cyberspace as an operational domain represents, therefore, just a further step in the evolution of NATO's understanding of the importance of cyberspace as an aspect of collective defense.

It was a bold decision, in that it was a strong commitment to incorporate into the Alliance's framework for military operations a domain that is relatively immature in terms of doctrine and capabilities, hampered with vaguely defined terms and concepts, and widely misunderstood. This paper outlines key challenges to implementing the Council's decision to recognize cyberspace as an operational domain, including the lack of common understanding of cyberspace itself and of the nature of the threats in cyberspace. It then outlines the two most important facets of NATO military efforts in cyberspace: mission assurance at the operational and tactical levels, and collective defense at the strategic level. Finally, it argues that the two are inextricably linked and must be approached in an integrated manner to ensure that the Alliance keeps pace with its cyber threats.

2. UNDERSTANDING CYBERSPACE

Of all the challenges, first and foremost is the lack of understanding of what is meant by cyberspace and what constitutes "cyberspace" as an operational domain. The Warsaw Summit declaration itself did not include a definition of the term, nor has it been included in the official NATO Glossary of Terms and Definitions (AAP-6). NATO is not alone in struggling with this. General Michael Hayden (2011, p. 3), who was at the center of the initial development of US cyberspace operational capabilities as Director of the National Security Agency and Director of Central Intelligence, commented: "Rarely has something been so important and so talked about with less clarity and less apparent understanding". Daniel Kuehl (2009, p. 3) listed thirteen different definitions of the term, and Peter W. Singer and Allan Friedman (2014, p. 13) were able to identify twelve different definitions that had been used within the US Department of Defense. For NATO's purposes, however, a good working definition can be established by appropriating a term already used in basic Alliance operational doctrine: "information environment". Allied Joint Publication 3, Allied Doctrine for the Conduct of Operations, defines the information environment as: "[t]he entire infrastructure, organization, personnel, and components that collect, process, store, transmit, display, disseminate, and act on information" (NATO 2011, pp. 4-5).

This definition overcomes the limitations of equating cyberspace with the "global grid" of the Internet and public telecommunications networks. While the Internet is certainly the largest "land mass" of cyberspace, there are many other cyberspaces – closed or largely isolated networks – of which national security, intelligence, law enforcement and classified military networks are the most obvious examples. Deployed military operations rely heavily on operational and tactical communications and networks over radio and satellite links and often secured through a variety of encryption systems. Despite the increasing use of Internet Protocol-based systems and the gradual phasing out of analogue and older digital systems among military forces, the types of attacks and sources of threats seen in the global grid are not necessarily – and certainly not automatically – directly or immediately applicable in the context of all military instantiations of cyberspace.

Neither is cyberspace purely a virtual environment. It has what have been referred to as "littorals" – points at which it overlaps with other environments, much as land and sea converge in the littorals in which amphibious operations take place. These include:

> physical infrastructure, cabling and electrical power; the electromagnetic spectrum that data traverses; electro-mechanical processes under computer control; and the senses and cognition of computer users (Withers 2015, p. 133).

These cyberspace littorals play a significant role in considering the military aspects of cyberspace – again, particularly in the context of deployed operations, where radio and satellite communications provide the primary transmission systems. The new functions of cyberspace operations in a deployed context must be coordinated or integrated with the existing functions of spectrum management, electronic warfare and what some forces call "electromagnetic operations".

A key concern in implementing cyberspace as an operational domain, therefore, is to establish an accurate understanding of the actual "territory" that comprises the cyberspace upon which a military operation depends. In NATO operations, the area in which a designated Joint Force Commander plans and executes a specific mission at the operational level is referred to as the Joint Operations Area (JOA). The boundaries within which a JOA is defined are contingency or mission-specific, and intended to focus and enhance military activities within that area (NATO 2011, pp. 1-23).

For cyberspace operations to be integrated as part of a NATO operation, therefore, it would be necessary to identify the cyberspace JOA – the specific elements of what one might call the total geography of cyberspace relevant to the operation. This would certainly include the communications and information systems used to carry out the command and control of operational forces. In addition, it would include all supporting systems, including intelligence, logistics, medical, civil-military cooperation, information and psychological operations, and force protection, as well as all long-haul reach-back communications and any systems and networks in the static infrastructure at home that support the deployed operation. It might include commercial systems, given that military forces increasingly augment limited military satellite communications with commercial satellite services. And it would almost certainly include the Internet and any interfaces to it, since the Internet has become the primary medium through which news reports, morale and welfare communications, social media discussions, strategic communications, information exchanges with non-governmental organizations and supporting financial, procurement and transportation arrangements will be conveyed.

A good illustration of the complexity and geographically dispersed nature of cyberspace is offered by the example of one capability likely to be employed in future NATO operations: the NATO Alliance Ground Surveillance (AGS) system. For this one system alone, a complex set of communications and information system assets, including deployed ground stations, military and satellite communications, air-ground tactical data links, air command and control systems, a Main Operating Base in Europe, and dissemination links to national and multinational intelligence analysis centers, is required to provide just part of the overall intelligence, surveillance and reconnaissance (ISR) support to the military operation. And some elements, such as the Main Operating Base, are one-deep resources that cannot be dedicated to a single operation and must always be viewed as strategic assets.

When one then considers the number of similarly complex and distributed systems supporting the NATO and national forces involved in a military operation of even moderate scale, it should become clear that it is difficult to draw clear boundaries that distinguish the area within which a NATO operational commander would have authority to take military decisions from that which is subject to civil or political authorities. As Scott Applegate (2012, p. 192) summed it up: "One difficulty in defining borders in cyberspace is that the physical geography of cyberspace does not even remotely match the logical geography". Indeed, one can argue that, given the heavy use of reach-back capabilities and Alliance and national strategic assets, it would be difficult to define a cyberspace JOA that is purely mission-specific and does not overlap substantially with the strategic Area of Responsibility (AOR) for which the Supreme Allied Commander Europe (SACEUR) is assigned sole responsibility (NATO 2011, pp. 1-23). Because of this, the principles of levels of command and delegation of authority that can be applied within the physical boundaries of a JOA for the air, land and maritime domains, where clear boundaries can be established, may be difficult to implement within a cyberspace JOA. And this is not the only unique aspect of cyberspace as an operational domain.

3. UNDERSTANDING THE NATURE OF CYBERSPACE THREATS

A second challenge in implementing cyberspace as a domain of military operations is the nature of cyberspace threats – in particular, the asymmetries in the relationships between attackers and defenders. As illustrated in the description of a cyberspace JOA above, military operations conducted by NATO and its member nations are hugely dependent upon a complex set of supporting military, governmental and commercial networks and systems. No NATO operational commander can automatically assume that the parties responsible for these networks and systems will allow their use by military cyberspace forces to launch attacks or take other measures that might put them at risk. Attribution can be extremely difficult, or simply impossible. Even with certain attribution, the organizations attacked will often lack the legal authority or appropriate tools to strike back directly or in kind.

But there are other aspects of the asymmetric nature of cyberspace threats that represent complex challenges when considering the use of military forces and capabilities in defense of cyberspace. Cyber attacks are not always immediate in effect, let alone easily attributable. Analogies between cyber and air defense fail: cyber defenders will not be watching incoming attackers on a "cyber radar" and launching cyber weapons in response. Instead, cyberspace defenders are more likely to be sifting through log files and employing sophisticated analytical tools, searching not only to pick up the trail of the attacker but even just to detect what the actual impact of the attack might have been.

Although the trend has been improving in recent years, the fact remains that there is often a significant delay between the initiation of an attack and its detection. According to one recent report, the median number of days between a cyber attack and its discovery was 146, which is well beyond the timescale traditionally associated with tactical operations in other domains (FireEye/Mandiant Consulting 2016, p. 4). As Dr Jan Kallberg (2016, p. 103) wrote recently: "In reality […] cyber-attacks would be over before any leadership understood the strategic landscape". In addition, cyber attackers may not consider themselves obliged to comply with the rule of law, to limit themselves to launching attacks after a formal declaration of war or to confine their attacks to clearly designated military targets. Indeed, they may be able to achieve a desired effect simply by asserting that successful attacks have taken place, or by manipulating perceptions via social media, as has been seen in support of Russian incursions into Ukraine (NATO Strategic Communications Centre of Excellence 2016, pp. 11-12).

If the tactics of cyber attackers may take months to detect, may have effects that are difficult to assess, are problematic to attribute to specific sources or involve exploitation of systems clearly outside military or governmental control, fall outside periods of formally-declared operations, and involve manipulation of social media, then the resources available to an operational commander in theatre are simply inadequate to develop effective responses or mitigations. As with the difficulty of establishing boundaries in cyberspace for NATO operations, these aspects of potential cyber threats suggest that the traditional distinctions between tactical, operational and strategic levels of command that apply in other domains may not be appropriate for cyberspace as a domain. Unlike the kind of threats a NATO operational commander may face in the air, land or maritime domains, in cyberspace there is a good chance that adversaries will undertake attacks against which there are no mature and well-understood responses. Unlike most responses in the other domains, cyberspace responses lack measures of effectiveness that have been established and proven through extensive use in exercises, if not actual operations, and that would allow them to be apportioned within clear rules of engagement at the tactical or operational level. This approach applies as well when considering how to achieve effective mission assurance in cyberspace.

4. UNDERSTANDING MISSION ASSURANCE IN CYBERSPACE

A primary argument made by NATO military authorities in advocating for the recognition of cyberspace as an operational domain was that it would improve mission assurance for NATO's joint military forces in accomplishing their core tasks. As with "cyberspace", NATO has not yet agreed a formal definition of "mission assurance". In its *Mission Assurance Strategy*, the US Department of Defense has defined the term as:

> [a] process to protect or ensure the continued function and resilience of capabilities and assets – including personnel, equipment, facilities, networks, information and information systems, infrastructure, and supply chains – critical to the performance of DoD Mission-Essential Functions (US Department of Defense 2012, p. 1).

What is particularly interesting about this definition is that it positions mission assurance as a supporting consideration to that of actually performing mission-essential functions. This understanding is crucial to ensuring the proper focus of cyberspace as a domain of military operations. According to Colonel William Bryant of the US Air Force (2016, p. 6): "Mission assurance in and through cyberspace is not fundamentally an IT problem, but a mission problem that requires a mission focus and approaches that go beyond what we have come to think of as traditional cybersecurity".

Mission assurance, in Michael Jay Lanham's (2015, p. 24) words, requires acceptance that "bad things will happen to an organization, despite the various avoidance, mitigation, retention and transfer measures in place". Mission assurance in cyberspace focuses on assuring that an organization's mission capability can be maintained not only by preventing degradations but by minimizing effects and orchestrating rapid responses when they do occur (Pritchett 2012, p. iv). Unlike cyber security, which strives to protect all information systems and assets, mission assurance seeks to ensure that the mission can be carried out even if some systems have failed.

One paper concluded that: “mission-critical assets do not have to be perfectly secure; they just have to be secure enough to reliably accomplish their mission” (Peake, Underbrink & Potter 2012, p. 30). A military operation with a strong level of cyber mission assurance is one capable of continuing its mission-essential functions even in the presence of cyber attacks, not one that simply aims to prevent these attacks. As Internet security expert Dan Geer (1998) once phrased it: “The ability to avoid loss never makes up for the ability to absorb loss”.

What cyberspace operators need to focus on, instead of protection and prevention measures, are effects. These should include both the negative effects that disruptions, compromises, outages, exploitations or other degradations in the cyberspace supporting the operation might have, and the positive effects that defensive or mitigation measures might have to ensure effective command and control. And, where appropriate, authorized and made available by contributing nations, the enabling effects that offensive cyberspace capabilities might contribute. However, given the scale and complexity of the cyberspace elements supporting a typical NATO operation, cyberspace operators need to consider the full scope of the communications and information systems involved, not just those deployed in theatre, and all possible threats against them including self-inflicted disruptions. At least until tools, techniques and tactics mature, this may be a task best approached at the strategic level.

Of course, the best time to do this analysis is before the operation begins, through exercises, simulations and training. This is why in recent years the US has made a priority of the exercise of its capabilities to operate effectively in “denied, manipulated, and/or contested cyberspace” (Chairman Joint Chiefs of Staff 2014, p. 3). This is also why it is important to separate cyberspace operations from traditional network operations. A good network operator will always strive to provide the most secure and reliable service possible and will be disinclined to arbitrarily cause outages or disruptions. But a responsible cyberspace operator should always want to test the operational force’s ability to deal with the unexpected when there is still time to learn where the key constraints in the supporting cyberspace are and mitigate the impacts of their disruption or loss. Placing the responsibility for cyberspace operations outside the network operations function improves the ability of cyberspace operators to focus on mission assurance. Without explicit planning and exercising for cyberspace incidents – whether hostile or self-inflicted – the effectiveness of detection and response measures, particularly those that involve multiple organizations or force elements, is significantly undermined (Lanham 2015, p. 50).

NATO has already begun to put a greater emphasis on integrating cyber defense into its military exercise and training program, but as it now implements cyberspace as an operational domain, that emphasis will need to shift from focusing on information assurance to focusing on mission assurance. Exercises should incorporate more scenarios aimed at testing the ability of coalition forces to operate effectively under constrained cyberspace conditions and in the presence of a variety of cyber attacks, including attacks that target strategic elements of Alliance cyberspace, supporting critical infrastructures and even social media. NATO cyberspace doctrine development should also explore techniques and tactics to enable forces to recover more quickly in the event of attacks or other disruptions.

And in implementing cyberspace as a domain, NATO should recognize that, unlike other domains, cyberspace enjoys the blessing (as well as the curse) of being an environment in which attacks and responses to their effects are part and parcel of everyday business for the NATO Command Structure and other elements of the NATO enterprise, for national military and government organizations and for commercial and non-governmental organizations. Every nation in the Alliance has some equivalent to the NATO Computer Incident Response Centre (NCIRC), dealing with cyber incidents on a daily basis. Every one of these incidents is an opportunity to get smarter about the techniques and capabilities of cyber attackers, to better understand how to eliminate or compensate for the vulnerabilities exploited, to improve response mechanisms, and to make the affected organization better able to cope with future attacks. As a recent survey of the arrangements for management of major cyber incidents in a number of European and Asian countries found, expecting "issues of command and responsibility to be resolved during the evolution of the crisis […] will likely have a negative impact on the effectiveness of the response" (Boeke, Heinl & Veenendaal 2015, p. 73). Given the lack of clear boundaries in cyberspace and the high probability that cyber attacks may occur outside the context of approved NATO operations, mission assurance in the cyberspace domain is not just something to worry about when drawing up an operational plan, it is a continuous concern for every organization operating in cyberspace. That is why mission assurance in cyberspace cannot be divorced from the issue of collective defense in cyberspace.

5. THE MILITARY ROLE IN COLLECTIVE DEFENSE OF CYBERSPACE

In recognizing cyberspace as an operational domain, the North Atlantic Council inevitably opens a dialogue over the appropriate role for NATO military commands and forces in the collective defense of cyberspace. This is still a highly-charged issue for some. They recognize that the portion of cyberspace supporting military operations represents only a fraction of the total cyberspace geography upon which their government, commercial organizations and private citizens depend. Some share Martin Libicki's (2012, p. 321) view that "cyberspace is not a warfighting domain".

In committing to the Cyber Pledge at the Warsaw Summit, the Council agreed to enhance the cyber defenses of national infrastructures and networks and recognized the indivisibility of Allied security and collective defense (NATO, Cyber Defense Pledge 2016). Although the Council recognized that "[o]ur interconnectedness means that we are only as strong as our weakest link", it also qualified that co-operative efforts such as "multinational projects, education, training, and exercises and information exchange" were only "in support of national cyber defense efforts". However, the Council did not make a direct connection between the operational domain of cyberspace and any standing role it might play in collective defense outside the context of military operations.

Indeed, perhaps the primary obstacle to acknowledging a role for the NATO Command Structure in the collective defense of cyberspace is the tendency to view this role in terms of analogies with collective defense functions in other domains. NATO maintains, for example, under its Air Component Command, a peacetime collective defense Air Policing mission that safeguard the integrity of the Alliance members' national airspace. NATO's Maritime Command has a standing task to maintain maritime situational awareness across national and international waters. When it comes to the defense of cyberspace, however, some nations are reluctant to entertain the possibility of the military, let alone an international organization such as NATO, controlling a cyber defense force with access to national networks or even with access to information about national cyberspace vulnerabilities. According to Cavelty (2012, p. 151): "Protecting them [critical information infrastructures] as a military mandate is an impossibility and considering cyberspace as an occupation zone is an illusion".

Yet even without such a commitment, the NATO Command Structure plays a significant role in the defense of what are known as the NATO enterprise networks. Allied Command Operations (ACO) funds over 80% of the annual operating costs of these networks and 100% of the cost of operating the NCIRC and other cyber defense capabilities such as the Malware Information Sharing Platform (MISP). In addition, ACO maintains a full-time strategic cyber defense situational awareness capability under a multi-disciplinary team known as Task Force Cyber, part of SACEUR's standing task to provide the Council with military advice on indications and warnings of threats to collective security. Allied Command Transformation (ACT) supports joint cyberspace doctrine development and sponsors Cyber Coalition and other events intended to improve cyber incident response capabilities. And as a high-visibility organization operating on an international scale, NATO represents an attractive target for cyber-attackers interested in making headlines.

There is a natural role, therefore, for the NATO Command Structure to support collective cyber defense through information-sharing and situational awareness. This is not the cyberspace equivalent of air policing. This is not a matter of any NATO organization gaining access to national systems or networks or commanding the application of defensive measures within those networks. Protection of national networks is and remains strictly a matter of national responsibility. Establishing a clear role for the NATO Command Structure in collective cyber defense is a matter of recognizing that the complex and distributed nature of cyberspace – which cuts across all levels of systems and command, from tactical to strategic – and the unique characteristics of cyber threats – which tend to involve timescales, employ techniques, and include targets that go well beyond the domain of traditional military operations – requires an operational approach involving continuous collaboration, rather than simply a capacity to gear up a cyberspace mission assurance capability in the event of a crisis.

Every day, some part of the Alliance experiences attacks or disruptions that provide live scenarios far better than could be constructed in any exercise or simulation. Without a clear task to participate in, which will foster the kind of information-sharing that can both improve collective defense and better prepare the command to address mission assurance in cyberspace, the NATO Command Structure can only maintain a limited cyberspace situational awareness function, organize a small number of exercises each year, and develop its cyberspace operational doctrine in relative isolation.

Sensitivities over the role of an international military organization in cyber defense, however, have left the responsibilities for these functions vaguely and inconsistently defined. The Charter of the NATO Communications and Information Organisation (NCIO) assigns the responsibility for protection of NATO's communications and information infrastructures to the NATO Communications and Information Agency (NCIA) (NATO 2012, pp. 1-12). Although the expectation that cyberspace will be addressed in the context of NATO operations is clear from the decision to recognize it as a domain, the NATO Command Structure has not been assigned specific tasks or authorities to establish formal cyber defense information-sharing mechanisms with equivalent elements in the NATO Force Structure.

The value of information-sharing, under SACEUR's leadership, to improve the state of collective cyber defense cannot be overestimated. A recent study by the RAND Corporation found that "[i]nformation exchange is an indispensable element in the improvement of cybersecurity" (Meulen 2015, p. viii). The NATO Command Structure already plays key roles in maintaining effective information exchange with Alliance national military forces regarding capabilities, readiness, certification, doctrine, training and a wide array of other common interests in the other operational domains. With the recognition of cyberspace as an operational domain, it only makes sense to add cyber defense to this list. NATO and national military cyber defense organizations should maintain a strong partnership, not just on cyber incidents and malware, but also on advancing doctrine, techniques, tactics and procedures and improving collective situational awareness capabilities.

Collective defense can only be improved by encouraging collective cyber defense throughout the Alliance through a healthy, open and lively exchange of information between military cyber defense organizations and, of course, with their civil and commercial counterparts. This should not be considered just a "nice to have", but an essential element of the collective cyber defense strategy. One can hardly imagine how Alliance leaders expect collective cyber defense to be improved by remaining mute on the question of the military's role. As Thomas Rid (2012, p. 29) stated: "The world's most sophisticated cyber forces have an interest in openness if they want to retain their edge, especially on the defensive. […] Only openness and oversight can expose and reduce weaknesses in organization, priorities, technology, and vision".

6. CONCLUSION

Recognition of cyberspace as an operational domain certainly presents considerable challenges for NATO. The Alliance must achieve a more accurate understanding of the scale and complexity of cyberspace, particularly as it applies to the support of NATO operations, and its inherent reliance on strategic information assets and even critical infrastructures. It must acknowledge the nature of cyber threats, which often involve timescales, targets or effects that can only be addressed effectively at the strategic level. In their assessment of the implications of recognizing cyberspace as an operational domain, NATO military authorities identified improving mission assurance for joint operations as a key benefit, and this perspective will help to shift the focus of cyberspace operators from information assurance and cyber security. But the best way to improve NATO mission assurance in cyberspace is to recognize the opportunities presented by the cyber incidents dealt with on a daily basis by cyber incident response centers and the networked organizations they support. And this means to recognize the legitimate role for the NATO Command Structure to act as an enabler for collective defense in cyberspace through partnership and information-sharing.

REFERENCES

Applegate, S. D. (2012). The Principle of Maneuver in Cyber Operations. *2012 4th International Conference on Cyber Conflict* (pp. 183-195). Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence.

Boeke, S., Heinl, C. H. & Veenendaal, M. (2015). Civil-Military Relations and International Military Cooperation in Cyber Security: Common Challenges & State Practices Across Asia and Europe. *7th International Conference on Cyber Conflict* (pp. 69-80). Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence.

Bryant, W. D. (2016, Winter). Mission Assurance through Integrated Cyber Defense. *Air & Space Power Journal*, 5-17.

Cavelty, M. D. (2012). The Militarisation of Cyberspace: Why Less May Be Better. *2012 4th International Conference on Cyber Conflict* (pp. 141-153). Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence.

Chairman Joint Chiefs of Staff. (2014). CJCS Notice 3500.01, 2015-2018 *Chairmans Joint Training Guidance*. Retrieved from US Department of Defense, Chairman Joint Chiefs of Staff: www.dtic.mil/cjcs_directives/cdata/unlimit/n350001.pdf.

FireEye/Mandiant Consulting. (2016). *M-Trends 2016*. Retrieved from FireEye Corporation: https://www2.fireeye.com/rs/848-DID-242/images/Mtrends2016-NEW.pdf.

Geer, D. (1998). Risk Management is Where the Money Is. *Risks-Forum Digest*, 20(6). Retrieved from http://cseweb.ucsd.edu/~goguen/courses/275f00/geer.html.

Hayden, M. V. (2011, Spring). The Future of Things Cyber. *Strategic Studies Quarterly*, 5(1), 3-7.

Kallberg, J. (2016). Strategic Cyberwar Theory–A Foundation for Designing Decisive Strategic Cyber Operations. *The Cyber Defense Review*, 1(1), 113-125.

Kuehl, D. T. (2009). From Cyberspace to Cyberpower: Defining the Problem. In F. D. Kramer, S. Starr, & L. K. Wentz, *Cyberpower and National Securit*y. Washington, DC: National Defense University Press.

Lanham, M. J. (2015). *Rapid Mission Assurance Assessment via Sociotechnical Modeling and Simulation* (Published Dissertation). Pittsburgh, PA: Institute of Software Research, Carnegie Mellon University.

Libicki, M. C. (2012). Cyberspace Is Not a Warfighting Domain. *I/S: A Journal of Law and Policy for the Information Society*, 8(2), 321-336.

Meulen, N. v. (2015, October 14). *Investing in Cybersecurity*. Retrieved from Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC): https://english.wodc.nl/onderzoeksdatabase/2551-investeren-in-cyber-security.aspx.

NATO. (2011). Allied Joint Publication 3. *Allied Joint Doctrine for the Conduct of Operations*.

NATO. (2012). *C-M (2012)0049, Charter of the NATO Communications and Information Organisation*. Retrieved from NATO Communications and Information Agency.

NATO. (2016, July 9). *Cyber Defence Pledge*. Retrieved from NATO: http://www.nato.int/cps/en/natohq/official_texts_133177.htm.

NATO. (2016, July 9). *Warsaw Summit Communiqué*. Retrieved from NATO HQ: http://www.nato.int/cps/en/natohq/official_texts_133169.htm.

NATO Strategic Communications Centre of Excellence. (2016). *Social Media as a Tool of Hybrid Warfare*. Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Peake, C., Underbrink, A. & Potter, A. (September-October 2012). Cyber Mission Resilience: Mission Assurance in the Cyber Ecosystem. CrossTalk, pp. 29-33.

Pritchett, M. D. (2012). *Cyber Mission Assurance: A Guide to Reducing the Uncertainties of Operating in a Contested Cyber Environment*. Retrieved from Air Force Institute of Technology: http://www.dtic.mil/dtic/tr/fulltext/u2/a563712.pdf.

Rid, T. (2012). Cyber War Will Not Take Place. *Journal of Strategic Studies*, 35(1), 5-32.

Singer, P. W., & Friedman, A. (2014). *Cybersecurity and Cyberwar: What Everyone Needs to Know*. New York City, NY: Oxford University Press.

US Department of Defense. (2012). *Mission Assurance Strategy*. Washington, DC.: Deputy Secretary of Defense.

Withers, P. (2015, Spring). What is the Utility of the Fifth Domain? *Air Power Review, 18*(1), 126-150.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Core Illumination: Traffic Analysis in Cyberspace

Kenneth Geers
Comodo Group
Toronto, Canada

Abstract: The information security discipline devotes immense resources to developing and protecting a core set of protocols that encode and encrypt Internet communications. However, since the dawn of human conflict, simple traffic analysis (TA) has been used to circumvent innumerable security schemes. TA leverages metadata and hard-to-conceal network flow data related to the source, destination, size, frequency, and direction of information, from which eavesdroppers can often deduce a comprehensive intelligence analysis. TA is effective in both the hard and soft sciences, and provides an edge in economic, political, intelligence and military affairs.

Today, modern information technology, including the ubiquity of computers, and the interconnected nature of cyberspace, has made TA a global and universally accessible discipline. Further, due to privacy issues, it is also a global concern. Digital metadata, affordable computer storage, and automated information processing now record and analyse nearly all human activities, and the scrutiny is growing more acute by the day. Corporate, law enforcement, and intelligence agencies have access to strategic datasets from which they can drill down to the tactical level at any moment. This paper discusses the nature of TA, how it has evolved in the Internet era, and demonstrates the power of high-level analysis based on a large cybersecurity dataset.

Keywords: *traffic analysis, malware, cyber operations, geopolitics*

1. INTRODUCTION: TRAFFIC ANALYSIS

Core Internet security protocols, including encryption, protect users from a wide range of threats. However, attackers can use traffic analysis (TA) to defeat almost any level of security precaution, as long as they have visibility and a capability to collect and analyse data. In fact, TA is a necessary precursor to cryptanalysis, and it is where strategic signals intelligence (SIGINT) almost always begins.

Digital TA relies on the examination of network flow and metadata, can be effective even against unbreakable encryption, and is often sufficient to act as a basis for both tactical and strategic intelligence insight. Basic information begins with source and destination addresses, message type, count, timing, frequency, length and other 'externals.' With only these data points, it is possible to deduce a surprising amount of intelligence regarding the communicants, including their identity, location, movement, behaviour, capabilities, intentions and morale.

From a military perspective, TA can determine chain of command, order of battle (OB), security level, indications and warning (I&W) and more. With such intelligence in hand, it may then be possible to jam, censor, or deceive an adversary.

TA even has some advantages over having complete access to the adversary's unencrypted, plaintext messages, including:

- speed – TA can be automated;
- cost – content analysis and language translation capabilities are expensive disciplines for which there is never enough expertise or manpower;[1] and
- surprise – TA yields many discoveries, some of them unexpected.

TA and counter-TA have been used throughout political, military, and economic history. Consider three famous examples from World War II: prior to its surprise attack on Pearl Harbour, the Japanese navy broadcast false in-port communications to fool American eavesdroppers; in Operation Quicksilver, the Allies played a similar game against German intelligence to divert Hitler's attention from Normandy; and in helping to crack the Enigma machine, Alan Turing discovered weaknesses in Nazi communications, first by TA, then by cryptanalysis.

SIGINT and Electronic Warfare (EW) have always been key elements of military planning and operations, and TA has always been a precursor and a critical piece of both SIGINT and EW. For example, in a military setting, the extreme secrecy surrounding submarine deployments means that boat captains must balance the benefit of connecting to the chain of command with the risk of being located by adversary vessels using direction finding. Thus, submarines must follow the strictest communications standards and procedures.

Traffic analysis can be performed on anything, from pizza deliveries at the Pentagon to noting the tail numbers of suspicious airplanes. Law enforcement and counterintelligence routinely tally the electricity use and bill payment methods of suspected criminals and spies. In contrast to military and intelligence agencies, civilian enterprises and individual citizens can be particularly vulnerable to TA, as they may not have adequate (or any) operations security (OPSEC) training or experience.

Any smart TA researcher with a large digital dataset has enormous possibilities, including for tracking advanced persistent threat (APT) actors, and even predicting certain future events. There is a well-understood cyber 'kill chain': phishing, for example, is both a principal component of, and a precursor to, most significant cyber attacks. Likewise, the prepositioning of hacker tools on industrial control systems may be considered a latent national security threat.

[1] Once a target has been selected at the strategic level for closer scrutiny at the tactical level, more expensive resources can be used: anything from computer hacking to human surveillance and physical destruction.

TA can place these cyber incidents in a wider context, and make them understandable even to non-technical decision-makers.

TA enhances traditional malware analysis. Take Stuxnet or the Democratic National Committee hack; no one researcher, company, or even nation provided us with our understanding of these attacks. Strategic analysis and insight stemmed from the work of hundreds of researchers in different countries – most of whom did not know each other personally – performing not only tactical malware analysis but strategic TA as well. TA can help chip away at the attribution problem, or the challenge of solving the problem of the 'last hop', by collating data points from many different types of sensors that exist in disparate legal jurisdictions, and can be used to discover operations by even the world's most secretive three-letter agencies.

2. LITERATURE REVIEW

One of the author's goals in this paper is to bring the significance of TA to a wider audience, as this topic has historically been confined to relatively small circles of specialised analysts. Another goal is to demonstrate just how quickly large digital datasets which encompass communications from around the world can be leveraged for both tactical and strategic insight.

This section references several dozen papers that discuss TA in a wide variety of settings. For example, TA has a strong history in economics, such as for the relief of road congestion [1] [2]. Digital TA dates from at least the early 1990s [3], with one company offering commercial counter-TA solutions as early as 2000 [4].

Digital TA has been performed at every scale, from the size of a computer chip, where the evaluation is mostly physics [5], to botnet research that encompasses software behaviour characterisation, honeypots, virtual machines, counterintelligence [6] and natural language analysis in chat rooms as a form of Turing Test [7]. TA covers both the temporal [8] and spatial examination of data [9].

Some authors have written academic TA overviews that blend historical viewpoints and modern information technology [10] to demonstrate the evolution of the concept. TA is a huge topic, with attacks ranging from time correlation to statistical disclosure and *a priori* knowledge [11]. But the basic idea of TA involves data collection, organising information into network flows [12] and putting it all together into an intelligence framework [13].

Counter-TA is also a mature discipline. Research has focused on how best to protect the location and operational logic of base stations [14, 15, 16], how to disguise network protocols [17], how to falsify traffic, how to move secretly between communications channels [18], how to hide communications within public key cryptography [19], how to pad traffic, how to reroute messages and how to simulate network entropy [20]. Alas, many of these solutions will depend on the quality of the technical personnel and the size of the IT budget.

There are many reasons why counter-TA fails. Effective TA tools can be freely downloaded from the web [21], and highly intrusive TA can be performed remotely. Encrypted tunnelling of HTTP traffic, for example, is vulnerable to TA attacks on time and bandwidth that can identify who you are and where you are going on the Internet [22]. The myriad ways in which TA is possible means that cyberspace is, ultimately, tough terrain for both privacy and human rights [23].

Digital TA begins at the connection level[2] [24], and covers every network protocol, from ADSL [25], to SNMP [26], IPTV [27], peer-to-peer (P2P) [28], DNS, and HTTP [29]. All Internet Service Providers are well-positioned to perform TA against the gamut [30], and as humans begin to live in virtual worlds, TA research has followed them, from Second Life [31, 32] to World of Warcraft [33, 34], and more [35, 36]. For white hat TA, part of the goal is to discover whether another online character is a real human or a game bot [37].

What all this means is that so-called 'anonymous' communications are not as secure as one might think, and when users flock to promising new systems, attackers will turn their sights in that direction [38]. Tor, for example, is a low-latency framework that is considered secure enough for normal web browsing, but likely insecure against TA from a strategic adversary such as a nation-state. These same TA strategies and tactics also work against covert channels [39].

Finally, while it is true that the volume of modern digital communications would seem overwhelming to any analyst, a variety of hardware [40] and software tools [41] have been created for big data analysis, including both licensed and open source network visualisation tools [42, 43].

3. TRAFFIC ANALYSIS 2.0

Modern information technology (IT), including the convergence of most communication streams over the same digital networks, has transformed TA. Due to the ubiquity of computers, and the interconnected nature of networks, cyberspace is now a reflection of all human affairs. Everything from politics to romance, business to crime, and espionage to military invasions, can be seen by anyone who has network access and the knowledge to translate Internet protocols into human language.

There are myriad forms of computer hardware and software today, but a basic requirement for interoperability is that most of them must use the same network 'stack', which in turn makes them vulnerable to capture and analysis by eavesdroppers, even when message content is encrypted. For the most part, Internet routing is transparent, and despite the astonishing number of communication devices on the Internet, there are many hardware and software tools that can intercept, store, process and analyse captured data.

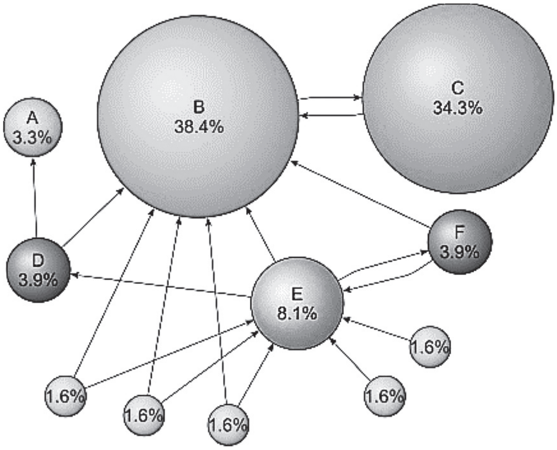
TA is effective at small and large scale. At the micro level, one can identify specific devices[3] and software configurations; it is also possible to map networks by pinging and probing

[2] E.g., source IP address, destination IP address, source port number and destination port number.

[3] Eavesdroppers can even remotely analyse the 'drift' of a digital clock. Web servers often perform timing measurements in order to make inferences about site visitors.

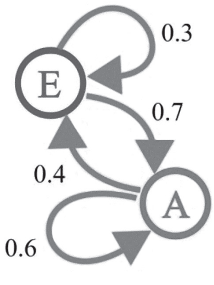
unfamiliar network space.[4] At the macro level, strategic datasets and algorithms are used on a daily basis. For example, the Google PageRank algorithm (see Figure 1) can evaluate the relative significance of web pages across the Internet, by counting the number and quality of hyperlinks pointing to a given site.

FIGURE 1. MATHEMATICAL PAGERANK MODEL (SOURCE: WIKIPEDIA)



Markov chain models (see Figure 2) are used to predict where Internet users will go next, based upon where the user currently is, and the use of probability theory.

FIGURE 2. MARKOV CHAIN MODEL (SOURCE: WIKIPEDIA)



These and similar TA algorithms are regularly exploited to identify users and predict what they will buy and for whom they will vote.[5]

Core Internet security protocols are also vulnerable to TA. Secure Shell (SSH), which encrypts all communication streams over unsecured networks, is vulnerable to attacks that simply count the number and timing of network packets.[6] Transport Layer Security (TLS) and its predecessor

[4] Nmap (Network Mapper) can identify operating systems, open ports, running services and more.

[5] Social network analysis was famously used to locate Saddam Hussein via his tribal and family links.

[6] Keyboards have fixed layouts that sometimes allow attackers to guess passwords based on the time it takes for human fingers to move between individual keys.

Secure Sockets Layer (SSL) do not obscure communications in a way that conceals their message sizes; this allows TA to discover which webpages a user has accessed.

A. Metadata

All digital communications generate activity records in the form of log files,[7] which can, to a large degree, indicate what takes place in traditional geopolitical space.[8]

Therefore, TA can illuminate physical, terrestrial activity by examining individual log files. However, it is far more effective to concatenate log files from numerous sources, which is akin to having multiple witnesses testify in a court trial. In this way, even the most secure networks are vulnerable to TA. For example, it often happens that a target is invisible on one layer in the network stack, but not on another.

Despite the enormous volume of metadata, TA specialists can automate much of their work, including by using advanced mathematical models that discover previously unknown correlations and anomalies between any two objects in a large dataset. Analysts may seek any number of interesting network patterns, but they often include political, military and economic intelligence.

Again, TA does not demand the availability of any messages in their plaintext (unencrypted) forms. For automation purposes, human conversations are in fact notoriously difficult for computers to understand.[9] Message content is even tricky for human analysts to follow, since many words are culture- and context-specific, and associated with events with which only the communicants are familiar.[10]

In the event that an adversary has employed so many special security protocols that they are nearly invisible, it is also possible that such extreme measures will themselves be discovered as anomalous, and only serve to pique the interest of third parties and raise the level of TA to which any such network is subjected.

B. Geolocation

The implications of digital TA are serious: if researchers or attackers can identify you, they also might be able to find you in the real world. Historically, TA employed radio frequency (RF) direction finding, with a line of bearing to a transmitter. However, today there are newer technologies such as the Global Positioning System (GPS) and Time Difference of Arrival (TDOA), which triangulate network users via satellite and cell phone towers.

Many browser-based tools can plot digital communications on a real-world map.[11] The most common method is to look up an Internet Protocol (IP) address in a Whois database, where

[7] Eavesdroppers may find interesting logs anywhere, from browser caches to web server and router logs.

[8] It is beyond the scope of this paper, but computing devices also emit physical signatures that can be measured and captured with the right equipment.

[9] The difficulty of passing the 'Turing Test,' for example, highlights how difficult it is for computers to appreciate human language in context.

[10] This is why, for example, that any type of censorship is difficult to perform accurately, and almost perforce leads to over-censorship.

[11] Standards include ISO 3166, ISO/IEC 19762-5:2008, FIPS, INSEE, Geonames, IATA, ICAO, American National Standards Institute (ANSI) Codes, WOEID (Where on Earth IDentifier), NAC Locator, geotagging, location-based services, mobile phone tracking, W3C Geolocation API, geolocation video and more.

anyone can see the IP's registrant, physical address, associated domain names, business name and more.[12] Sophisticated geolocation can be performed on MAC addresses, RFID, embedded code, Wi-Fi positioning systems, device GPS coordinates, archival tags, microchip implants, data storage tags, and more. Using social media, it is also possible to geolocate images, videos, and comments, most of which are self-disclosed. Finally, online search engines and mapping software such as Google Earth can be used to refine and display geo-coordinates.

A recent trend has been the collaborative efforts of cybersecurity researchers and firms worldwide to investigate attacks using a wide range of the free tools described above. For example, Bellingcat and Vice News have tracked Russian military forces in Ukraine; and international, crowdsourced efforts have reverse-engineered major cybersecurity incidents such as Stuxnet and the 2016 attack on the US Democratic National Committee.

C. Attribution

In cyber defence analysis, the biggest challenge is typically attribution, or determining the true source of an attack. This is also known as the problem of the 'last hop', and refers to the known IP address with which a hacker interacts with a victim computer, for reconnaissance or exploitation purposes. As a rule, it is a compromised computer that lies within a broader attack infrastructure, and which the attacker uses as a temporary stepping stone across the Internet.

Strategic TA is one of the primary ways in which cyber defenders can correlate attack data, by connecting disparate communication streams from logs collected at different points on the Internet. This may be done passively by analysing the log files, or actively, by, for example, creating honeypots that place homing beacons on stolen files that can later call home and de-anonymise the attackers. Similar measures can be used against so-called 'anonymous' communication channels like Tor, given that the maximum latency of human communication is quite bounded.

In light of the power of strategic TA, there is little reason for Internet users to imagine that their online activities will remain private forever. Eventually, it must be assumed that disconnected communication segments will be reassembled into one complete stream. However, practically speaking, it is important to note that human communications are not random, and on many occasions they can even be predicted, either manually or automatically. With such intelligence in hand, eavesdroppers, criminals, and spies may already be lying in wait.

[12] In many cases, attackers can hide behind fairly opaque IP ranges, but Whois is a good place for cyber defenders to start, and it is usually possible to lodge an abuse complaint to a Point of Contact listed here.

4. TRAFFIC ANALYSIS IN PRACTICE

TA is a strategic tool that can overcome many tactical defences, including encryption and covert channels. This is because cyberspace is a ubiquitous medium in which there are many ways that an eavesdropper can piece together otherwise obscure relationships and activities. At the very least, it is usually possible to detect that some type of communication is taking place, at which point an analyst can begin to isolate and evaluate specific data.

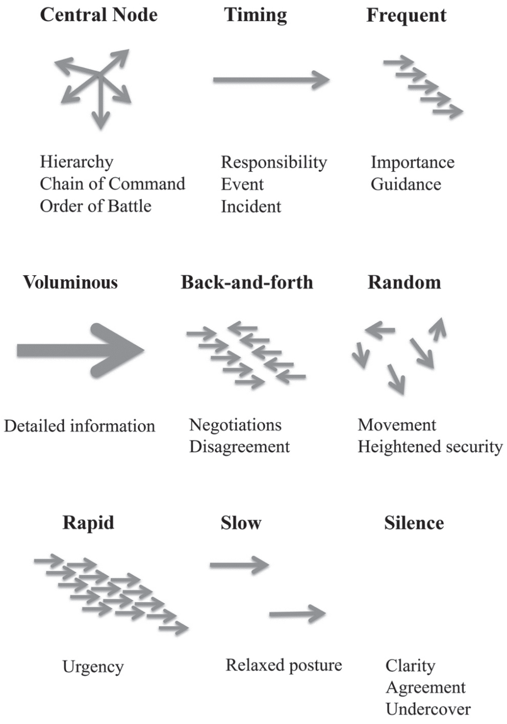
TA is both science and art. First, an analyst must collect sufficient data, from which they will develop a baseline for what appears to be normal traffic. Then the analyst seeks interesting patterns to examine further, often in the form of correlations and anomalies. In a sense, TA is analogous to speedreading, in that it provides a summary of a large volume of information from which the analyst can drill down and make intelligent discoveries.

Over time, TA researchers can solve even highly challenging problems, such as finding stealthy insiders, advanced cyber criminals and even Advanced Persistent Threat (APT) or nation-state actors. At this point, aggressive action may take place against an adversary, before the adversary is even aware that he or she has been discovered.

The list below (and in Figure 3) shows some of the ways in which communication patterns might be used to give away real-world activities.

- Central node
    - Hierarchy, chain of command, order of battle
- Timing
    - Responsibility for an event or incident
- Frequent
    - Importance or guidance
- Voluminous
    - Detailed information
- Back-and-forth
    - Negotiations or disagreement
- Random
    - Movement or heightened security
- Rapid
    - Urgency
- Slow
    - Relaxed posture
- Silence
    - Clarity, agreement, or undercover

FIGURE 3. TRAFFIC ANALYSIS: COMMON PATTERNS



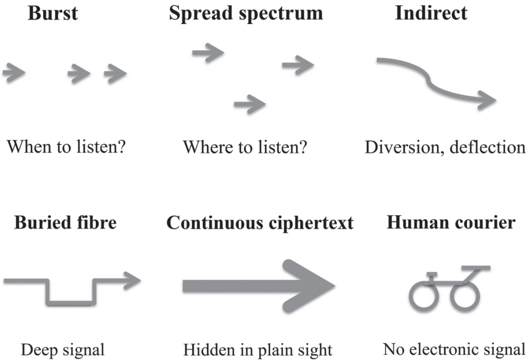
A. Counter-Traffic Analysis

Historically, governments and militaries have defended against TA in many ways, especially by altering network traffic. The list below and Figure 4 summarise some of the most common methods.

- Bursts
  - So the attacker does not know when to listen
- Spread spectrum
  - So the attacker does not know where to listen
- Indirect routing
  - Diversion, deflection, appears intended for another party
- Buried fibre
  - Deep signal so the attacker cannot hear
- Continuous ciphertext
  - So the message is hidden in plain sight
- Human courier
  - So there is no electronic signal to capture.[13]

[13] In 1998, a DARPA Challenge was issued, which recommended traffic padding and rerouting communications through long alternative network paths.

FIGURE 4. COUNTER-TA STRATEGIES AND TACTICS



The basic goal is to minimise exposure so that attackers cannot sense, monitor, analyse, jam, manipulate, or otherwise react to sensitive communications. However, high-level security can be impractical, as it is expensive, requires a determined effort to maintain and often attracts increased scrutiny. In general, only governments and large corporations can afford it.

On the Internet, there are many resources to defend against online attacks including TA: from Mixmaster, to Mixminion, Java Anon Proxy, and Tor, which obscure data such as email headers, web caches and network routing. With hackers, Snowden and Big Brother so often in the news, many researchers feel that developing digital security products, including to protect against TA, is a public good. For example, it is easy to understand that some levels of anonymity and privacy are needed for the proper functioning of democracy. That said, it is also clear that defence against strategic TA is difficult in the short run and may be impossible eventually.

B. Ethics and TA

In the age of digital society and electronic government, the future of TA has profound implications for the world. Anyone with access to network logs has the capability to perform TA; however, governments, including law enforcement and intelligence agencies,[14] possess the capability to conduct strategic TA at will. Governments typically have 'backdoor' relationships with telecommunications providers, and businesses are mandated to retain logs far longer than typical business needs require.[15] Finally, commercial firms and advertising agencies have either collected or purchased granular TA for targeted purposes including tailored advertising.

At the international level, TA is fraught. Technology always outpaces policy and law. However, between nation-states, legal harmonisation is especially difficult in an environment where investigations impinge on another state's sovereignty and governments have become addicted to cyber espionage. The concern for privacy, democracy, and human rights is not uniform across the planet. When does routine surveillance become Big Brother? There is no doubt that decentralisation, privatisation and the fragmentation of telecommunications, including the

[14] Articulating a rationale is easy: crime, espionage, and terrorism.

[15] In the US, businesses have been required to maintain logs since September 2007 and Internet Service Providers (ISP) since March 2009.

creation of innumerable online web services that incorporate technologies such as Voice over Internet Protocol (VoIP), have complicated law enforcement and its jurisdictions; however, it is equally true that many governments have a vested interest in knowing too much about their citizens' private lives.

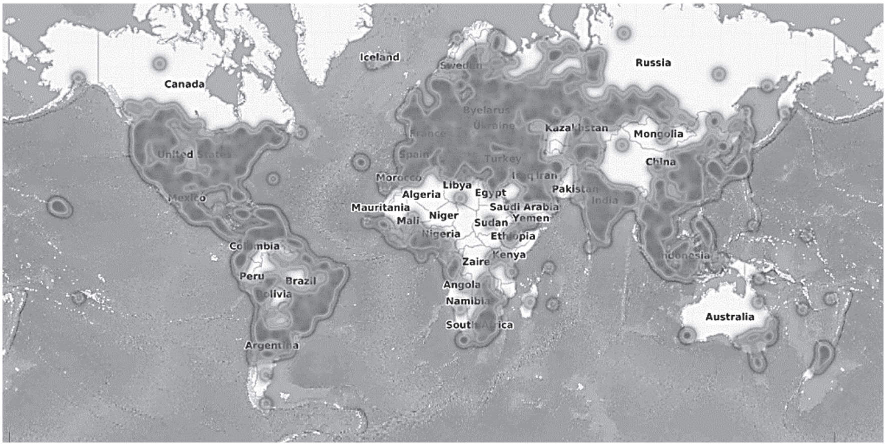
5. DATA ANALYSIS

In this section, the author will demonstrate some basic TA concepts against a large malware dataset, in order to show how quickly any analyst – for political, military, business, or personal reasons – can deduce an endless variety of real-world activities from computer log files.

Several layers of malware analysis will be presented, from the strategic to the tactical level.

Figure 5 displays a heat map of nearly a billion rows of malware data from the first six weeks of 2017.[16] In this dataset, nearly all 253-country code top-level domains (ccTLD) are represented.[17] The dark areas indicate where the author's firm detected malware; the white portions of the map have little malware due to their light concentrations of human settlement.

FIGURE 5. WORLD MAP OF DETECTED MALWARE, EARLY 2017

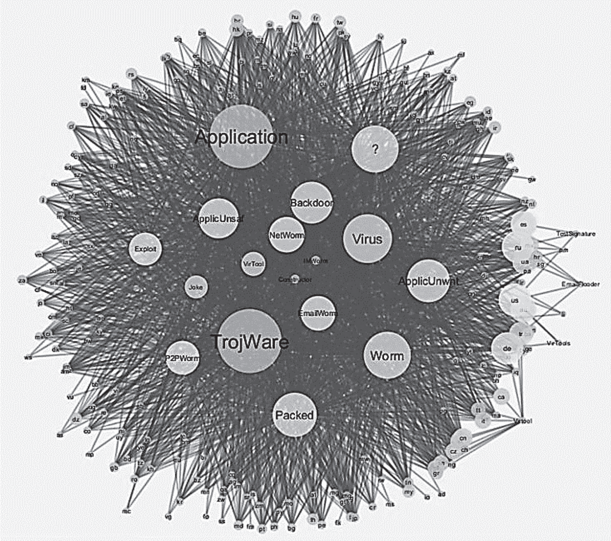


However, the heatmap is just a precursor to TA, which is more about connections and relationships. So, let us take a different view of the data. In Figure 6, the heatmap has been turned into a network chart, which shows a more logical relationship between malware categories and nation-states.

[16] The author's cybersecurity firm analyses over 10 million potential pieces of malware every day, manages over 85 million cybersecurity software installations and works with hundreds of thousands of business customers and global partners around the world.

[17] There are currently 250 ccTLDs, which are Internet top-level domains reserved for countries, sovereign states, or dependent territories. All ASCII ccTLD identifiers are two letters long, and all two-letter top-level domains are ccTLDs.

FIGURE 6. NETWORK MAP: MALWARE TO COUNTRY



Here, we can see that cyberspace really is to some degree a borderless domain in which attackers and malicious code move seamlessly between administrative and legal jurisdictions. The large circles represent malware categories, such as viruses and worms, while the smaller circles are the affected nation-states, identified by their ccTLDs.

TA is all about context. So, let us look at a more detailed description of the malware represented in Figure 6. In Figure 7, below, we can see that different types of malware affect different types of targets. The countries are ranked according to the ratio by which the ‘backdoor’ and ‘worm’ malware categories affected them in the first six weeks of 2017.

FIGURE 7. MALWARE COMPARISON: BACKDOORS VS. WORMS

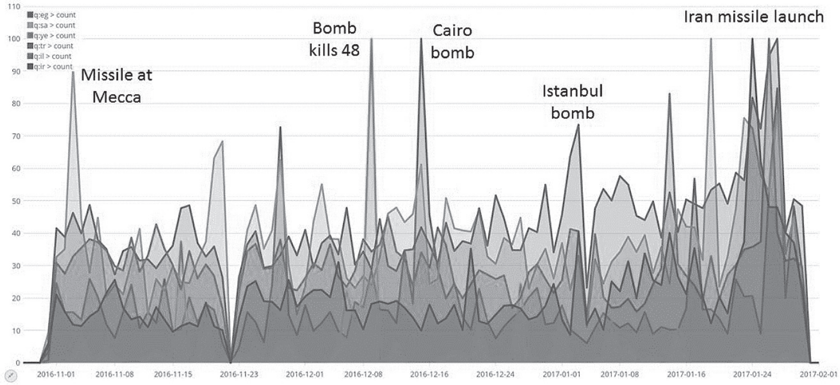
Country	Backdoor Ratio Rank	GDP Rank (IMF)
Kuwait	1	6
Belgium	2	25
UAE	3	9
Portugal	4	47
Spain	5	34
Hong Kong	6	12
Iceland	7	18
Singapore	8	4
Bahrain	9	17
Madagascar	10	179

Country	Worm Ratio Rank	GDP Rank (IMF)
Nigeria	1	131
Ethiopia	2	166
Congo	3	124
Somalia	4	N/A
Maldives	5	84
Rwanda	6	169
Philippines	7	120
Bangladesh	8	141
Yemen	9	158
Moldova	10	135

In Figure 7, we see that backdoors tend to afflict richer countries, which may be better protected against random attacks, and require more targeted strategies and tactics on the part of the attacker. The latter category, worms, afflicted poorer socioeconomic countries, which may run older, outdated, and therefore unsupported software, and are vulnerable to random digital attacks.

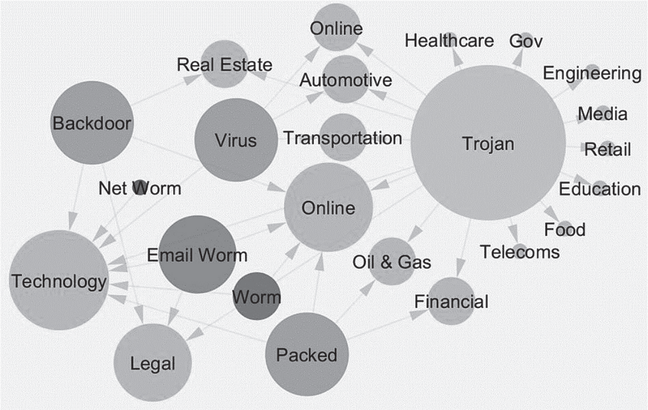
Time is the best friend of a TA researcher. Human conversations are necessarily confined within the boundaries of human patience. Figure 8, below, displays the overall number of malware incidents the author's firm detected in Egypt, Saudi Arabia, Yemen, Turkey, Israel and Iran during the first six weeks of 2017. Above the highest spikes for each country, the author has placed a key event from its national news, which took place on or about the same day. The malware incidents and geopolitical events are not necessarily related, but they may indicate law enforcement, intelligence, hacktivist, or criminal actions in cyberspace that may be correlative.

FIGURE 8. TIMELINE OF MALWARE ACTIVITY WITH GEOPOLITICAL OVERLAY



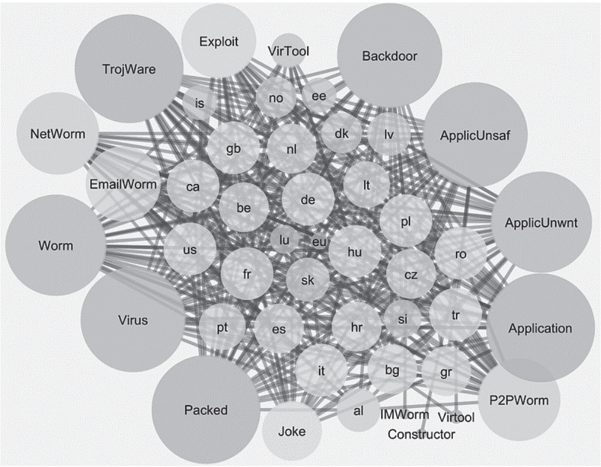
With TA, it is important to apply as much real-world logic as possible to an otherwise opaque dataset. As seen above, malware authors tailor code to a variety of targets, and not every economic sector, or 'vertical', is affected by every type of malware. In Figure 9, the author has paired his firm's top malware types against their target verticals, which highlight two quick conclusions: 1) trojans can be found in every vertical, and 2) attackers are using every type of malware to target the technology sector. This latter conclusion is not surprising, as the technology sector holds the keys to the virtual kingdom of cyberspace; in other words, compromising a particular software, website or protocol can lead to the compromise of all who use it.

FIGURE 9. NETWORK CHART: MALWARE TYPES VS. VERTICALS



Since this paper is written for the NATO CCD COE International Conference on Cyber Conflict (CyCon), it is useful to go back to the malware-to-country network chart in Figure 6, and delete the non-NATO countries (see Figure 10, below). A TA researcher can analyse the entire world (extrapolating from a sample dataset), or choose any part of it for dissection. Below, we can see at a glance that all NATO countries suffer from nearly all categories of malware.

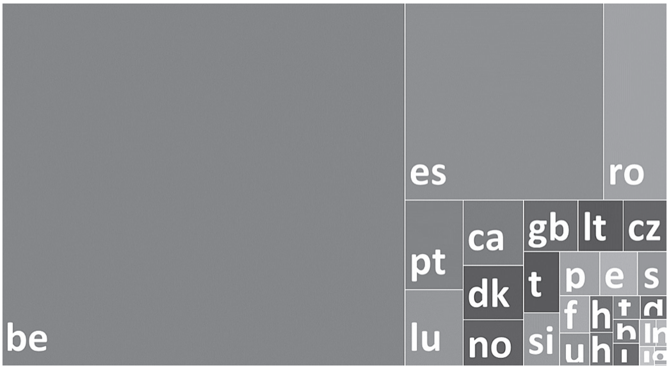
FIGURE 10. MALWARE MAP: NATO COUNTRIES



However, while the data in this network chart has been greatly simplified, it is still not granular enough. Therefore, I created an index for all of the NATO countries, based on the prevalence (by ratio) of each type of malware. In general, the malware-to-country pairings had somewhat similar profiles.

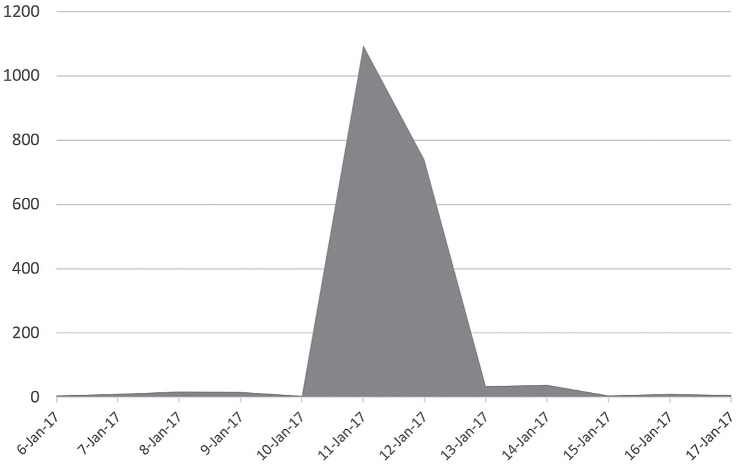
However, close TA almost always yields anomalies to investigate further. In this case, Belgium, which had a malware ratio that ranked near the bottom in nearly every malware category, scored a surprising first place in the 'backdoor' category. Figure 11, below, clearly shows Belgium's domination of this subset of the malware data.

FIGURE 11. BACKDOOR DETECTIONS IN NATO COUNTRIES: EARLY 2017



Next, we should find out precisely when the backdoor detections took place. The timeline in Figure 12 shows this quite clearly: 10-13 January 2017. At this point, we do not know what caused the sudden increase in backdoor detections. It is possible that new security signatures simply found older, previously installed malware. It is also possible that there was a targeted attack against one or more enterprises in Belgium, and that the backdoors were installed with the aid of phishing, a worm, social engineering, or the unwise installation of a malicious application.

FIGURE 12. TIMELINE: BACKDOOR DETECTION IN BELGIUM



At this stage, TA has successfully taken the analyst from a strategic to a tactical level. However, at this point, there are still a variety of ways for a researcher or an investigator to move forward. For example, this information could simply be given to network system administrators for device isolation, digital forensics, software patching, reinstallation and further attack mitigation. Alternatively, investigators may keep this digital TA quiet (leaving the backdoors open temporarily), and begin a process of real-world correlation, in the hope of ensnaring a sophisticated adversary. Such an investigation would include asking hard questions such as: What else happened during this time period that may shed light on the malware's ultimate purpose? If the victim was a business, were there any important trade deals happening at the time? If the victim was a government, did the sharp spike occur just prior to an election, or a national security incident?

Digital TA can inspect log files for any kind of correlation, of a political, military, criminal, business, or personal nature. With enough data in hand, sometimes gathered over many years of painstaking intelligence collection, TA can even help to solve the attribution problem, or that of the anonymous hacker. There are many prominent historical examples, including The Cuckoo's Egg, Moonlight Maze, Stuxnet, Sony, and the Democratic National Committee (DNC).

In a geopolitical context, no stone will be left unturned. TA will not only encompass the temporal, spatial, directional, and logical character of network traffic, but will incorporate intelligence from other domains as well, such as human intelligence (HUMINT), signals intelligence (SIGINT), and open source intelligence (OSINT).

6. CONCLUSION

Historically, TA has been used to circumvent a wide range of core security protocols including encryption. Businesses use TA for market research and advertising; governments collect foreign and domestic intelligence; researchers analyse countless streams of data for academic papers. As we grow more dependent on the Internet – and give IP addresses to everything from toasters to the brakes on our cars – the power of IT will strengthen, magnify, and amplify TA as never before.

This paper has sought to bring TA, especially its digital version, to a wider audience, by describing not only its famous achievements during World War II, but how modern computer log files are essentially a record of all human activity, and can be mined for virtually any kind of intelligence value, from the strategic to the tactical level. Because cyberspace is merely a reflection of human affairs, all major geopolitical events, from elections to invasions, have digital analogues that are just waiting to be discovered.

TA has limitations. Computer log files can provide convincing evidence of real world activity, but once the analysis is complete, traditional investigative practices, such as physical (and network) forensics must complement TA. For example, even the most famous cyberattack case studies, from Moonlight Maze to Estonia and Sony to the DNC, have remained mired in the 'attribution' controversy for years, despite the fact that many analysts believe a preponderance of evidence, including TA, points to a guilty party. At the end of the day, TA is just one tool in a larger toolbox, but it can serve to complement other, more traditional, tools in striking ways.

Looking forward, TA is an established discipline, but research gaps will continue to remain due to the rapid and dynamic evolution of all things IT. Future research should investigate the effect of cloud computing, autonomous systems, artificial intelligence (AI) and more. More analysis of the legal and ethical aspects of TA, especially given that there is only one Internet and one cyberspace which encompasses every jurisdiction and sovereignty on Earth, is also long overdue.

REFERENCES

[1] M. Fathy and M. Y. Siyal. 'An image detection technique based on morphological edge detection and background differencing for real-time traffic analysis.' *Pattern Recogn. Lett*. 16, 12, 1321-1330. December 1995.

[2] L. Wischhof, A. Ebner, and H. Rohling. 'Self-organizing traffic information system based on car-to-car communication: Prototype implementation.' *International Workshop on Intelligent Transportation (WIT)*. Hamburg. March 2004.

[3] Chevalier and L. M. Wein. 'Scheduling Networks of Queues: Heavy Traffic Analysis of a Multistation Closed Network.' *Operations Research* 41(4), August 1993.

[4] J.-F. Raymond. 'Traffic Analysis: Protocols, Attacks, Design Issues and Open Problems.' *Zero-Knowledge Systems*, December 19, 2000.

[5] G. Varatkar and R. Marculescu. 'Traffic Analysis for On-chip Networks Design of Multimedia Applications.' *DAC 2002*, New Orleans, Louisiana, USA, June 10-14, 2002.

[6] W. Lee, C. Wang, and D. Dagon (Eds). *Botnet Detection: Countering the Largest Security Threat*. New York: Springer, 2008.

[7] C. Mazzariello. 'IRC Traffic Analysis for Botnet Detection.' *The Fourth International Conference on Information Assurance and Security*, IEEE, 978-0-7695-3324-7/08. 2008.

[8] G. Danezis. 'The Traffic Analysis of Continuous-Time Mixes.' In: D. Martin, A. Serjantov (eds) *Privacy Enhancing Technologies*. Lecture Notes in Computer Science, 3424. Springer, Berlin, Heidelberg. PET 2004.

[9] M. Crovella and E. Kolaczyk. 'Graph Wavelets for Spatial Traffic Analysis.' BUCS-TR-2002-020, Office of Naval Research. July 15, 2002.

[10] G. Danezis and R. Clayton. 'Introducing Traffic Analysis.' Chapter in *Digital Privacy*. Auerbach Publications, ISBN: 9781420052176. January 21, 2007.

[11] N. Mathewson, R. Dingledine. 'Practical Traffic Analysis: Extending and Resisting Statistical Disclosure.' In: D. Martin, A. Serjantov (eds) *Privacy Enhancing Technologies*. Lecture Notes in Computer Science, 3424. Springer, Berlin, Heidelberg. PET 2004.

[12] M.-Sup Kim, Y. J. Won, and J. W. Hong. 'Characteristic analysis of internet traffic from the perspective of flows, Computer Communications.' Vol. 29, Issue 10, 1639–1652, *Monitoring and Measurements of IP Networks*. June 19, 2006.

[13] J. R. Goodall, W. G. Lutters, Rheingans, and A. Komlodi. 'Focusing on Context in Network Traffic Analysis.' *Visualization for Cybersecurity*, 0272-1716/06, IEEE Computer Society. March/April 2006.

[14] J. Deng, R. Han, and S. Mishra. 'Countermeasures Against Traffic Analysis Attacks in Wireless Sensor Networks.' *Technical Report CU-CS-987-04*. December 2004.

[15] J. Deng, R. Han, and S. Mishra. 'Decorrelating Wireless Sensor Network Traffic To Inhibit Traffic Analysis Attacks.' *Elsevier Pervasive and Mobile Computing Journal*, Special Issue on Security in Wireless Mobile Computing Systems, 2, issue 2, 159-186. April 2006.

[16] J. Deng, R. Han, and S. Mishra. 'Intrusion tolerance strategies in wireless sensor networks.' In *Proc. of IEEE 2004 International Conference on Dependable Systems and Networks (DSN'04)*. 2004.

[17] C.V. Wright, S.E. Coull, F. Monrose. 'Traffic morphing: an efficient defense against statistical traffic analysis.' In: *Proc. of ISOC Network and Distributed System Security Symposium (NDSS)*. 2009.

[18] N. Mathewson and R. Dingledine. ‘Practical traffic analysis: Extending and resisting statistical disclosure.’ In *Proceedings of Privacy Enhancing Technologies workshop (PET 2004)*, LNCS. May 2004.

[19] D. L. Chaum. ‘Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms.’ *Communications*, 24 ACM Number 2 0001-0782/81/0200-0084. February 1981.

[20] R. E. Newman, I. S. Moskowitz, Syverson, and A. Serjantov. *Metrics for Traffic Analysis Prevention*. Office of Naval Research. 2003.

[21] M. Qadeer, A. Iqbal, M. Zahid, and M. Siddiqui. ‘Network Traffic Analysis and Intrusion Detection using Packet Sniffer.’ *Second International Conference on Communication Software and Networks*, 978-0-7695-3961-4/10 IEEE. 2010.

[22] K. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton. ‘Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail.’ *IEEE Symposium on Security and Privacy*. 2012.

[23] X. Gong, N. Kiyavash, and N. Borisov. ‘Fingerprinting Websites Using Remote Traffic Analysis.’ *CCS’10*, Chicago, Illinois, USA. ACM 978-1-4503-0244-9/10/10. October 4-8, 2010.

[24] S. Sarvotham, R. Riedi, and R. Baraniuk. ‘Connection-level Analysis and Modeling of Network Traffic.’ *IMW’OI*. San Francisco, CA, USA, ACM l-581 13-435-5. November l-2, 2001.

[25] N. Ben Azzouna and F. Guillemin. ‘Analysis of ADSL traffic on an IP backbone link.’ *France Telecom R&D*, 0-7803-7975-6/03. 2003.

[26] J. Schonwalder, A. Pras, M. Harvan, J. Schippers, and R. van de Meent. ‘SNMP Traffic Analysis: Approaches, Tools, and First Results.’ 1-4244-0799-0/07 *IEEE*. 2007.

[27] T. Silverston, O. Fourmaux, A. Botta, A. Dainotti, A. Pescapé, G. Ventre, and K. Salamatian. ‘Traffic analysis of peer-to-peer IPTV communities.’ *Computer Networks*. Elsevier. 2008.

[28] M.-Sup Kim, H.-Jung Kang, and J. W. Hong. ‘Towards Peer-to-Peer Traffic Analysis Using Flows.’ In: Brunner M., Keller A. (eds) *Self-Managing Distributed Systems*. Lecture Notes in Computer Science, 2867. Springer, Berlin, Heidelberg. DSOM 2003.

[29] C. Rossow, C. Dietrich, H. Bos, L. Cavallaro, M. van Steen, F. C. Freiling, and N. Pohlmann. ‘Sandnet: Network Traffic Analysis of Malicious Software.’ *BADGERS 2011 Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, ACM. Salzburg. 2011.

[30] S. J. Murdoch and Zielinski. ‘Sampled Traffic Analysis by Internet-Exchange-Level Adversaries.’ In *Privacy Enhancing Technologies (PET)*. 2007.

[31] J. Kinicki and M. Claypool. ‘Traffic Analysis of Avatars in Second Life.’ *NOSSDAV '08*, ACM 978-1-60588-157-6/05/2008. 2008.

[32] S. Fernandes, R. Antonello, J. Moreira, D. Sadok, and C. Kamienski. ‘Traffic Analysis Beyond This World: The Case of Second Life.’ *NOSSDAV’07*, Urbana, Illinois USA, ACM 978-1-59593-746-9/06/2007. 2007.

[33] Svoboda, W. Karner, M. Rupp. ‘Traffic Analysis and Modeling for World of Warcraft.’ *ICC 2007 proceedings of the IEEE Communications Society*, 1-4244-0353-7/07. 2007.

[34] M. Suznjevic, M. Matijasevic, and O. Dobrijevic. ‘Action specific Massive Multiplayer Online Role Playing Games traffic analysis: Case study of World of Warcraft.’ *NetGames ‘08*, 2008 ACM 978-160558-132-3-10/21/2008. 2008.

[35] W.-chang Feng, F. Chang, W.u-chi Feng, and J. Walpole. ‘Provisioning On-line Games: A Traffic Analysis of a Busy Counter-Strike Server.’ *OGI CSE Technical Report*, CSE-02-005, Portland State University. May 15, 2002.

[36] K.-Ta Chen, Huang, and C.-Laung Lei. ‘Game traffic analysis: An MMORPG perspective.’ 1389-1286, *Computer Networks* 50 3002-3023 Elsevier. 2006.

[37] K.-Ta Chen, J.-Wei Jiang, Huang, H.-Hua Chu, C.-Laung Lei, and W.-Chin Chen. ‘Identifying MMORPG Bots: A Traffic Analysis Approach.’ *EURASIP Journal on Advances in Signal Processing*. 2009.

[38] A. Back, U. Moller, and A. Stiglic. ‘Traffic analysis attacks and trade-offs in anonymity providing systems.’ In I.S. Moskowitz, editor, *Information Hiding*, pp. 245-257. Springer-Verlag, LNCS 2137. 2001.

[39] S. J. Murdoch and G. Danezis. ‘Low-Cost Traffic Analysis of Tor.’ 2005 *IEEE Symposium on Security and Privacy*, Oakland, California, USA. May8-11, 2005.

[40] F. Fusco and D. Luca. ‘High Speed Network Traffic Analysis with Commodity Multi-core Systems.’ IMC’10, Melbourne, Australia, ACM 978-1-4503-0057-5/10/11. November 1-3, 2010.

[41] Y. Lee, W. Kang, and H. Son. ‘An Internet Traffic Analysis Method with MapReduce.’ *2010 IEEE/IFIP Network Operations and Management Symposium Workshops* 978-1-4244-6039-7/10. 2010.

[42] L. Xiao, J. Gerth, and Hanrahan. ‘Enhancing visual analysis of network traffic using a knowledge representation.’ *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pp. 107-114. 2006.

[43] J. R. Goodall, W. G. Lutters, Rheingans, and A. Komlodi. ‘Preserving the Big Picture: Visual Network Traffic Analysis with TNV.’ *Workshop on Visualization for Computer Security*, 0-7803-9477-1/05 IEEE. 2005.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Trading Privacy for Security in Cyberspace: A Study Across the Dynamics of US Federal Laws and Regulations Between 1967 and 2016

Ido Sivan-Sevilla
The Federmann School of
Public Policy and Government
The Hebrew University of Jerusalem
Jerusalem, Israel
sivan018@umn.edu

Abstract: How does the legislative and regulatory agenda in the US trade between security and privacy in cyberspace? How can we explain the shift in the agenda towards more security and less privacy in the past 20 years? In order to answer these questions, I use an original dataset (N=85) of US federal laws and regulations on security and privacy between 1967 and 2016. Within the database, each policy event is classified according to the extent that security and privacy compete or complement each other. The findings indicate: (1) a shift in US federal policies towards greater security at the expense of privacy since the mid-1990s; and (2) a consistent lack of mandatory cyber security and privacy protections in the private sector. I explain this shift through emphasising: (1) the broken interest alliance over privacy between businesses and civil society organisations; (2) the increased power of the executive vis-à-vis Congress; and (3) the institutional position of security agencies and internet monopolies in the online environment. The contribution of this study stems from its empirical focus on the ambivalent role of the state in cyberspace over time. I trace the way the state promotes cyber security and privacy while increasingly collecting information or allowing others to do so at the expense of privacy and cyber security. Understanding how the state chooses between security and privacy increases our understanding of how governments manage cyberspace risks in the digital age.

Keywords: *cyber security, privacy, surveillance, resilience, regulation*

1. INTRODUCTION

This study asks how and why the US federal government has traded security and privacy over the past fifty years. The promotion of these goals complements and contradicts at the same time, and reveals the dual role of the state. Indeed, the growing dominance of cyberspace in modern life[1] requires the state to protect digitally stored personal information against new security and privacy threats. At the same time, emerging technologies and the reliance of modern societies on functioning digital systems allow the state and private actors to use cyberspace to collect personal information for security and economic purposes. This undermines privacy and threatens data security. Despite the puzzling nature of these two conflicting policy goals, little attention has been paid to the policy processes that decide between security and privacy in cyberspace.

I explore federal legislation, Executive Orders, Presidential Directives, federal register rules, federal policy guidelines, strategy documents, and novel court rulings (N=85) in the US federal arena between 1967 and 2016 and classify them to three distinct categories in which security and privacy compete with or complement each other. The findings indicate a policy shift towards greater government and privacy sector surveillance since the mid-1990s. This trend is evident throughout all branches of the federal government: (1) *the executive*, which was under close scrutiny by Congress over domestic surveillance issues in the 1970s, has been authorising surveillance without proper checks and balances and through ‘temporary’ tools that quickly became permanent. In the private sector, the executive led the way for unregulated market actors that rely on commodification of personal information for revenue; (2) *the legislature* has shifted from promoting policies that compromised between security and privacy during the 1970s and 1980s to passing legislation that encourages government information collection and letting private actors to collect personal information as they please; (3) and *the judiciary*, which was able to set the policy tone of restricting government surveillance in the 1970s, has been mostly unsuccessful in influencing the agenda in the same direction since.

The literature on the policy processes behind the conflicting roles of the state in cyberspace is surprisingly narrow. While only a few scholars address this discrepancy, they provide very limited explanations for the causes of each policy trend (Mendez and Mendez 2009; Deibert and Rohozinski 2010). Other policy scholars address either privacy (Flaherty 1989; Regan 1995) or cyber security (Etzioni 2011; Hiller and Russel 2013; Harknett and Stever 2011), but provide a limited and rather outdated empirical analysis. This study, however, considers both security and privacy as important elements of the whole, and traces them over the course of fifty years to reveal policy trends.

This paper challenges the common wisdom that emphasises the 9/11 terrorist attacks as the main significant cause for policy-makers to choose security over privacy in cyberspace. Indeed, the US Department of Justice (DOJ) issued guidelines in 2002 that indicate a strategy shift from mitigation to prevention of security threats. In addition to minimising damages, the department has devoted much of its focus to the prevention of threats altogether through massive information collection practices. But the 9/11 explanation does not completely fit with our findings, which

[1] Any attempt to function in a modern society without employing digital practices is considered to be eccentric. Zuboff (2015) argues that ‘It is impossible to imagine effective social participation - from employment, to education, to healthcare - without Internet access.’ She asserts that this phenomenon has ‘happened quickly and without our understanding or agreement.’

recognise a policy trend of increased government and private sector surveillance already in the mid-1990s. The 9/11 explanation also assumes that all federal authorities and market players act as a single actor with a unified post-9/11 strategy of increased surveillance. By separately analysing the actions of each federal authority and assessing the policy efforts of market players over time, this study enriches our understanding of surveillance drivers. Regarding private sector surveillance, the 9/11 explanation does not address the continuous deterioration of privacy by market actors. Thus, while the war on terrorism had catalysed surveillance, it was business interests, the increased power of the executive vis-à-vis Congress, and the institutional position of security agencies and information monopolies that contributed to the distinct policy shift from a compromise between security and privacy in the 1970s and 1980s to increased surveillance since the mid-1990s.

The article is organised in four sections. The next section clarifies the interplay between the concepts of privacy, security, surveillance, and resilience for the argumentation of the paper. Next, I present the analytical framework and methodology through which I test the promotion of security and privacy via cyberspace in the US federal arena. The third section discusses the findings through two sub-sections on each role of the state: (1) using cyberspace to ensure security through undermining privacy and cyber security; and (2) protecting cyberspace through cyber security and privacy measures. The shift towards greater security and less privacy is discussed through the role of each federal authority and business group in crafting this balance. The last section concludes by assessing the implications of understanding the dual role of the state in cyberspace and discussing the limitations of this research.

2. CONCEPTUAL CLARIFICATIONS

Privacy, security, surveillance, and resilience are four fundamental concepts for the analysis and arguments in this paper. These concepts can be used and understood in many ways, and the purpose of this section is to clarify their meanings in the paper.

First, I will discuss the concept of privacy. In contrast to its insufficient promotion in the public policy arena (Regan 1995), privacy as a concept has a rich history of definitions and understandings. Some stress the importance of an isolated location and space in order to enjoy the right to privacy. Other definitions tie privacy with control over personal information. A few scholars further argue that privacy is about the body and mind of the individual, rather than its location or personal information. Bygrave (2002) provides a promising framework to understand this blend of definitions. He divides the debate over the definition of privacy into four distinct groups. The first is scholars who take non-interference as their starting point and argue that individuals cannot be exposed to the public unless they choose (Warren and Brandeis 1890). The second group includes scholars who attach privacy to the levels of control over personal information (Westin 1967; Fried 1968; Rachels 1975; Laudon 1996; Lessig 1999). Scholars who focus on the degree of access to a person argue that privacy is about the body and mind and make up the third group. Gavison (1980) defines this amount of access across three dimensions – secrecy (personal information), solitude (physical access to a person), and anonymity (attention to a person). This broader notion of privacy considers mental health, autonomy, growth, creativity, and the capacity to create meaningful relations as fundamental to the definition of privacy. With no privacy, this approach would argue, we are no longer the primary controllers of our self-presentation and do not decide on the term of our social interactions. Finally, the fourth group is made up of scholars who attach privacy to intimate or sensitive information. Julie Innes promotes this privacy approach by claiming that privacy 'is the state of possessing control over a realm of intimate decisions which include decisions about intimate access, information, and actions' (Innes 1992, p. 140). For the purpose of this paper, I choose to follow the second group of definitions and argue that privacy is about individuals' ability to control their personal information. At the same time, I acknowledge that privacy holds broader implications and I do not assert that the four groups of definitions are independent. Some may lead to others, as privacy is a dynamic concept that is determined by social relations over time. Nevertheless, for the simplicity of the argument, I would assert that violation of privacy is practically any illegal and non-transparent collection of personal information, even without a proof of harm to the individual.[2]

Second, I will clarify my understanding of the broad concept of security. In contrast to the conception of privacy, security has long been viewed as a dominant policy concept that guides public policies, public opinion, and the distribution of money and power (Rothschild 1995). Hobbes (1642) views security as one of the traditional roles of the sovereign. Waldron (2006) broadens the definition of security beyond physical safety. According to him, security provides certainty, freedom from fear, and the assurance for individuals that they will not be harmed. It creates an essential platform through which individuals can enjoy other values. Locke (1689) was maybe the first to notice and define the tension between security and liberty. The possession of basic liberty rights is insufficient without the security to exercise them, but if security seriously compromises liberty, one might wonder whether security retains its fundamental value.

Waldron (2006) takes this one step further and distinguishes between two types of security. *Individual security* is defined as the security of fundamental human rights (such as privacy) and is exercised by state institutions. Individuals acknowledge that in order to sustain social and state structures that keep them protected, they have to pay a tax. Such individual security is not only about physical security, but also addresses the security of cultural, social, and institutional attributes that allow individuals to live the way they choose. *Collective security* addresses the security of the nation, its institutions, and the distribution of security across populations. It introduces questions to individuals regarding constrains they are willing to carry for the sake of the 'security of everyone'. To ensure collective security, individuals might have to carry burdens that would not necessarily improve their own security status, but rather contribute to the individual security of others. Waldron's distinction between individual and collective security is useful, and will be used in this paper to assess the tension between security and privacy in cyberspace.

Third, I will address the concept of surveillance. Common perception ties surveillance to modernity and uses the concept to better capture the contemporary privacy problems in society (Lyon 2001; Regan 2011; Bennett 2011). Surveillance is not attached to data capturing in a private space, but rather refers to the systematic monitoring and analysis of individuals that exists on all level across institutions, social practices, and modern life. It became a useful

[2] As opposed to Hughes (2015), who argues that a privacy claim turns into privacy right only when harm is unjust.

tool for government and the private sector to develop disciplinary power and new forms of governance. It is used instrumentally by the state and justified to increase collective, individual, or infrastructural security against terrorism or other break downs of public order (Regan 2011). The effects of surveillance on individuals do not just reduce privacy, they also alter opportunities and life style. The intensity of surveillance hampers freedoms, ethical principles, and even democracy itself (Raab, Jones, and Szekely 2015). The phrase 'privacy invasion' is too limited to encompass what has become a distinguishing and disquieting feature of modern life. Privacy was suited to a time when society was moving from paper records to large computerised databases, not a time of decentralised data capturing every movement on wireless devices (Regan 2011). For the purpose of this paper, I will use the term surveillance to describe a systematic violation of privacy by state institutions and private corporations.

Finally, I will discuss the term resilience. Raab, Jones, and Szekely (2015) tried to capture this concept through an extensive overview of policy documents that use 'resilience' as their end goal. They realise that this concept is identified with 'all kinds of natural or social phenomena where threats to the integrity and identity of physical objects, social goods, ethical values, or social relationships are introduced.' (Raab, Jones, and Szekely 2015, p. 23) They further find practical implications for resilience. They describe it as a coherent set of measures that include 'protecting, detecting, and responding to the consequences of threats, attacks, disasters, and other adverse events [...] that put vital interests such as national security, food supply, and community functioning at risk.' (p. 23) Strategy to achieve resilience usually relies on planned and coordinated efforts across organisations at various levels and among participants with separate roles and responsibilities. Politically, the term enjoys a certain political appeal, possibly because it suggests strength and robustness. Following these distinctions, Raab, Jones, and Szekely (2015) suggest an important conceptual boundary between the uses of the term 'resilience.' The term can be used as a property of community, individuals, or sphere, or as a set of activities undertaken to bounce back a threat. For the latter, resilience and surveillance complement each other. Governments conduct surveillance as part of a resilience strategy. For the purposes of this paper, however, I embrace the definition of resilience as a property that reflects the integrity and robustness of the digital sphere. I view resilience as a sustained and systematic process that includes capacity-building and institutional development that increases the stability and integrity of a sphere.

3. ANALYTICAL FRAMEWORK AND METHODOLOGY

The complex relations[3] between security and privacy are a subset of a broader theoretical scholarship over security and liberty in modern Western societies (Dworkin 1977; Waldron 2003; 2006; Zender 2003). The distinction between collective and individual security reveals some of the puzzle. While collective security is perceived as the 'platform' through which individuals can enjoy their liberties (Waldron, 2006), the intrusive means that political systems tend to adopt against collective security threats undermine liberty and paradoxically, some argue, lead to individual insecurity (Zender 2003; Waldron 2003; 2006). Thus, security and privacy are not logically independent and hold a social and collective importance for societies (Waldron 2003; 2006; Regan 1995, Hallsworth and Lea 2011).

[3] Waldron (2003), Zender (2003) and others argue that security and privacy are much more parallel than we tend to think. Others, like Etzioni (2014), suggest a more utilitarian approach and assert that societies should consider scenarios in which security overrides the privacy of some for the security of others.

With the expansion of cyberspace and the increasing reliance of modern societies on robust and secure digital infrastructures, it became challenging to preserve the policy goals of security and privacy. Traditional threats have evolved and adapted themselves to the characteristics of the digital sphere. Cyber criminals, commercial hacking firms, and states' cyber-espionage arms have increased cyber insecurity and required response from the state. At the same time, governments and commercial organisations have been taking advantage of new technological capacities to monitor individuals and promote security, efficiency, and economic revenue at the expense of privacy. This study focuses on these often-conflicting goals. It traces the way the state promotes cyber security and privacy, but also increasingly collects or allows the collection of information at the expense of privacy and cyber security for greater national security, law enforcement, and market goals.

Surprisingly, the dual role of the state in promoting or impeding security and privacy in the digital age has not been fully explored in the literature. We are still puzzled by how security and privacy relationships are constructed by policy-makers. Deibert and Rohozinski (2010) highlight this discrepancy by differentiating between risks 'to the security cyberspace' (hacking, cyber crime etc.) and risks 'through cyberspace' that are generated by states through cyber technologies in order to promote other policy goals. This can be achieved through political depression and the violation of privacy to ensure the stability of regimes or address security threats on the state. They recognise the contradiction between increasing cyber security and using cyberspace for surveillance, but do not take us further to understand how this discrepancy is constructed and where it comes from in the policy-making process.

Mendez and Mendez (2009) shed more light on the policy process behind these conflicting goals. They consider government laws and regulations that either protect or threaten privacy, and argue that both policy fields have experienced increased federal concentration of power. Their explanation has two limbs. They emphasise the increasing threat to US commerce posed by strict EU privacy directives in the 1990s and view it as a federal incentive for changing the sectorial 'hands-free' privacy approach of the US government towards a more centralised federal approach in the form of a privacy monitoring agency (the Federal Trade Commission). They also argue that salient policy issues with 'a dangerous external threat', like the 'war on terror' post 9/11, led to even more centralised solutions by federal actors and paved the way for federal acts that violate privacy with very little scrutiny by Congress. Their findings raise an immediate puzzle; are these contradictory roles of the state advanced equally across federal powers? Since Mendez and Mendez (2009) base their conclusions on rather narrow empirical foundations,[4] we are still puzzled regarding the paradoxical role of the state in cyberspace. The authors' empirical analysis does not address the federal arena over time, and fails to consider cyber security policies as a tool for promoting privacy as well. While the federal arena lacked any privacy promotions during the 2000s, Congress and the Executive did pass significant privacy protections before that. These empirical shortcomings do not allow explanations other than 9/11 terrorist attacks for violations of privacy, and do not fully explore the role of business interests in this contradictory policy process. If the 9/11 terrorist attacks explain the expansion of US surveillance policies, why did we witness this expansion in the 1990s? If, according to Mendez and Mendez (2009), the likelihood of passing acts for federal privacy protections is high, why do we constantly see failed attempts to pass federal privacy legislation? The role of

[4] They only focus on the rise of the Federal Trade Commission (FTC) as the U.S. privacy regulator in the 2000s in light of two significant laws that violate privacy in the same post 9/11 period.

businesses in government surveillance policies, and the failed attempts to pass federal privacy protections for the private sector, are not explored despite their significance in these policy processes.

Other scholars have addressed either privacy or security to study one aspect of the state's role, but did not analyse these attempts as part of the whole. Privacy policy scholars explain lack of privacy protections by either policy-makers' perceptions of privacy as an individual value that is subordinate to other collective values[5] (Regan 1995) or the lack of institutional capacities in the US to adequately promote privacy (Flaherty 1989). While these studies enrich our understanding on the policy processes that lead to insufficient privacy protections, they are rather outdated and focus on the 1970s and 1980s in the US federal arena. What these scholars viewed as insufficient privacy protections is nowadays viewed as the 'golden age' of privacy that was followed by significant privacy erosions by the US government and the private sector. A more recent study by Newman and Bach (2004) analyses the incentives behind the self-regulation model of privacy protections in the United States. Newman and Bach (2004) argue that latent threats and the potentially costly federal regulation dictate close collaborations within industries to avoid government regulation. While Newman and Bach (2004) shed light on why this 'hands-free' federal approach over privacy persists, we still lack an understanding of how and why this approach was decided on in the first place. This self-regulatory model is not contrasted with the emerging private sector surveillance that was created from this lack of privacy scrutiny by the government.

Finally, scholars of security in cyberspace shed even less light on the policy process and the contradictory role of the state. Etzioni (2011) explains the implications behind the reluctance of private actors to accept mandatory cyber security regulations, while Hiller and Russel (2013) vaguely explain the self-regulatory model of private sector cyber security through the 'regulatory culture' in the US that is traditionally skewed in favour of businesses. None of these scholars, however, address the dual role of the state empirically over time, or link security and privacy to provide explanations on the policy processes.

To explore the relations between security and privacy in cyberspace through the US federal arena, I have created an original data set with policy events[6] (N=85) from the years 1967 – 2016 that address:

(1) Government information collection for security purposes at the expense of privacy;
(2) Limitations to government information collection that increase privacy at the expense of security; and
(3) Cyber security and data protection measures that advance security and privacy at the same time.

The methodological approach of this study does not only include components of the traditional cyber security and data protection regulation, but also covers the promotion of national

[5] Such as national security, law enforcement, or business efficiency.

[6] Federal Legislation, Executive Orders, Presidential Orders and Directives, National Security Directives, Federal Register Rules from federal agencies, Policy Guidelines from federal agencies that provide additional interpretation to federal statues, Strategy documents from the White House and federal agencies that provide voluntary recommendations and best practices on the related issues, novel FISA Court rulings that further advance the understanding and practices over government surveillance, and novel Supreme and District Court rulings that provide new interpretation to the regulatory regimes of security and privacy.

security and law enforcement goals that dictate the extent to which the government can collect information in cyberspace. This allows for a broad understanding on the dynamics between security and privacy and the contradictory role of the state. The starting point of the policy events was chosen to be *Katz v the United States (389 U.S. 347)*, a Supreme Court landmark ruling from 1967 which overturned a decision from 1928 and provided constitutional privacy protections from government information collection.[7] This Supreme Court decision had triggered policy-making processes over security and privacy issues that shape the regulatory arena as we know it today.

Each gathered policy event was then classified to one of three policy categories according to the relationships it dictates between security and privacy. The events were conceptually mapped according to the following table:

TABLE I. THE CONCEPTUAL MAPPING OF POLICY EVENTS ACCORDING TO THE RELATIONSHIPS BETWEEN SECURITY AND PRIVACY IN EVERY EVENT

		PRIVACY	
		+	–
SECURITY	+	Security & Privacy complements (N=33): Cyber security and data protection practices that strengthen the security of personal information systems and advance the right to privacy of the associated data subjects	Security > Privacy (N=31): National Security or Law Enforcement policies that increase the collection of personal information and weaken digital infrastructures for security purposes
	–	Privacy > Security (N=21): Privacy practices that limit government information collection for security purposes and promote privacy at the expense of potential security risks that may arise from lack of collected information	

The *first* category includes policy-making events from 1984–2016 that strengthen security and privacy in cyberspace at the same time (N=33). These are mainly cyber security and data protection measures that strengthen the security of information systems as well as the privacy of individuals whose personal information is processed by those systems. These policies reflect one role of the state as strengthening the resilience of cyberspace. The *second* category includes policy-making events from 1976–2015 that open avenues for information collection by the government to advance national security and law enforcement at the expense of privacy (N=31). The *third* category includes policy-making events from 1967–2016 that deal with limitations to government surveillance, and thus strengthen privacy at the expense of security (N=21). These latter two categories reflect the extent to which the state uses cyberspace to increase security over privacy.

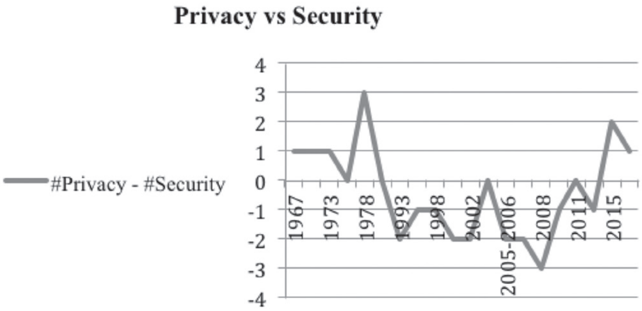
[7] The court decided that using a telephone bug without a court order violates privacy according to the Fourth Amendment and determined that the right to privacy is entitled to people rather than places.

4. FINDINGS

A. Security ≠ Privacy: The State Uses Cyberspace to Advance Security Over Privacy

Figure 1 reflects the shift in the role of the state that uses cyberspace. Since the mid-1990s, the US federal arena has been crafting more laws and regulations to collect information from cyberspace, and thus, prioritise security over privacy. For each year, the figure reflects the yearly number of federal measures that *limit* government information collection minus the yearly amount of federal measures that *encourage* government information collection. While in the 1970s and 1980s the federal arena had more privacy than security measures (the blue line is above the horizontal axis), since the mid-1990s, the line is mostly under the horizontal axis, which quantitatively reflects a clear priority for national security and law enforcement measures over privacy measures in cyberspace.

FIGURE 1. ([PRIVACY MEASURES] – [SECURITY MEASURES]) PER YEAR OVER TIME [1967-2016]



In the last four decades, the government's information collection desires have been regulated through a variety of policy instruments and significantly affected the way security and privacy risks are managed in cyberspace. In the 1970s and 1980s, Congress was able to limit governmental desires for information collection. Up to the mid-1990s, Congress constructed a compromise between security and privacy, backed by public legitimacy, following political surveillance disclosures of anti-Vietnam war activists and the Watergate scandal (1974). The 1990s, however, brought the technological breakthrough of mass encryption and communication capabilities. Consequently, law enforcement agencies were worried that their surveillance capabilities would be undermined. This also changed the alliance of interests between businesses and civil society. In contrast to pro-privacy policies that were advanced by private businesses for the sake of their products, the 1990s brought close ties between government and telecoms businesses to increase surveillance. The 9/11 terrorist attacks and the official strategy of the administration, reflected in the 2001 Patriot Act, to collect anything 'tangible', was accompanied by the lack of significant pushback by Congress.

Figure 1 demonstrates a shift towards security at the expense of privacy in the use of cyberspace over time. This shift is evident throughout US federal authorities and relevant business groups. First, I will briefly summarise the actions of the executive. Since the 1990s, the executive branch had gradually eroded checks and balances over the use of government power to collect information from cyberspace. During the 1960s and 1970s, however, significant public outcries led to a rather weak executive. Congress was able to form investigative committees (the Church and Pike Committees of 1976) that eventually limited government surveillance and posed serious constraints on the ability of the executive in this area. Following demands from the American public, the executive itself initiated steps to limit surveillance through Executive Orders and Attorney General Guidelines (1976). However, since the mid-1990s, the legitimacy to exercise executive power has changed. With no major public scandals over surveillance and with the mind-set of the 'war on terror', there was almost complete federal silence over pro-privacy issues. This led to weak oversight mechanisms and lowered the standards and judicial safeguards for government information collection. This trend in the executive started in the 1981 Executive Order #12333, in which President Regan ordered the protection of privacy when collecting information, but only through self-regulation mechanisms rather than an external oversight. This trend significantly increased through several Attorney General Guidelines over the years (1983, 1989, and 2002) which broadened surveillance authority within the US with no negotiations or a complete policy process through Congress. In addition, from 2001–2007, President Bush solely relied on his own judgement to launch and then secretly continue to operate unlawful surveillance programmes, without notifying Congress. In 2001, President Bush launched these programmes on a 'temporary' basis to gain legitimacy for their use, and then extended them for seven years. Finally, the Administration had gradually expanded the authority of National Security Letters (NSLs) from the Department of Justice (DOJ). This policy instrument was originally launched as a national security exception to required data protection practices, but soon became one of the main avenues for information collection, taking the tool completely out of its original purpose.

Second, the role of Congress in the shift for greater security over privacy is quite striking. During the 1970s and 1980s, the legislature was highly active and pressured the administration to limit surveillance through its investigative committees and unprecedented legislation (Foreign Intelligence Surveillance Act (FISA) 1978 and Electronic Communications Privacy Act (ECPA) 1986[8]). However, since the mid-1990s, Congress has been unable to pass any significant legislation to limit information collection.[9] Privacy stakeholders in Congress became weak and were brought down by competing national security and law enforcement forces. Congress had shifted its role from pushing for a compromise between security and privacy, to backing the administration and supporting the collection of data at the expense of privacy. Since the mid-1990s, Congress allowed the passing of surveillance measures that were marketed as 'temporary' but soon became permanent. These include the Patriot Act 2001 provisions and the 2007 and 2008 amendments to FISA that significantly weaken privacy in favour of greater surveillance. Congress had also allowed the erosion of the legal barrier between surveillance for national security purposes and surveillance for law enforcement purposes. While national security surveillance was limited by previous laws, US intelligence agencies circumvented these limitations through collecting national security information in the name

[8] Both these acts significantly limit the collection of personal information for national security (FISA) and law enforcement (ECPA) purposes.

[9] The US Freedom Act that was passed in 2015 was the first legislation after 30 years that limits government bulk information collection.

of ‘law enforcement purposes.’ Congress approved the erosion of this ‘wall’ policy through the Patriot Act 2001, since it was related to one of the September 11 intelligence failures. Recently, however, following disclosures on government surveillance by Edward Snowden, Congress passed the US Freedom Act 2015 that has imposed some limitations on government surveillance after several decades.

The judiciary also had a role in the shift towards greater surveillance. Throughout the 1970s and 1980s, the courts had a significant role in promoting pro-privacy legislation to the federal policy-making agenda. This is evident through two examples: (1) the 1967 *Katz vs. United States (389 U.S. 347)* court ruling that was the basis of the Omnibus Crime Control Act 1968, which limited government information collection for the first time; and (2) the 1976 *United States vs. Miller (425 U.S. 435 1976)* case, which limited privacy rights when information was shared with third parties by individuals, was the basis for a counter-response by policy-makers in the form of the Electronic Communications Privacy Act (ECPA) 1986. At the same time, the influence of the courts on limiting government information collection since the 1990s was only expressed through amendments to restrict the use of NSLs by the FBI in the 2000s. Despite additional court rulings on the illegality of surveillance, the court was unable to influence the agenda and advance legislative measures to promote privacy over security. In fact, since the 1990s, decisions from the special Foreign Intelligence Surveillance Courts (FISC) for surveillance authorisation have contributed to the surveillance agenda of the government. These courts have secretly ruled on controversial issues without oversight from external judges. Moreover, FISC judges have also allowed controversial temporary surveillance authorisations to become permanent. Overall, the judiciary has been unable to limit surveillance in the past two decades, and in fact had a role in opening more avenues for government surveillance at the expense of privacy through FISA rulings.

Finally, the role of business groups from the telecom industry in fuelling the policy shift of security over privacy is also significant. Since privacy was viewed as an economic advantage during the 1970s and 1980s, business groups had strongly supported limiting government surveillance. An alliance over privacy between businesses and civil society was formed and successfully advanced pro-privacy legislation such as the Right to Financial Privacy Act 1978 and the Electronic Communication Privacy Act (ECPA) 1986. But in the mid-1990s, this alliance was broken after industry leaders paved the way for the Communication Assistance for Law Enforcement Act (CALEA) 1994 to pass. The government had provided a significant compensation for telecoms businesses to make their communications infrastructure ‘surveillance friendly’ for the government, and had weakened privacy at the expense of security. This trend of close ties between businesses and government continued throughout the 1990s and 2000s.

Most of this cooperation is hidden, but what we do know is that businesses cooperated with the government over controversial uses of National Security Letters (NSLs) to collect information. In addition, Internet Service Providers (ISPs) were allowed by law (Patriot Act 2001) to conduct surveillance based on their own judgement, while probable surveillance causes were lowered to suspicion only. Recently, however, since the Snowden disclosures, civil society and business interests have converged again. Examples include the refusal by Apple to break iPhone encryption for national security purposes; the pushback of the industry that was able to postpone CALEA II legislative proposals by the FBI; and the opposition from Microsoft to turn in personal records of its clients from servers outside US jurisdiction. Privacy is gradually re-becoming a competitive economic advantage and a way to satisfy consumer demands.

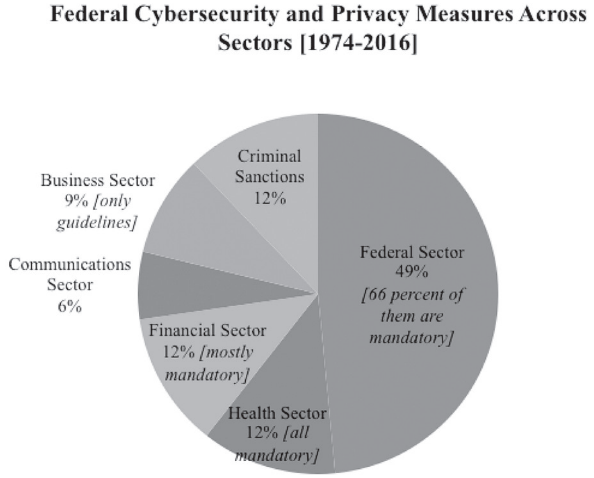
B. Security = Privacy: The State Increases the Security and Privacy of Cyberspace

At the same time, and in contrast to its efforts to undermine privacy and cyber security for national security and law enforcement purposes, the state has passed laws and regulations to increase the resilience of cyberspace. Three elements stand out from these policy efforts. First, these protection efforts are narrow in scope, mostly voluntary for the business and communications sectors, and thus do not seriously limit surveillance practices by the government or the private sector. Second, the government's agenda from the very early days of the Internet was to leave it unregulated. This had created the institutional conditions for the massive private sector surveillance for revenue we see today. Third, even within this policy category that mostly promotes security and privacy while increasing the resilience of cyberspace, we can see instances of tension over promoting cyber security through the collection of personal information and the violation of privacy.

1) Insufficient Privacy and Cyber Security Protections in the Private Sector

The advancement of security and privacy in cyberspace became a salient issue for federal policy-makers in the mid-1980s (Warner 2012). Nonetheless, despite the growth of the Internet's economy and the reliance on private networks, the promotion of security and privacy in the private sector is inefficient, sectorial, and only partially advances cyber security and privacy in cyberspace. The federal government is unable to extend its reach and regulate segments other than the federal, health and financial sectors. Most mandatory regulations pose strict requirements on critical sectors and the federal government, while non-critical private sectors are only addressed through declarative policy-making as their protection is mainly based on self-regulation models. Figure 2 demonstrates the lack of federal government efforts to increase the resilience of cyberspace for the business and communications sectors. Almost half of the regulatory effort is channelled to federal government networks, but while the federal government has to protect itself, it leaves other significant sectors to rely on voluntary practices (Newman and Bach 2004).

FIGURE 2. MANDATORY AND VOLUNTARY CYBER-SECURITY AND PRIVACY MEASURES ACROSS SECTORS [1974-2016]



Businesses were able to skip costly and mandatory regulations from the very beginning. The Privacy Act 1974 was enacted by Congress after long debates over the importance of protecting private information from federal authorities. Within these debates, private industry argued that there was little concrete evidence of abuses in the information practices of private businesses. They claimed that they were already overburdened by government regulation and the proposed regulation by the federal government was unnecessary and costly (Regan 1995, p. 78). Their strategy was to urge companies to enact voluntary protections for personal information in order to lessen the pressure for government regulation. The private sector also opposed establishing a federal agency to oversee information collection and use. A more rigorous Senate bill that included the creation of a privacy protection commission and suggested tighter restrictions on personal information was rejected in favour of a weaker House bill. In fact, the Privacy Act 1974 encompassed the minimum protection that was advocated at that time (Regan 1995).

In the mid-1990s, the federal government responded to the growth of the Internet by regulating the information security and privacy of specific private networks. Yet again, businesses were able to skip strict federal privacy regulations. In 1997, The Clinton Administration published *The Framework for Global Electronic Commerce (Clinton and Gore 1997)* which described businesses as essential to the new economy. The Administration did not want to limit businesses' expansion by posting costly and mandatory regulations over their operations. The framework also called for applying self-regulation models over privacy protections, leaving privacy decisions to the private sector. This practically mandated the private sector to set the standards of privacy protections by itself. This trend emerged in the beginning of the Internet age and has not changed since. It allowed the business practice of the commodification of personal information to evolve, and led to extensive private sector surveillance for economic purposes by information monopolies like Google and Facebook.

Despite dozens of bills to regulate the private sector, Congress ended up passing legislation solely on the health and financial sectors. Health records were recognised as sensitive and critical in the digital age and the Health Insurance Portability and Accountability Act (HIPAA) 1996 was the first time that information security standards were codified. After long debates and the expressed concerns of private players over the cost and complexity of the regulation, the act became a binding federal register rule in 2003. The policy-makers then turned to the financial sector. Through the Gramm-Leach-Bliley Act (GLBA) 1999, federal policy-makers sought to ensure the security and privacy of information in the financial sector. The financial sector enjoyed an additional indirect boost through the Sarbanes-Oxley Act (SOX) 2002 that was introduced following the collapse of Enron and WorldCom in order to restore public trust in US corporations. The legislation changed the way publicly traded companies manage their audit, financial reporting, and internal controls. While information security is not specifically discussed in the Act, reviews of companies' controls include information security controls that have to follow the same strict restrictions.

In 2010, after a decade of federal laws and regulations that mainly dealt with the protection of federal networks and critical infrastructures, the regulatory agenda was shifted toward regulating the private sector. But instead of changing the 40-year trend towards mandatory and strict requirements, the Department of Commerce continued the trend of posing no mandatory security and privacy requirements on the private sector. The Department published two strategy documents to address the privacy and security problems of 'non-critical' sectors. The first[10] goes a long way by suggesting a baseline for consumers' privacy and security protections in the business sector. Specifically, the strategy recommends businesses to adopt the Federal Information Processing Standard (FIPS) – a privacy standard that was enacted in 1974 by the Privacy Act – and suggests that the federal government pass federal breach notification rules.[11] The strategy also calls for the establishment of a privacy oversight office under the Department of Commerce. A similar policy suggestion on the oversight of the privacy of federal networks through a dedicated office was made in the 1970s, but had been unable to get on the agenda since. The second strategy document[12] calls for collaboration between the public and private sector and the promotion of trust and multi-stakeholder processes in order to develop security best practices and make them industry standards. The suggested framework is voluntary, and the aim of the strategy was to find the proper security protection for each 'sector and sub-sector in the economy.' (Cybersecurity, Innovation, and the Internet Economy p. 2) It defines a new sector – the Internet and Information Innovation Sector (I3S) – and provides recommendations

[10] The referred policy document is the *Commercial Data Privacy and Innovation in the Internet Economy: A dynamic policy framework*. The document argues that 'many key actors, due to the sectorial privacy and cyber-security approach of the U.S., operate without specific statutory obligations to protect personal data.' (p. 12).

[11] These are rules that require companies to report and face financial consequences in case of a data breach. Currently, the U.S. has 47 versions of breach notification laws across its states and was unable to pass a unified federal legislation despite many attempts in the last 15 years. There is controversy over issues like – federal preemption, desired policy goals, scope of notification, and effectiveness of policy.

[12] The referred policy document is the *Cybersecurity, Innovation, and the Internet Economy* by the Department of Commerce Internet Policy Task Force.

on technical security standards and incentives[13] to deal with cyber security threats that are integrated in the culture of each firm. Despite these important efforts, the cyber security and privacy of businesses remained almost completely a product of self-interest and judgement, bound only to what is considered 'fair trade practices' that could be enforced by the FTC.

Since 2013, however, we have seen a few significant steps at the federal level to ensure private sector security and privacy. The health and financial sectors are now required to adopt further protections through guidelines from federal agencies, while the Federal Communications Commission (FCC) is also increasing its role as a privacy and security regulator in cyberspace. The agency published a strategy document[14] with voluntary recommendations to communication providers on how to mitigate cyber security risks and comply with the National Institute for Standards and Technology (NIST) network security framework. Additionally, the 3rd U.S. Circuit Court of Appeals in Philadelphia had recently taken a stand on the authority of the Federal Trade Commission (FTC) to enforce cyber security protections in the private sector *(FTC vs. Wyndham Worldwide Corporation 2015)*. This was a significant ruling. Previously, the FTC had relied on the reasonableness of companies' security practices and enforced regulation based on unfair business practices. This enforcement power was authorised to the FTC by section 5 of the FTC Act 1914. Following this ruling, the FTC has a new mandate and institutional power to enforce cyber security and privacy protections. This trend continued in 2016, with the FCC moving from recommendations to actions. It published a new rule that requires Internet Service Providers (ISPs) to protect their consumers against information collection practices and require full transparency in personal information processing. However, with the recent change in the Administration and the appointment of a new FCC Chair by President Trump, these mandatory privacy guidelines have already been partially reversed.[15]

These inadequate privacy protections have paved the way for the lack of barriers for government surveillance, but more importantly, laid the foundations for the commodification of personal information and the massive surveillance practices we see today in the private sector. In its role to promote cyber security and privacy, the state does not only lack the ability to protect the private sector, but has also created the conditions that have allowed private sector surveillance to emerge and thrive.

2) Contradictions between Cyber Security and Privacy in the State's Efforts

Beyond the lack of sufficient mandatory protections for the private sector, the federal government also risks privacy to increase the security of cyberspace. Even though increasing the levels of cyber security would also increase the protection of privacy, many state initiatives include massive information collection practices. This tension between cyber security and privacy reflects a shift towards less privacy protections in cyber security measures from 2001 onward.

Concerns over the means and the potential violations of privacy to ensure information security were already evident in 1984. Through a National Security Directive in 1984, President Reagan authorised the National Security Agency (NSA) to protect all classified government

[13] Such incentives can be achieved through national breach notification law, information sharing and liability protections, and insurance mechanisms.

[14] The referred strategy document is the *Cybersecurity Risk Management and Best Practices Working Group 4: Final Report March 2015*, FCC.

[15] The new FCC chairman, Ajit Pai, blocked FCC requirements from ISPs to apply common sense security practices to protect personal information. See more here: https://www.eff.org/deeplinks/2017/02/new-fcc-chairman-begins-attacks-internet-privacy.

networks. This decision increased surveillance concerns of US policy-makers and in 1987, Congress responded and enacted the Computer Security Act. The purpose of the Act was to assign responsibilities for the protection of federal networks and appoint the civilian National Institute for Standards and Technology (NIST) rather than the intelligence National Security Agency (NSA) as the sole body responsible for protecting federal networks. However, NIST was never powerful enough institutionally and was bound to certain arrangements with the NSA that questioned its influence over the entire process (Flaherty 1989). In 2001, an unexpected contributor to the promotion of cyber security and privacy was the passage of the Patriot Act. While the Act significantly weakens privacy protections for information collection that serves national security interests, it also increases defensive capabilities against cybercrime. The act creates, for the first time, the definition of a ‘computer trespasser’. It also allows law enforcement officials to trace the communications of computer trespassers and improves their ability to track cybercrime activities. Section 220 allows a single court order with jurisdiction over the cybercrime offence to issue a search warrant for electronic evidence anywhere in the country. These increased cybercrime capabilities threaten privacy. The types of collected data on computer trespassers is broad; section 210 expands the information that can be obtained from communications providers to include means and sources of payments as well as session times and temporarily assigned network addresses, while section 216 applies tracking devices of meta-data (e.g. pen registers and trap-and-trace devices) to any communication facility in the country.

Tension between the right to privacy and promoting cyber security was also expressed in the 2008 *Comprehensive National Cyber Security Initiative (CNCI)*. The White House published this strategy in order to ensure that federal networks are resilient to the dynamic nature of cyber-attacks. The initiative encourages the use of intelligence and advances information collection of foreign intelligence. It also encourages the use of decryption capabilities by the NSA that risk privacy. Another instance of tension between cyber security and privacy is the recent passage of the Cyber Information Sharing Act (CISA) 2015. This allows non-transparent information collection from the private sector on cyber threats for the sake of greater cyber security. It signals a new phase in the tension between privacy and cyber security as it removes liability from businesses that choose to share information, and thus encourages information collection by the government without a court order. To conclude these trends, even in its role as protecting cyberspace, the US has undermined privacy in cyber security at an increasing rate since 2001.

5. CONCLUSION

Over the course of fifty years, the US federal government has held a contradictory role with regards to the promotion of security and privacy in cyberspace. In the past twenty years, the power of the executive branch vis-à-vis Congress and the close cooperation between businesses and security agencies have increased surveillance and weakened the privacy and cyber security of cyberspace. At the same time, the reluctance of businesses to accept mandatory regulations and the power of the executive and security agencies to advance cyber security at the expense of privacy have led to lax cyber security and privacy protections. This has encouraged the trend of deteriorating privacy and created the institutional conditions for massive private sector surveillance by powerful information monopolies that undermine privacy for economic revenue.

While the US excels in undermining privacy and cyber security to achieve 'greater' security goals, it is less successful in strengthening the resilience of cyberspace as a common good. After thirty years of policy experience with regards to cyberspace, this paper illustrates the urgency of changing the policy discourse on cyber security. In order to advance cyberspace as a common good, the current policy framing that places cyber security as a property of systems rather than people has to change.

Traditional definitions of cyber security usually involve the protection of digital systems against hackers, and the dominant actors in this policy domain are security agencies and private interests. This creates a discourse over cyber security according to which 'surveillance powers are expanded, encryption is limited, backdoors are installed, and accountability structured are weakened' (Puddephatt and Kaspar 2015 p. 3). This dramatically opposes individual security and questions the move towards greater reliance on digital practices that weakens the fabric of society.

As demonstrated empirically in this paper, the overarching policy trends over security and privacy in cyberspace are a source of concern. The policy discourse over those issues should emphasise the security rights of end users rather than just the security of systems or the promotion of national interests. This might mean giving ownership back to data subjects, guarantee end-to-end encryption and public education over privacy, and include stronger accountability and oversight mechanisms for necessary data collection by ensuring that the scope of such powers is narrowly defined. Voices other than those of the security agencies should be involved in the policy debate to ensure that individual and collective cyber security are jointly advanced.

Finally, it is worth mentioning the limitations of my research conclusions. There is a tension in this paper between the amount of the data (85 policy events) and the explanatory power of my arguments that relies on general trends in regulation over time. The next logical step would be to focus on three or four cases that provide a representative sample of the data and analyse them comparatively to reveal the mechanisms of business influence and executive power over the way the US government manages risks in cyberspace.

REFERENCES

Bennett, C.J., 'In Defense of Privacy: The Concept and the Regime', *Surveillance & Society*, 2011, Vol. 8, No. 4.

Bygrave, L.A., *Data Protection Law – Approaching its Rationale, Logic, and Limits*, 2002, Kluwer Law Intl.

Clinton, W.J., and Gore, A., *A Framework for Global Electronic Commerce*, 1997, Washington, DC.

Deibert, R.J., and Rohozinski, R., 'Risking Security: Policies and Paradoxes of Cyberspace Security', *International Political Sociology*, 2010, Vol. 4, Issue 1, 15-32.

Doyle, C., ‘The USA Patriot Act: A Legal Analysis’, *Congressional Research Services (CRS) Report for Congress*, 2002.

Dworkin, R., *Taking Rights Seriously*, 1977, Harvard University Press.

Etzioni, A., ‘Cybersecurity in the Private Sector’, *Issues in Science and Technology*, 2011, Vol 28, Issue 1.

Etzioni, A., ‘A Cyber Age Privacy Doctrine: A Liberal Communitarian Approach’, *Journal of Law and Policy For the Information Society*, 2014, Vol 10, Issue 2.

Federal Trade Commission, ‘Protecting Consumer Privacy in an Era of Rapid Change’, *Preliminary FTC Staff Report*, 2010.

Flaherty, D.A., *Protecting Privacy in Surveillance Societies: The Federal Republic of Germany, Sweden, France, Canada, and the United States*, 1989, UNC Press.

Fried, C., ‘Privacy’, *Yale Law Journal*, 1968, Vol. 77, 475-93.

Gavison, R., ‘Privacy and the Limits of Law’, *Yale Law Journal*, 1980, 89: 442.

Hallsworth, S. and Lea, J., ‘Reconstructing Leviathan: Emerging contours of the security state’, *Theoretical Criminology*, 2011, 15(2), 141-157.

Harknett, R.J., and Stever, J.A., ‘The New Policy World of Cybersecurity’, *Public Administration Review*, 2011, 455-460.

Hiller, J.S., and Russel, R.S., ‘The challenge and imperative of private sector cybersecurity: An international comparison’, *Computer Law & Security Review*, 2013, Vol. 29, 236-245.

Hobbes, T., *De Cive*, 1642.

Hughes, D.R.L., ‘Two Concepts of Privacy’, *Computer Law & Security Review*, 2015, Vol. 31, 527-537.

Innes, J., *Privacy, Intimacy, and Isolation*, 1992, New York: Oxford University Press.

Laudon, K.C., ‘Markets and Privacy’, Association for Computing Machinery, 1996, *Communications of the ACM*, Vol 39, Issue 9.

Lessig, L., *Code and Other Laws of Cyberspace*, 1999, Basic Books.

Locke, J., *Two Treatises of Government*, 1689, London.

Lyon, D., *Surveillance Society: Monitoring Everyday Life*, 2001, Open University Press.

Mendez, F., and Mendez, M., ‘Comparing Privacy Regimes: Federal Theory and the Politics of Privacy Regulation in the European Union and the United States’, *The Journal of Federalism*, 2009, Vol. 40, Issue 4, 617-645.

Newman, A.L., and Bach, D., ‘Self-Regulatory Trajectories in the Shadow of Public Power: Resolving Digital Dilemmas in Europe and the United States’, *Governance*, 2004, Vol 17, No. 3, 387-413.

Puddephatt, A., and Kaspar, L., ‘Cybersecurity is the new battleground for human rights’, *OpenDemocracy.net*, 2015, access: https://www.opendemocracy.net/wfd/andrew-puddephatt-lea-kaspar/cybersecurity-is-new-battleground-for-human-rights.

Raab, C., Jones, R., and Szekely, I., ‘Surveillance and Resilience in Theory and Practice’, *Media and Communication*, 2015, Vol. 3, Issue 2, 21-41.

Rachels, J., ‘Why Privacy is Important’, *Philosophy & Public Affairs*, 1975, Vol. 4 No. 4, 323-333.

Regan, P., Legislating Privacy: *Technology, Social Values, and Public Policy*, 1995, UNC Press.

Regan, P., 'Response to Bennett: Also in defense of privacy', *Surveillance & Society*, 2011, Vol. 8, No. 4.

Rothschild, E., 'What is Security?' *Daedalus*, 1995, Vol. 24, No. 3, 53-98, MIT Press.

Waldron, J., 'Security and Liberty: The Image of Balance', *Journal of Political Philosophy*, 2003, 11 (2): 191-210.

Waldron, J., 'Safety and Security', *Nebraska Law Review*, 2006, 85: 454-507.

Warner, M., 'Cyber Security: A Pre-history', *Intelligence and National Security*, 2012, Vol 27, Issue 5, 781-799.

Warren, S.D., and Brandeis, L.D., 'The Right to Privacy', *Harvard Law Review*, 1890, 4: 193-220.

Westin, A., *Privacy and Freedom*, 1967, New York: Atheneum.

The White House, 'National Strategy for Trusted Identities in Cyberspace', *White House Strategy*, April 2011.

Zender, L., 'Too Much Security?', *The Journal of Sociology of Law*, 2003, Vol. 31, Issue 3, 155-184.

Zuboff, S., 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilisation', *Journal for Information Technology*, 2015, Vol 30, 75-89.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

The Role of International Human Rights Law in the Protection of Online Privacy in the Age of Surveillance

Eliza Watt
Westminster Law School
University of Westminster
London, UK
elizawatt@googlemail.com

Abstract: Whilst the political dust on mass surveillance is slowly settling down, what has become apparent is the uncertainty regarding the interpretation and application of the right to privacy norms under Article 17 of the International Covenant on Civil and Political Rights 1966 in the context of cyberspace. Despite the world-wide condemnation of these practices by, *inter alia*, the United Nations and international human rights organisations, little consensus has been reached on how to bring them in line with international human rights law. This paper proposes that the most pragmatic solution is updating Article 17 by replacing General Comment No.16. There are many issues that require attention. The paper focuses on two fundamental aspects of this process, namely the development of more detailed understanding of what is meant by the right to privacy in the 21st century, and the challenge posed by foreign cyber surveillance to the principle of extraterritorial application of human rights treaties. To that end, the paper identifies that the 'effective control' test, developed by international human rights courts and bodies adopted to determine jurisdiction, is unsuitable in the context of state-sponsored cyber surveillance. The paper considers a number of suggestions made by legal scholars, which hinge on the control of communications, rather than the physical control over areas or individuals. Such a 'virtual control' approach seems in line with the jurisprudence of the European Court of Human Rights, according to which extraterritorial obligations may arise when states exercise authority and control over an individual's human rights, despite not having physical control over that individual. The paper argues that the 'virtual control' test, understood as a remote control over the individual's right to privacy of communications, may help to close the normative gap that state intelligence agencies keenly exploit at the moment.

Keywords: *cyber surveillance, privacy, extraterritorial obligations, 'effective control' test, 'virtual control' test*

1. INTRODUCTION

One of the starkest lessons to be learned from the 2013 Edward Snowden revelations is the need for a global solution regarding state sponsored communications surveillance,[1] conducted in particular by the coalition of the so-called Five Eyes states.[2] Undoubtedly, these activities breach the right to privacy of communications[3] enshrined in Article 17 of the International Covenant on Civil and Political Rights 1966 (ICCPR)[4] and Article 8 of the European Convention on Human Rights 1950 (ECHR).[5] However, despite numerous calls from international organisations and human rights courts and bodies condemning mass surveillance, to date there is no consensus on how to bring these activities in line with human rights law.

This paper will address some of these challenges, focusing on legal solutions within the existing international human rights framework, as achieving a legally binding agreement remains elusive. To that end, the first part will outline some recent developments from the United Nations (UN) organisations[6] and human rights bodies;[7] it will conclude that current state practice in the form of transboundary state-sponsored cyber espionage[8] and long-standing disagreements regarding the future of Internet governance[9] make the negotiation from scratch of a new UN privacy treaty for the digital environment unlikely. However, there are other solutions, discussed in part two, such as the long overdue modernisation of the existing privacy norms under Article 17 ICCPR. The paper will focus on two important aspects of this process, namely the updating of the notion of privacy and the extraterritorial application of human rights treaties in the context of cyber surveillance. This part will outline the approach adopted in the international human rights law jurisprudence, which in certain circumstances holds a state accountable for human rights violations conducted extraterritorially, based on the 'effective control' test. It will be shown that this model of extraterritorial jurisdiction, currently articulated as physical power or control over either an area or a person, is not well suited to the cyber environment and needs therefore to be adapted for the transboundary context of digital mass surveillance. A number of possible solutions have been proposed by legal scholars, and this paper will outline some of their rationales. It will conclude that the exercise of power or authority over an individual's

[1] For a definition of communications surveillance see UNHRC 'Report by the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue' (2013) UN Doc A/HRC/23/40, para 6.

[2] The Five Eyes comprises the US National Security Agency, the UK General Communications Headquarters, Canada's Communications Security Establishment Canada, the Australian Signals Intelligence Directorate and New Zealand's Government Communications Security Bureau.

[3] UNGA, 'Report of the Office of the United Nations High Commissioner for Human Rights the Right to Privacy in the Digital Age' (2014) UN Doc A/HRC/27/37, para 20; *Privacy International v Secretary of State for Foreign and Commonwealth Affairs* [2016] UKIPTrib 15_11-CH.

[4] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 17.

[5] Convention for the Protection of Human Rights and Fundamental Freedoms (opened for signature 4 November 1950, entered into force 3 September 1953) 213 UNTS 222 (ECHR), art 8.

[6] UNGA Res 68/167 (18 December 2013) UN Doc A/RES/68/167; UNGA Res 69/166 (18 December 2014) UN Doc A/RES/69/166.

[7] OHCHR Report, supra note 3; UNHRC, 'Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Ben Emmerson QC' (2014) UN Doc A/69/397; Report of the Special Rapporteur La Rue, supra note 1.

[8] Russell Buchan, 'The International Legal Regulation of State-Sponsored Cyber Espionage', in Anna Maria Osula and Henry Roigas (eds.), *International Cyber Norms: Legal, Policy and Industry Perspective* (NATO CCD COE Publications, Tallinn 2016) 65-86.

[9] *The Guardian*, 'ITU and Google Face-off at Dubai Conference over Future of the Internet' (3 December 2012).

right to privacy through ‘virtual control’ over their communications may constitute a workable way forward in preventing states from avoiding their human rights responsibilities ‘simply by refraining from bringing those powers within the bounds of the law’.[10]

2. HUMAN RIGHTS PROTECTION IN CYBERSPACE

Efforts to construct a global coordination and policy-making framework for the Internet began in the mid-1990s and to date remain unsuccessful.[11] There is no single state, or international body formally in overall charge of ensuring compliance with the law in respect of the way the Internet works.[12] Nor is there an overall treaty applicable to the Internet, although there are national laws and international treaties that are applicable to activities on the Internet.[13] In the context of international security, a broad consensus has been reached by the UN Group of Governmental Experts that, in principle, international law and in particular the Charter of the United Nations apply in cyberspace.[14] The UN Human Rights Council, in adopting Resolutions in 2012, 2014 and 2016,[15] together with the UN General Assembly (GA) adopting Resolutions in 2013 and 2014 on the right to privacy in the digital age, have asserted that international human rights law applies as much offline as online.[16] To that end, Resolution 69/166 called upon member states to review their practices and legislation on the interception and collection of personal data, including mass surveillance, to ensure the full and effective implementation of their obligations under international human rights law. Resolution 28/16 in 2015 also urged states to provide ‘an effective remedy’ and encouraged the Human Rights Council to identify ‘principles, standards and best practice’ for protection of privacy.[17]

The protection of human rights online has been a subject of international Internet governance[18] discourse for some time, including during the World Summit for the Information Society in 2003 and 2005. Not until the Snowden revelations, however, did the need for increased privacy protection gain importance and, subsequently, calls for the setting of international norms in relation to the interception of communications and data protection intensified. In 2013, the President of the Republic of Brazil, Dilma Rousseff, made a compelling case for the creation of ‘multilateral mechanisms for the worldwide network that are capable of ensuring principles such as freedom of expression, privacy of individuals and respect for human rights’.[19] Germany, leading a coalition of states, also proposed to enshrine digital privacy in an international human

[10] Supra note 3, para 33.

[11] Milton Mueller, et al. ‘The Internet and Global Governance: Principles and Norms of a New Regime’ (2007) 13 Global Governance.

[12] Council of Europe Commissioner for Human Rights, ‘The Rule of Law on the Internet and in the Wider Digital World’ (2014), 36.

[13] Council of Europe, Convention on Cybercrime (23 November 2001) ETS No 185; Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (1 October 1985) ETS 108.

[14] UNGA, ‘Report by Group of Governmental Experts on Development in the Field of Information and Telecommunications in the Context of International Security’ (24 June 2013) UN Doc A/68/98; UN Doc A/70/174 (22 July 2015).

[15] UNHRC Res A/HRC/RES/20/8 (16 July 2012); UNHRC Res A/HRC/RES/26/13 (14 July 2014); UNHRC Res A/HRC/RES/32/13 (18 July 2016).

[16] UNGA Res 68/167 (n 6).

[17] UNGA Res 28/16 (26 March 2015) UN Doc A/HRC/28/16.

[18] For a definition of Internet governance, see World Summit on Information Society, ‘Tunis Agenda for Information Society’ (2005) WSIS-05/Tunis/Doc/6(Rev. 1), 4.

[19] Statement by H.E. Dilma Rousseff, President of the Federative Republic of Brazil at the Opening of the General Debate of the 68th Session of the United Nations General Assembly (24 September 2013).

rights treaty by means of a new additional protocol to Article 17 ICCPR for the 'digital sphere'.[20] The idea, put forward at the 35th International Conference of Data Protection and Privacy Commissioners, was overwhelmingly supported by most of the privacy authorities, except for the United States (US).[21] Nevertheless, the opening of the negotiations on the additional protocol to Article 17 ICCPR conducted by the Special Rapporteur on Privacy, Professor Cannataci has begun.[22] The additional protocol is not envisaged, however, as 'one new global all-encompassing international convention covering all of privacy or Internet governance'.[23] The Special Rapporteur adopted a realistic approach, expecting that protection of privacy could be increased by incremental growth of international law through the clarification and eventually the extension of existing legal instruments. This seems to be a pragmatic solution, bearing in mind the number of unsuccessful attempts to reach an international agreement regarding the setting out of norms regulating state behaviour in cyberspace, in particular those of the Shanghai Cooperation Organisation in 2011 and 2015 introducing the *International Code of Conduct for Information Security*[24] to the UN General Assembly.

3. MODERNISING ARTICLE 17 ICCPR

In 1988, at the time when the General Comment No.16[25] on Article 17 ICCPR was adopted, the impact of advances in information and communication technologies on the right to privacy was barely understood, as the Internet was in its infancy. The paradigm shift in the way we communicate and the aggressive collection of personal information by many states have significantly undermined this right in recent decades. Consequently, there have been a number of calls for the Human Rights Committee (HRC) to draft a new general comment, most notably from UN Special Rapporteur Frank La Rue,[26] by the General Assembly,[27] and by civil society.[28] There are a number of reasons for updating General Comment No.16, and the fundamental starting point of this process must be articulating what the right to privacy actually means and protects, together with the scope of extraterritorial obligations of states under the human rights treaties. Both of these aspects will be discussed in turn below.

A. The Meaning of Privacy

The first step in modernising Article 17 must be the development of a better, more detailed and universal understanding of what is meant by 'right to privacy' in the 21st century.[29] The absence

[20] Ryan Gallagher, 'After Snowden Leaks, Countries Want Digital Privacy Enshrined in Human Rights Treaty', Slate (26 September 2013).

[21] 35th International Conference of Data Protection and Privacy Commissioners, Resolution on Anchoring Data Protection and the Protection of Privacy in International Law, (23-26 September 2013).

[22] UNHRC 'Report of the Special Rapporteur on the Right to Privacy, Joseph A. Cannataci' (8 March 2016) UN Doc A/HRC/31/64 para 46(j).

[23] Ibid.

[24] UNGA International Code of Conduct for Information Security (14 September 2011) UN Doc A/66/359; UNGA, 'Letter Dated 9 January 2015 from the Permanent Representative of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations Addressed to the Secretary General' (2015) UN Doc A/69/723.

[25] UNHRC, 'General Comment No.16: Article 17 (Right to Privacy). The Right to Respect of Privacy, Family, Home and Correspondence and Protection of Honour and Reputation' (8 April 1988) UN Doc HRI/GEN/1/Rev.

[26] UNHRC (n 1).

[27] UNGA Res 68/167 (n 6).

[28] UNHRC 'Written Statement by Reporters Without Borders International, a Non-Governmental Organisation in Special Consultative Status' (4 September 2013) UN Doc A/HRC/24/NGO/31.

[29] UNHRC (n 22), para 46(a).

of a universally agreed and accepted definition, and the different rates of economic development and technology deployment in diverse geographical locations, mean that the principles relating to privacy that were established fifty years ago at the time of drafting the ICCPR need to be further developed and supplemented to make them more relevant and useful to the realities of the modern era.[30] The debate on the understanding of what privacy is and should be has only just begun. However, some aspects for discussion regarding this concept have been put forward by the Special Rapporteur Cannataci as a useful starting point. Several countries, including Brazil and Germany, have written into their constitutions an overarching fundamental right to dignity and to free, unhindered development of an individual's personality.[31] Existing rights such as privacy, freedom of expression and freedom of access to information also constitute a tripod of enabling rights and, together with the fundamental right to dignity and the free and unhindered development of one's personality, would help to articulate how the concept of privacy should be understood in the modern age.

The definition of privacy must also encompass the idea of autonomy and self-determination, which in some countries such as Germany gives rise to a constitutional right to 'information self-determination'.[32] This idea is also referred to as 'informational privacy' and is concerned with the interest of individuals in exercising control over access to information about themselves.[33] This is in part already reflected in the current general comment to Article 17, according to which 'the gathering and holding of personal information on computers, databanks and other devices by public authorities or private individuals or bodies, must be regulated by law'.[34] The Human Rights Committee has applied this framework in several of its Concluding Observations[35] and this practice is also present in the jurisprudence of the European Court of Human Rights (ECtHR). The Court has held that the notion of 'private life' is 'not susceptible to exhaustive definition'[36] and has found on numerous occasions that 'protection of personal data is of fundamental importance to a person's enjoyment of respect for his or her personal data and family life'.[37] Article 8 of the Charter of Fundamental Rights of the European Union, explicitly recognises the right to protection of personal data separately and in addition to the right to privacy under Article 7.[38] In this regard, the Court of Justice of the European Union (CJEU) delivered a landmark decision in *Schrems v Data Protection Commissioner*,[39] holding that 'legislation permitting the 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 rights guaranteed by Article 7 of the Charter'.[40]

A new general comment should therefore affirm that Article 17 applies to informational privacy, which is understood as the individual's right to access and control personal data.

[30] Ibid.

[31] Id., para 25.

[32] Ibid.

[33] American Civil Liberties Union, 'Information Privacy in the Digital Age' (February 2015) <https://www.aclu.org/files/assets/informational_privacy_in_the_digital_age_final.pdf>

[34] UNHRC (n 25), para 10.

[35] UNHRC 'Consideration of Reports Submitted by States Parties under Article 40 of the Covenant, Concluding Observations, Spain' (2009) UN Doc CCPR/C/ESP/CO/5, para 11.

[36] *Bensaid v the United Kingdom* (App No 44599/98) (2001) ECHR para 47; *Botta v Italy* (App No 21439/93) (1994) ECHR.

[37] *MK v France* (App No 19522/09) (2013) ECHR; *S and Marper v the United Kingdom* [GC] (App Nos 30542/04 and 30566/04) (2008) ECHR.

[38] Charter of Fundamental Rights of the European Union, arts. 7 and 8, 2000/C 364/01 (12 December 2000).

[39] *Maximilian Schrems v Data Protection Commissioner* (6 October 2015) Case C-362/14.

[40] Id., para 94.

B. Extraterritorial Application of Human Rights Treaties

Governments may and do carry out surveillance both within and beyond their borders. However, the extent of the mass surveillance abroad, together with the international cooperation and the intelligence sharing among the Five Eyes partners, raises questions regarding these states' extraterritorial obligations under international law.

The jurisdictional scope of the ICCPR is set out in Article 2(1) of the Treaty and obliges member states 'to respect and to ensure' the rights recognised in the treaty 'to all individuals within its territory and subject to its jurisdiction'.[41] Similarly, Article 1 of the ECHR provides that state parties must secure to everyone within their jurisdiction the Convention's rights and freedoms.[42] The legislative frameworks, pursuant to which global surveillance of the Five Eyes operates, make a distinction between external and internal communications (e.g. UK Regulation of Investigatory Powers Act 2000 (RIPA)),[43] and the communications of nationals and non-nationals.[44] These laws differentiate between the obligations owed to nationals and those within the state's territory, and non-nationals who are outside state borders. For example, under ss.8 (1) and (2) RIPA, 'internal' communications may only be intercepted under a warrant which relates to a specific individual or address and may be granted on the basis of a suspicion of unlawful activity.[45] In cases of interception of 'external communications', defined as 'means of communication sent or received outside the British Islands',[46] ss.8(1) and (2) do not apply, which means that there is no need to identify any particular person who is to be the subject of the interception, or a particular address that will be targeted. The definition of 'external' communications, by the UK government's own admission, seems to encompass all activities of UK residents conducted through such platforms as Facebook, Twitter and Google, as their headquarters are located in the US.[47] This gives the UK intelligence agencies *carte blanche* to intercept all communications in and out of the UK, and means that UK residents are being deprived of the essential safeguards that would otherwise apply to them. Consequently, both UK residents' and foreigners' communications may be monitored indiscriminately under a 'general warrant' on the basis of s.8(4) RIPA 2000. The UN Human Rights Committee commented on this differentiation in its 2015 *Periodic Report on the United Kingdom*, stating that:

> the Regulation of Investigatory Powers Act 2000 (RIPA), that makes a distinction between 'internal' and 'external' communications, provides for untargeted warrants for the interception of external private communications and communication data, which are sent or received outside the United Kingdom without affording the same safeguards as in the case of interception of internal communications.[48]

[41] ICCPR (n 4), art 2(1).

[42] ECHR, (n 5), art 1.

[43] UK Regulation of Investigatory Powers Act 2000 s.8(4); Investigatory Powers Act 2016 s 136(3); New Zealand Government Security Bureau Act 2003 s.15A.

[44] US Foreign Intelligence Surveillance Act 1978 s 1881a(a); Australian Intelligence Services Act 2001 s 9; Canadian National Defence Act 1985 s 273.64(1).

[45] RIPA, (n 43), s 8(2).

[46] Id., s 20.

[47] *Privacy International v GCHQ*, Witness Statement of Charles Blandford Farr on Behalf of the Respondent (16 May 2014) IPT/13/92/CH.

[48] UNHRC 'Concluding Observations on the Seventh Periodic Report of the United Kingdom of Great Britain and Northern Ireland' (17 August 2015) UN Doc CCPR/C/GBR/CO/7.

The HRC urged the UK to:

> review the regime regulating the interception of personal communications and retention of communications data […] with the view to ensuring that such activities both within and outside the State party, conform to its obligations under the [International Covenant of Civil and Political Rights], including Article 17.[49]

Despite this recommendation, the new Investigatory Powers Act 2016 s.136(3) which seeks to reform the regime under which the UK law enforcement and security agencies perform their functions, allows that bulk interception warrants be issued to collect 'overseas related communications'.[50]

The issue of the extent of the human rights obligations of states' intelligence agencies conducting surveillance in cyberspace remains far from settled. The US government has consistently denied that it is bound by its obligations under the ICCPR, which the US ratified in 1992, with respect of acts done outside its physical territory.[51] It is therefore not legally bound to comply with the ICCPR in relation to its surveillance over non-US communications, or Internet activities. The US government's positon is that the Covenant obligations are restricted to situations when a person is both within a state's territory *and* subject to its jurisdiction.[52] This means that foreigners who do not satisfy both those conditions simultaneously do not benefit from the protection of the ICCPR.[53]

In the context of the UK state cyber surveillance, the Investigatory Powers Tribunal (IPT), which oversees the working methods of the intelligence agencies, has recently considered the issue of the extraterritorial human right obligations of the UK in *Human Rights Watch and Others v The Secretary of State for the Foreign and Commonwealth Office and Others (HRW v Secretary of State)*.[54] The case related to the interception, storage and use of information and communications by GCHQ of two groups of applicants – those resident in the UK and those who are not. Regarding the latter, the IPT ruled that the UK:

> owes no obligation under Article 8 ECHR to persons [who] are situated outside its territory in respect of electronic communications between them, which pass through that state.[55]

The IPT reasoned that foreigners not physically present in the UK, but subject to GCHQ surveillance under s.8(4) RIPA, do not have a right to privacy under Article 8 ECHR because they have not enjoyed a private life in the UK and therefore under Article 1 ECHR the UK is

[49] Ibid.

[50] RIPA (n 43).

[51] UNHRC 'Summary Record of the 1405th Meeting' (24 April 1995) UN Doc CCPR/C/SR.1405, para 20; UNHRC 'Consolidation of Reports Submitted by States Parties under Article 40 of the Covenant' (2005) UN Doc CCPR/C/USA/3.

[52] Id., para 20.

[53] US Department of State 'Second and Third Periodic Report of the United States of American to the UN Committee on Human Rights Concerning the International Covenant on Civil and Political Rights' (21 October 2005), Annex I.

[54] *Human Rights Watch Inc. and Others v The Secretary of State for the Foreign and Commonwealth Office and Others* [2016] ALL ER (D) 105 (May).

[55] Id., [60].

under no obligation to respect it.[56] In rejecting the extraterritorial application of the ECHR, the IPT adopted a conservative approach, based on *Bankovic v Belgium*[57] whereby, as a general principle of international law, jurisdictional competence of states is primarily territorial. The IPT was thus unwilling to 'extend the bounds of the UK Courts' jurisdiction under Art 8'.[58]

Ultimately, the issue of UK mass cyber surveillance abroad will be for the ECtHR to resolve in this and other cases.[59] Nonetheless, *HRW v Secretary of State* and the US consistent rejection of the extraterritorial application of the ICCPR obligations highlight the acute lack of transnational legislative instruments capable of addressing this issue. Suggestions have been made, however, that the 'effective control' over digital communications infrastructure, discussed in more detail below, may give rise to states' human rights obligations.[60]

C. Models of Extraterritorial Application of the ICCPR and the ECHR

The jurisdictional competence of a state is primarily territorial.[61] However, all major human rights courts and bodies, including the International Court of Justice (ICJ), the UN HRC, the Inter-American Commission on Human Rights (IACHR) and the ECtHR, agree that in some circumstances human right obligations may apply extraterritorially. This means that a state is bound by international human rights law in relation to individuals who may be not within its borders, but who are under its jurisdiction. To that end, a broadly similar approach, based on 'effective control', has been adopted to determine jurisdiction. Thus, the HRC has held that:

> a State Party must respect and ensure the rights laid down in the [International] Covenant [of Civil and Political Rights] to anyone within the power, or effective control of that State Party, even if not situated within the territory of the State Party.[62]

Similarly, the IACHR has established that, to determine whether a person is within a state's jurisdiction or not:

> the inquiry turns not on the presumed victim's nationality, or presence within a particular geographical area, but on whether under specific circumstances, the State observed the rights of a person subject to its authority and control.[63]

In conceptualising when and how the international human rights obligations may arise outside a state's territory, two types of extraterritorial jurisdiction were distinguished, namely the spatial and the personal models. The spatial model sees jurisdiction as effective overall control over a geographical area, whereas the personal sees it as a physical control over an individual. The spatial model was articulated by the ECtHR in *Loizidou v Turkey*,[64] where the Court held that a state's responsibility was engaged when, as a consequence of lawful or unlawful military

[56] Id. [58].

[57] *Bankovic and Others v Belgium* (App No 52207/99) (2007) 44 EHRR, 57.

[58] *HRW v Secretary of State* (n 54), [58].

[59] *Big Brother Watch v the United Kingdom* (App No 58170/13); *10 Human Rights Organisations v the United Kingdom* (Index No IOR 60/1415/2015); *Bureau of Investigative Journalism and Alice Ross v the United Kingdom* (App No 62322/14).

[60] OHCHR (n 3), para 34; Emmerson (n 7), para 41.

[61] *Bankovic* (n 57).

[62] UNHRC 'General Comment No. 31. The Nature of the General Obligations Imposed on State Parties to the Covenant' (2004) UN Doc CCPR/C/21/Rev.1/Add1326 May 2004, para 10.

[63] *Alexandre v Cuba*, Case 11.589, (1999) IACHR Report No. 109/99, para 37.

[64] *Loizidou v Turkey* (App No 15318/89) (1995) 20 EHRR 99.

action, it exercised effective control of an area outside its national territory. A similar approach was adopted by the ICJ in the *Wall* Advisory Opinion[65] and in *DRC v Uganda*,[66] where it was held that the ICCPR applies extraterritorially when a state is occupying territory of another state. Whilst the spatial model has its merits, particularly in its clarity and in setting some limits on states' obligations, it also has some drawbacks.[67] According to Milanovic, 'a state is perfectly capable of violating the rights of individuals without controlling the actual area', for example by using drones for targeted killing thus dispensing with the need to have troops on the ground.[68]

The jurisprudence of the international human rights courts has also recognised that states have human rights obligations when exercising physical control over an individual. In *Lopez Burgos v Uruguay*[69] the HRC held that state parties are liable for the actions of their agents on foreign territory, as it would be

> unconscionable to so interpret the responsibility under Article 2 of the [ICCPR] as to permit a State party to perpetrate violations of the Covenant on the territory of another State, which violations it could not perpetrate on its own territory.[70]

In its General Comment No.31, the Committee established that:

> a State Party must respect and ensure the rights laid down in the Covenant to anyone within the power or effective control of that State Party, even if not situated within the territory of the State Party […] regardless of the circumstances in which such power or effective control was obtained.[71]

However, by far the most varied jurisprudence regarding the personal model is that of the ECtHR. In *Al-Skeini v UK*[72] the Court stressed the primarily territorial nature of jurisdiction under the ECHR, but recognised exceptions to that principle, namely where state agents exercise authority and control extra-territorially, and when a state exercises effective control of an area outside national territory. State agent authority is particularly pertinent in military operations where physical authority and control is exercised in formal detention centres, as was the case in the British-controlled facilities in *Al-Skeini*. However, the exercise of authority was also held to have occurred outwith a formal detention centre in *Öcalan v Turkey*.[73] The case concerned the handover in Kenya to Turkish authorities of an individual suspected in Turkey of terrorist-related crimes. The ECtHR noted that he was effectively under Turkish authority and therefore within its jurisdiction, even though Turkish officials at the time of the arrest exercised their authority outside Turkey.

[65] Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territories (Advisory Opinion) (2004) ICJ Reports 163.

[66] *Case Concerning Armed Activities on the Territory of the Congo (Democratic Republic of Congo v Uruguay)* (2000) ICJ Reports 111.

[67] Marko Milanovic, 'Human Rights Treaties and Foreign Surveillance: Privacy in the Digital Age' (2015) 56 Harvard International Law Journal 81, 114-115.

[68] Id., 113.

[69] UNHRC *Lopez Burgoz v Uruguay*, Communications No 52/1979 (17 July 1979) UN Doc CCPR/C/13/D/52/1979.

[70] Id., paras 12.2-12.3.

[71] UNHRC (n 62), para 10.

[72] *Al-Skeini and Others v United Kingdom* [GC] (App No 55721/07) (7 July 2011) ECHR 2011.

[73] *Öcalan v Turkey* (App No 46221/99) (2003) 41 EHRR 985.

In addition, and perhaps most notably, the ECtHR has recognised that the extraterritorial jurisdiction on the basis of state agent authority or control is not limited to situations of the physical custody of an individual, but may be engaged when state agents exercise authority and control over an individual's rights, as was the case in *Jaloud v the Netherlands*.[74] The case concerned the fatal shooting of Azhar Sabah Jaloud, who at the time was passing through a checkpoint manned by personnel under the command and direct supervision of a Royal Netherlands Army officer in Iraq. The ECtHR found that the Netherlands exercised its jurisdiction on the basis that Dutch troops asserted 'authority and control over persons passing through the checkpoint' because they exercised authority and control over his right to life at that moment. This gave rise to extraterritorial jurisdiction, despite not having physical control over Mr Jaloud. The case therefore marks the ECtHR moving away from an approach wherein jurisdiction is founded on the basis of pure factual authority, towards one based on the exercising of authority and control over an individual's rights.

D. Applicability of Human Rights Treaties to Extraterritorial Cyber Surveillance

If a state may be found to have human rights obligations because it exercises authority and control over an individual's right to life, as proposed in *Jaloud*, then by analogy the exercise of control over their right to privacy of communications should also give rise to extraterritorial obligations in cases of foreign cyber surveillance. Such an interpretation seems necessary, given that the 'effective control' test is unsuitable, outdated and narrow in the context of state-sponsored cyber surveillance operations. It is outdated, because it has been articulated by international human rights courts and bodies long before digital technologies begun to play such a pervasive role in the lives of millions of individuals around the world. The existing approach is entirely inadequate for the cyber and communications realm, as it places the emphasis on the exercise of physical control over persons or territory, which is difficult to relate to cyberspace.[75] The shortcomings of the effective control approach centre around the fact that some state intelligence services, particularly the NSA, exert effective remote, rather than physical, control over much of the communications of foreign nationals abroad.[76] This occurs through eavesdropping on those communications, filtering, or altering their content, and breaking many forms of encryption by installing 'back doors' in many software systems.[77] The NSA also has the capacity to gain control of computers not directly connected to the Internet due to implantation of transmitting devices in computers manufactured in the US and elsewhere.[78] In addition, the US has relationships with Internet and telecommunications companies that facilitate surveillance, and therefore have the capacity to access directly the undersea cables and other carriers of Internet and telephonic communications.[79] The US's virtual power is unprecedented,[80] and the narrowly defined standard requiring physical control means that states interfering with the right to privacy would continue to exploit this gap by circumventing their human rights obligations. There can be no doubt, therefore, that the 'effective control' test must be adapted to suit the realities of cyber surveillance operations.

[74] *Jaloud v the Netherlands* (App No 47708/08) (2014).

[75] Peter Margulies, 'The NSA in the Global Perspective: Surveillance, Human Rights and International Counterterrorism' (2014) 82 Fordham Law Review 2137.

[76] Id., 2151.

[77] Ibid.

[78] Ibid.

[79] Ibid.

[80] Ibid.

A number of suggestions have been made, and their overall tenet seems to hinge on the control of communications, rather than physical control over areas or individuals. Thus, Nyst argues that when data or communications are intercepted within a state's territory, the state should owe obligations to those individuals regardless of their location on the basis of 'interface-based jurisdiction',[81] that is not to interfere with communications that pass through its territorial borders.[82] This approach is broadly in line with that proposed by Milanovic, who distinguishes between the overarching positive obligation of states to secure or ensure human rights, and extends even to preventing human rights violations by third parties and negative obligations of states to respect human rights that only requires states to refrain from interfering with the rights of individuals without sufficient justification.[83] This model conceptualises jurisdiction as a negative duty to refrain from interference and would apply to all potential violations of negative obligations, for example to refrain from interfering with privacy.[84] In this sense, human rights treaties would apply to most, if not all foreign surveillance activities.[85] Both these approaches have their merits, in as much as they recognise the weaknesses of the personal and spatial models and emphasise the negative duty of states not to interfere with protected rights. However, the nature and scope of the Five Eyes surveillance seems to go beyond the interception, collection and storage of data. The partnership between the US and its allied services allows governments to easily engage in the so-called 'collusion for circumvention'.[86] For example, GCHQ is allowed to spy on anyone except British nationals, whilst the NSA on anyone but Americans.[87] Information-sharing partnerships enable each agency to circumvent its respective national restrictions protecting their countries' citizens, since they are able to access the data collected by others.[88] This reciprocity has important ramifications on the domestic level if it is used to circumvent domestic legislation and limits on the governments' ability to tap its own citizens' communications.[89] In this context, the negative duty not to interfere with privacy would only be discharged if the interference is also understood as 'collusion for circumvention', encompassing such information sharing arrangements.

Given that this is not entirely clear, a sound candidate for a model of jurisdiction may be the 'virtual control' test, proposed by Margulies.[90] This test would make the ICCPR and other human rights treaties applicable when a state can assert 'virtual control' over an individual's communications, even though it lacks control over the territory in which the individual is located, or over the 'physical person' of that individual.[91] 'Virtual control' in this context means the ability to intercept, store, analyse and use communications. Although it could be argued that mere surveillance does not constitute physical control, it may constitute virtual control, in that it not only stifles their right to privacy, but also has a chilling effect on other human rights, such as free expression, freedom of conscience and religion, free assembly and association, and health, to name but a few. It therefore affects and controls individuals' behaviour.

[81] Carly Nyst, 'Interface Based Jurisdiction Over Violations of the Right to Privacy' (21 November 2013) EJIL:Talk! <http://www.ejiltalk.org/interference-based-jurisdiction-over-violations-of-the-right-to-privacy/>

[82] Ibid.

[83] Milanovic (n 67), 126.

[84] Ibid.

[85] Id., 129.

[86] Parliamentary Assembly of the Council of Europe, 'Mass Surveillance' Doc 13734 (18 March 2015), paras 30-3.

[87] Ibid.

[88] Ibid.

[89] Ibid.

[90] Margulies (n 75), 2139.

[91] Ibid.

Although the ‘virtual control’ approach has been criticised for being new and ‘without support in patterns of generally shared legal expectations about personal jurisdiction’,[92] it has a number of advantages. First, it corresponds to the notion of control developed and required by human rights courts and bodies,[93] outlined above. Secondly, it responds to the jurisdictional challenges of human rights obligations in surveillance cases, because the intelligence agencies under scrutiny are perfectly capable of controlling lives and private information with the press of the button.[94] Thirdly, it is in line with the ECtHR reasoning in *Jaloud v the Netherlands*, where a more expansive approach was taken and extraterritorial jurisdiction was established because of the state agents’ exercise of authority and control over the individual’s right to life, which made their physical proximity unimportant. Fourthly, such an approach would ensure equal treatment of all individuals, irrespective of their nationality or physical location, because establishing ‘virtual control’ over someone’s communications would not depend on where the interference takes place, but rather on whether or not a state can assert such control even when it lacks authority or control over the territory or the physical person. Finally, it could also mean that governments’ ‘collusion for circumvention’ arrangements may fall within their obligations not to interfere with the privacy rights, as they would have an obligation derived from the human rights treaties in relation to the rights of all individuals whose communications fall within their control, either inside and outside their territories.

It still remains unclear how cyber surveillance may trigger the extraterritorial application of human rights law. Although there is a general endorsement from international organisations that human rights treaties apply to extraterritorial cyber surveillance, no human rights body has yet directly addressed how electronic surveillance affects the right to privacy in detail. The Human Rights Committee has engaged with this issue, suggesting that extraterritorial surveillance does affect the ICCPR, when addressing the NSA surveillance pursuant to s.702 of FISA, stating that:

> the Committee is concerned about the surveillance of communications in the interest of protecting national security conducted by the National Security Agency (NSA) conducted both within and outside the United States.[95]

The United Nations Office of the High Commissioner also addressed extraterritorial surveillance noting that:

> digital surveillance […] may engage a State’s human rights obligations if that surveillance involves the State’s exercise of power or effective control in relation to digital communications infrastructure, wherever found, for example through direct tapping or penetration of that infrastructure. Equally, where the State exercises regulatory jurisdiction over a third party that physically controls the data, that State also would have obligations under the Covenant.[96]

[92] Jordan J. Paust, ‘Can You Hear Me Now? Private Communications, National Security and the Human Rights Disconnect’ (2015) 15(2) Chicago Journal of International Law 612(2015), 625.

[93] Ilina Georgieva, ‘The Right to Privacy under Fire-Foreign Surveillance under the NSA and the GCHQ and Its Compatibility with Art. 17 ICCPR and Art. 8 ECHR’ (2015) 31(80) Utrecht Journal of International and European Law 104.

[94] Ibid.

[95] UNHRC, ‘Concluding Observations on the Fourth Periodic Report of the United States of America’ (23 April 2014) CCPR/C/USA/CO/4, para 22.

[96] OHCHR (n 3), para 34.

Similarly, the Special Rapporteur Emmerson observed that the:

> State's jurisdiction is not only engaged where State agents place data interceptors on fibre-optic cables travelling through their jurisdictions, but also where a State exercises regulatory authority over the telecommunications or Internet service providers that physically control the data.[97]

The United Nations General Assembly, in adopting Resolution 68/167, appears to support the view that the ICCPR applies to extraterritorial surveillance, expressing its deep concern:

> at the negative impact that surveillance […] including extraterritorial surveillance […] in particular when carried out on a mass scale may have on the exercise and enjoyment of human rights.[98]

These approaches seem to broadly correspond with legal academic opinion articulating jurisdiction being triggered on the basis of states' control over the individual's rights to privacy. However, they leave unanswered the question of what degree of control is necessary to establish that a state exercises 'power or effective control in relation to digital communications infrastructure'. In *Jaloud* the ECtHR indicated its approach to the issues of authority and control based on the actual exercise of such powers over an individual's rights. Whether or not it will apply this or a similar approach to the pending surveillance cases[99] remains to be seen.

There can be no doubt that, as currently defined, the 'effective control' test of extraterritorial jurisdiction is not well suited for application to cyber surveillance operations. Cyberspace is a transnational environment where information is deliberately routed through a number of jurisdictions to reach its destination. When interference is conducted remotely, physical control over an area or an individual ceases to be relevant. At the very least, it leaves a gap that intelligence agencies can exploit to circumvent the obligations under the human rights treaties through the use of intelligence sharing agreements. What becomes important in this context is the 'virtual control' over the individuals' right to privacy, regardless of where they are located or their nationality. How these obligations may apply to cases of cyber surveillance remains unclear, especially bearing in mind the 'inevitable ripple effects on other scenarios such as extraterritorial use of lethal force through, for example drone strikes',[100] if more permissive approach to this issue were to be adopted. This makes the task of the Human Rights Committee when drafting new general comment on Article 17 particularly challenging.

4. CONCLUSION

In the age of increased terrorist threat, the balance between the need for security and the right to privacy of innocent individuals is particularly difficult to achieve. The vulnerability of this and other rights in the face of an unprecedented interference by states' intelligence agencies conducting mass surveillance must be addressed. This paper concentrated on

[97] Emmerson (n 7), para 41.

[98] UNGA Res 68/167 (n 6).

[99] Listed at note 59.

[100] Marko Milanovic, 'UK Investigatory Powers Tribunal Rules that Non-UK Residents Have No Right to Privacy under the ECHR' (2016) EJIL: Talk!

one legal solution – the overhaul of Article 17 ICCPR, and in particular re-defining the concept of privacy and addressing the question of how and when states may be liable under international law for their surveillance activities, the effect of which may be felt beyond their borders. The paper has illustrated that the narrowly defined territorial limitations on human rights protection based on nationality (e.g. s.702 FISA), or geographical distinctions (s 8(4) RIPA; s.136 IPA) are meaningless when applied to highly integrated global communications networks. The surveillance conducted on these legal bases, coupled with the states' 'collusion for circumvention' places practically no limitation on the extent to which governments can access the communications of millions of individuals in their own and other countries. Yet the enjoyment of fundamental rights is not limited to citizens of particular states, but includes all individuals, regardless of nationality. Although the jurisprudence of the international human rights courts recognises that there are certain circumstances when extraterritorial human rights obligations will be engaged based on the 'effective control' test, this paper has highlighted its limitations in the context of cyber surveillance and has proposed that the 'virtual control' test – understood as a remote control over an individual's right to privacy – may be a solution to this problem.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

The Misuse of Protected Indicators in Cyberspace: Defending a Core Aspect of International Humanitarian Law

Jeffrey Biller, Lt Col, USAF
Stockton Center for the Study of International Law
U.S. Naval War College
Newport, RI, USA
jeffrey.biller@usnwc.edu

Abstract: International humanitarian law (IHL) imposes a complex array of laws regarding the use of markings, signals, symbols and other indicators. Protections related to indicators are also directly implicated in the laws of perfidy and ruses. Although these laws are generally well accepted in principle, practitioners struggle to apply these rules in the newer, man-made domain of cyberspace. Despite recent steps forward in the application of IHL to cyber, questions surrounding enemy, neutral, and protected indicators remain largely unresolved. This paper seeks to answer these thorniest of issues related to military cyber operations during international armed conflicts.

The article is divided into two sections. The first addresses protected and specially recognized indicators, particularly those of the UN and the Geneva Conventions. The IHL rules regarding these symbols are defined and applied in the context of cyber operations. This section also discusses perfidy and proximate causation in the cyber context. The second turns to the improper use of national indicators in cyberspace, particularly the definition of military emblems, which draws on a separate body of law than protected or specially recognized emblems. Although the misuse of indicators may also implicate international criminal law, this article focuses exclusively on IHL applicability.

Keywords: *cyberspace, markings, indicators, emblems*

1. INTRODUCTION

A core principle of international humanitarian law (IHL) is the protection of civilians and civilian objects.[1] This protection includes aid organizations such as the International Committee of the Red Cross (ICRC) and observer organizations such as the United Nations (UN). These groups, in addition to neutral states not party to the conflict, are distinguished through the use of various indicators[2] governed by an extensive body of law dating back to ancient times.[3] IHL divides these indicators into two primary categories: first, the protected and recognized indicators of the Geneva Conventions (GC), the UN, the white flag of truce, and other internationally recognized protective emblems, signs or signals;[4] and second, indicators of nationality, such as the uniforms or military equipment markings of neutral or adversary nations.[5] Related to the use or misuse of indicators under IHL is the prohibition on perfidy, which outlaws killing or injuring by resort to acts inviting the confidence of an adversary leading to a reliance on protection under the rules of IHL with the intent to betray that confidence.[6]

The basic notion of extending the body of IHL regarding these indicators into cyberspace is uncontroversial.[7] However, a full agreement does not yet exist as to what constitutes recognized indicators in the cyber domain and how to realize the protections signaled by these indicators.[8] This article examines the treaty and customary rules related to the use of indicators and then applies those rules to many of the network characteristics currently used to identify entities in cyberspace. Although some characteristics meet the definitions of relevant indicators under IHL, this article identifies gaps where markings indicating a trusted party could be used to conduct offensive cyber operations.

To that end, the article is divided into two sections. The first addresses protected and specially recognized indicators, particularly those of the UN and the Geneva Conventions (GC). The IHL rules regarding these symbols is defined and applied in the context of cyber operations. This section also discusses perfidy and proximate causation in the cyber context. The second section turns to the improper use of national indicators in cyberspace, particularly the definition of military emblems, which draws on a separate body of law than protected or specially recognized emblems. Although the misuse of indicators may also involve international criminal law, this article focuses exclusively on IHL applicability.

[1] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 48, June 8, 1977, 1125 UNTS 3 (hereinafter AP I).

[2] The word "indicators" will be used throughout this paper to generally encompass the set of uniforms, emblems, flags, etc. that are used to indicate nationality, status, special protections, or particular categories.

[3] Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, ¶ 1526 (Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., 1987) (hereinafter AP I Commentary).

[4] Convention (I) for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field art. 44, 53, Aug. 12, 1949, 6 UST 3114, 75 UNTS 31 (hereinafter GC I); AP I, art. 38; Convention for the Protection of Cultural Property in the Event of Armed Conflict, May 14, 1954, 249 UNTS 240 see also 1 Customary International Humanitarian Law, r. 58-61 (Jean-Marie Henckaerts & Louise Doswald-Beck eds., 2005) (hereinafter CIHL Study).

[5] AP I, art. 39; CIHL Study, r. 62-63.

[6] CIHL Study, r. 65.

[7] See *Tallinn Manual on the International Law Applicable to Cyber Warfare* r. 60-65 (Michael N. Schmitt ed., 2013) (hereinafter *Tallinn Manual*).

[8] Id.

2. PROTECTED AND RECOGNIZED INDICATORS

A. The Law Regarding Improper Use of Protected and Recognized Indicators

The long-standing IHL rules against the improper use of protected and recognized indicators such as the emblems of the GCs and the UN are well-established.[9] This law developed as recognition of the need to protect certain classes of individuals, organizations, and locations on the battlefield from targeting by combatants. As such, the law focuses primarily on these emblems' use as concrete, visible representations.[10] Although it is unlikely that the use of protected indicators in a purely electronic environment was initially envisaged, the language within the relevant articles is broad enough to encompass its extension into the cyber domain.

The First Geneva Convention (GC I) defines the emblem of the Red Cross and delineates its permissible use.[11] Specifically, GC I states that the emblem, and the words "Red Cross" or "Geneva Cross,"[12] "may not be employed either in time of peace or in time of war, except to indicate or to protect the medical units and establishments, the personnel and material protected by the present Convention and other Conventions dealing with similar matters."[13] Similarly, Article 38 of Additional Protocol I (AP I) prohibits the "improper use of the distinctive emblem of the red cross, red crescent or red lion and sun or of other emblems, signs or signals provided for by the Conventions or by this Protocol" and also "to make use of the distinctive emblem of the United Nations, except as authorized by that Organization."

The 2016 Commentary to GC I (GC I 2016 commentary) notes that the GC emblems may serve both as a protective device indicating protection under the Convention and as an indicative sign demonstrating a connection to the organization of the International Red Cross and Red Crescent.[14] Although the indicative use does not imply that the bearer holds protections under the Convention, its improper use is still prohibited.[15] AP I does not address the indicative use, focusing on the protective use,[16] which provides "a visible sign of the protection conferred by international law on certain persons and objects."[17] Unlike misuse of the emblem as an indicative sign, a misuse of the protective function could implicate the prohibition on perfidy.[18]

GC I Article 53 further expands the law relating to the GC emblems, prohibiting their use "by individuals, societies, firms or companies either public or private, other than those entitled

[9] See *CIHL Study*, r. 59-60; GC I, art. 44, 53. AP I, art. 38; Regulations Respecting the Laws and Customs of War on Land, annexed to Convention No. IV Respecting the Laws and Customs of War on Land art. 23(f), Oct. 18, 1907, 36 Stat. 2227, TS. No. 539 (hereinafter Hague Regulations).

[10] Commentary to Geneva Convention I for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field 325-330 (Jean Pictet ed., 1952) (hereinafter GC I 1952 Commentary).

[11] GC I, art. 38, 44, 53.

[12] Hereinafter, only the words "Red Cross" will be referenced, but both phrases are implicated.

[13] GC I, art. 44.

[14] International Committee for the Red Cross, Commentary to Geneva Convention I for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field ¶ 2661 (2d ed. 2016), art. 44 (https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=EC42EEC3274A7323C1257F7A0056F100) (hereinafter GC I 2016 Commentary).

[15] Id.

[16] AP I Commentary, ¶ 1540.

[17] International Committee for the Red Cross, Regulations on the Use of the Emblem of the Red Cross or the Red Crescent by the National Societies art. 1 (1991) (https://www.icrc.org/eng/resources/documents/article/other/57jmbg.htm).

[18] AP I Commentary, ¶ 1532.

thereto under the present Convention, of the emblem or the designation "Red Cross" or "Geneva Cross," or any sign or designation constituting an imitation thereof, whatever the object of such use."[19] By including "imitations thereof," Article 53 broadens the prohibition and suggests that abbreviations or approximations of the words "Red Cross" that are meant to imitate an official representation would violate this prohibition.[20] Although many practitioners view Article 53 as primarily applying to peacetime misuse of the emblem in its protective sense, the GC I 2016 commentary states that Article 53 applies "both in situations of armed conflict and in times of peace."[21] Additionally, the GC I 2016 commentary notes that Article 53 "encompasses both misuse of the emblem in its protective sense and when used as an indicative sign."[22]

AP I also prohibits the unauthorized use of the distinctive emblem of the UN.[23] However, the treaty law governing the UN emblem is less expansive than that of the Red Cross emblems, primarily in the scope of its definition. Unlike the words "Red Cross" under GC I, IHL protects neither the words "United Nations" nor approximations thereof.[24] Therefore any application of law regarding the GC and UN emblems in the cyber context must start with an understanding that protections against GC emblems will have broader applicability.

Additional categories of protected emblems, signs, and signals established under international law include the Hague IV and AP I prohibition against the "improper use of a flag of truce"[25] and the AP I prohibition against deliberate misuse in an armed conflict of "other internationally recognized protective emblems, signs or signals."[26] Recognized protected indicators include those markings that indicate objects or locations such as installations containing dangerous forces, cultural property, POW or civilian internee camps, civil defense, and hospital, safety, or neutralized zones.[27]

Unlike the prohibition on perfidy (see below), there is an absolute character to these prohibitions, meaning that there is no requirement for a particular result following the prohibited misuse.[28] Examples of the improper use of protected emblems, signs, and signals include their use by other than intended personnel while engaging in attacks, to favor one's own military operations, or to impede enemy military operations.[29] These examples encompass almost the entire potential range of military operations, as any relevant action undertaken by a military is likely to either favor their own, or disfavor the enemy's, operations.

One section of law that would initially appear relevant to cyber operations is the rules contained in Annex I to AP I governing "distinctive signals." However, the signals referenced in Annex I are very specific types of radio communication and light signals.[30] Although some

[19] GC I, art. 53.

[20] GC I 2016 Commentary, ¶ 3065, art. 53 (https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=57F199148260B5AFC1257F7A00579E9B).

[21] GC I 2016 Commentary, ¶ 3066, art. 53.

[22] Id.

[23] AP I, art. 38(2).

[24] Id.

[25] Hague Regulations, art. 23(f).

[26] AP I, art. 38.

[27] AP I, art. 59-60, 66; Office of the General Counsel, U.S. Department of Defense, *Law of War Manual* (hereinafter *DoD Law of War Manual*), para 5.24; CIHL Study, r. 61.

[28] AP I Commentary, ¶ 1532.

[29] *DoD Law of War Manual* § 7.15.4 (2016).

[30] AP I, Annex I, art. 6-14.

conceivable use in cyber operations is plausible, there are no generally applicable rules to cyber operations that flow from them and therefore will not be discussed. With an understanding of the prohibition related to protected and recognized indicators in place, the following section analyzes its application in cyberspace.

B. Improper Use of Protected and Recognized Indicators in Cyberspace

Extension of the basic rule prohibiting "mak[ing] improper use of the protective emblems, signs, or signals that are set forth in the law of armed conflict" into the cyber domain is uncontroversial.[31] However, protected indicators signal the ability to trust, and trust plays a prominent role in network security systems, which depends on forming trust relationships between parties before allowing access and sharing information. Masquerading as a party known to be trusted by a target system is a frequently used method of defeating network security. Therefore, rules related to the cyber indicators of trusted parties, such as the ICRC and UN, require detailed understanding. This section examines particular cyber methods that involve violations of the trust relationship, including the variations on phishing, Internet Protocol (IP) spoofing, and domain name spoofing, as a way of contextualizing and exploring these rules.

First is the use of phishing, a type of social engineering, to manipulate authorized system users into providing information and thus allowing unauthorized system access.[32] This manipulation occurs in the purely cyber context through the use of e-mail, e-messaging, or online communications. The Tallinn International Group of Experts (IGE) addressed this situation, citing the example of an adversary sending an email with the "bare assertion that the sender is a delegate of the International Committee of the Red Cross."[33] The IGE found no misuse in this example, despite the use of the words "Red Cross." Although GC I Article 44 specifically protects these words from unauthorized use, the presumed argument is that the operator's use of the words "Red Cross" is not formal enough to be considered as an emblematic identifier. However, if the words were employed in a more formal manner, such as an email signature block, letterhead to an attachment, or another manner formally indicating an official Red Cross document, there is a much stronger argument that the use violates the GC I Article 44 prohibition on use of the words "except to indicate or to protect the medical units and establishments, the personnel and material protected by the present Convention and other Conventions dealing with similar matters."[34]

The difference between the simple statement as to ICRC affiliation and a signature block is the formality involved. Signature blocks are typically used in official correspondence and serve to indicate the sender's representative status of the named organization, whereas simple statements, while misleading, lack that level of formality. As a Red Cross signature block "is used to show that a person or object has a connection with one of the organizations of the Movement, without implying protection under the Geneva Conventions or any intent to invoke them," this would be considered an indicative use, as opposed to protective use, of the words "Red Cross." Nonetheless, it constitutes a violation of GC I Article 44 and the customary IHL related to the use of the distinctive GC emblems.[35]

[31] *Tallinn Manual*, r. 62.

[32] Michael Gregg, *Certified Ethical Hacker Guide* 513 (2014) (hereinafter *CEH Guide*).

[33] *Tallinn Manual*, cmt. accompanying r. 62, ¶ 4.

[34] GC I, art. 44.

[35] CIHL Study, r. 59.

The second type of operation is a related type of phishing campaign, but with the aim of tricking the target operator into taking cyber-based self-defeating actions. This method uses spoofed emails, social media messaging, and websites to induce the target into either downloading malicious attachments or following web links to malicious websites.[36] Like other types of social engineering, these attacks rely on the target operator trusting the e-mail, website or attachment such that they will take the desired action. Protected emblems could easily be implanted into the e-mail, message or website to induce trust in the target. For example, the use of the UN emblem as a watermark in a downloadable document or the Red Cross emblem as a social media avatar could induce a target to follow a link to a site containing malware. As the actual protected emblem is clearly used in an unauthorized manner, this is a clear IHL violation. The Tallinn IGE came to the same conclusion on this question.[37]

A third method illustrating misuse of emblems is IP spoofing. Here, cyber operators attempt to gain unauthorized system access by creating a malicious message that appears to originate from a trusted machine, imitating its IP address.[38] For example, spoofing an IP address associated with the ICRC to defeat a firewall that relies on IP addresses for filtering.[39] The primary question is whether IP addresses should be viewed as a legal indicator of a protected organization. This appears logical, given the widespread use of IP addresses as a trust indicator by cyber operators. For example, a defensive operator may specifically program a firewall to permit connections from ICRC or UN IP addresses during an armed conflict. These connections may allow communications regarding the treatment of wounded or prisoners or war. If an adversary were to spoof these IP addresses, the network operator may be forced to block communications from these previously trusted sources.

Permitting a party to a conflict to represent a communication as coming from the ICRC or UN appears to run counter to the intent of IHL. Article 1 of AP I states it is a general principle that in cases not covered by AP I or other international agreements, customary law, "principles of humanity," and "dictates of public conscience" apply. Additionally, Article 31(1) of the Vienna Convention on the Law of Treaties (VCLT) states that a treaty should be interpreted partly "in the light of its object and purpose."[40] However, the VCLT also states that treaties should be interpreted "in accordance with the ordinary meaning to be given to the terms of the treaty in their context." Provisions governing use of the emblems suggest an element of general awareness or recognition of the emblem as such.[41] Thus, it is unlikely that a spoofed ICRC IP address could be considered an imitation of the emblem under the Article 53 standard given the lack of general awareness as to what the sequence of numbers in an IP address specifically indicates. The ordinary person is unlikely to mistake an IP address for an imitation of the words "Red Cross," even if a cyber operator may understand the connection. Despite their importance in identifying organizations in cyberspace, IP addresses do not meet any definition of the GC emblems. Given that the provision governing use of the UN emblem is less broad than that of the GC emblems, it is also doubtful UN-associated IP addresses would be considered as a protected indicator.

[36] Ed Skoudis & Tom Liston, *Counter Hack Reloaded* 566 (2006).

[37] *Tallinn Manual*, cmt. accompanying r. 63, ¶ 2.

[38] Skoudis, at 470.

[39] Id.

[40] Vienna Convention on the Law of Treaties art. 31(1), May 23, 1969, 1155 UNTS 331.

[41] The United Nations Flag Code and Regulations, ST/SGB/132, United Nations, January 1967; GC I 2016 Commentary, ¶ 2543, art. 38 (https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=8BF732335A87E6DCC1257F15004A27A5).

The fourth method for analysis involves the spoofing of e-mail addresses or domain names. By spoofing an email address, such as @ICRC.org, in the recipient's "From" field, the operator hopes to induce either the target system to allow the email through a firewall, or a target individual to trust the contents of the email. Once this trust is established, the operator may then use this connection to conduct the next phase of a cyber operation. Similarly, a Domain Name System hijacking operation may send an unwitting target who accesses the ICRC.org or UN.org websites to a spoofed website containing malicious links or false information.[42]

Here, the focus is on domain names which serve to provide users with a recognizable identity to resources found on the Internet. Although related to IP addresses, domain names differ in that they often contain an organization's name or abbreviation, as opposed to the numerical designator of an IP address. The narrower protection for the UN emblem, which does not include the name United Nations or approximations, eliminates its applicability from this analysis. The relevant question as to the Red Cross is whether a spoofed email address or domain name containing the words Red Cross, the acronym ICRC, or similar abbreviation, would constitute "an imitation thereof."[43] The Tallinn IGE "struggled with the issue," and laid out two potential approaches.[44]

The first approach argued that email address and domain names are not protected indicators because they do not constitute "electronic reproductions of the relevant graphic emblems."[45] This approach overlooks the prohibitions in Articles 44 and 53 on the unauthorized use of the words "Red Cross" or "an imitation thereof" when they function as an indicative or protective emblem. The 1952 Commentary to GC I expresses this concern well:

> It was obviously not enough merely to prohibit misuse of the red cross emblem. Protection had also to be extended to the words which form the official title of the great humanitarian institution known as the Red Cross. These words are as familiar to the public as the emblem, and must enjoy the same prestige.

The second approach found the key factor to be the "use of an indicator upon which others would reasonably rely in extending protection provided for under the law of armed conflict"[46] Thus, the imitation of the ICRC.org domain name or email address would be an unauthorized use because, as the IGE states, it "invites confidence as to the affiliation of the originator."[47] Although it does not reference Article 53, this view would be consistent with that article's inclusion of any "sign or designation constituting an imitation thereof." Given the ubiquitous use of the acronym ICRC, it would be hard to argue that it does not constitute an "imitation thereof." Therefore, the second approach of the IGE appears to be a more accurate reflection of IHL.

The various methods of phishing and spoofing are not the only types of cyber operations that implicate the rules against misuse of protected emblems. However, they highlight the most likely ways in which protected indicators may be used in a remote access cyber operation. They

[42] Skoudis, at 220-221.

[43] GC I, art. 53.

[44] Tallinn Manual, cmt. accompanying r. 62, ¶ 6.

[45] Id.

[46] Id., cmt. accompanying r. 62, ¶ 7.

[47] Id.

also serve to help identify which cyber indicators could constitute protected indicators and reveal gaps where adversaries could take advantage of the trusted nature of organizations such as the ICRC and UN to conduct offensive cyber operations.

C. Perfidy and Proximate Causation in Cyberspace

A discussion of the use of protected emblems naturally raises the issue of perfidy. Although the improper use of protected indicators is itself an IHL violation, it may also constitute an element of perfidy. Several questions surround the application of the rule against perfidy in the cyber context. However, this article focuses specifically on proximate causation between the perfidious act and the injury or killing of the adversary. Proximate causation is an act that directly produces an effect and without which the effect would not have occurred. Although there is no requirement for proximate causation written into the treaty law on perfidy, both the Tallinn Manual and commentators have extrapolated such a requirement from the definition of perfidy.[48]

Perfidy is well-defined in treaty and customary law. AP I Article 37(1) defines perfidy as:

> […] acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence.

Perfidious acts that result in the killing or wounding of the enemy constitute a war crime.[49] Included in the ICRC CIL study's list of acts that constitute perfidy is the "simulation of protected status" through the use of the Red Cross, United Nations emblems, or other protected emblems.[50] Simulation of status can be accomplished in any manner intending to inspire confidence, and is therefore distinct from the "improper use" standard discussed previously. Despite this distinction, as with the improper use of protected emblems, the concept of violating trust makes perfidy an important concern in cyberspace.

The proximate causation requirement appears to come from the language of AP I Article 37(1), stating that it is "prohibited to kill, injure or capture an adversary by resort to perfidy." This language suggests a direct link between the perfidious act and the end result, and is reflected in the *Tallinn Manual* commentary.[51] The proximate cause example in the *Tallinn Manual* focuses on proximate causation after the kinetic effect produced by the perfidious cyber action.[52] However, the cyber operation should also be evaluated to determine whether the perfidious act was the proximate cause of the kinetic effect. If there is a distinct, additional phase to a cyber operation required between the perfidious act and the kinetic effect, then there is no proximate causation between the perfidious act and resulting death or injury.

Cyber operations are rarely single-step operations resulting in a malicious cyber effect. Rather, they are multi-stage efforts, with most stages resulting in no discernible effect on the target. Initial stages may be limited to developing an understanding of, and then gaining access to,

[48] *Tallinn Manual*, cmt. accompanying r. 60, ¶ 5; Michael Bothe, Karl Josef Partsch & Waldemar A. Solf, New Rules for Victims of Armed Conflicts, Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949, at 244 (1982).

[49] CIHL Study, r. 65.

[50] Id.

[51] *Tallinn Manual*, cmt. accompanying r. 60, ¶ 5.

[52] Id.

a target network.[53] The perfidiously-gained access may be used for espionage purposes, for a sub-use of force non-cyber effect, or for no action whatsoever. Often a cyber operator has no other goal than to gain access to a system, with the use of that access to be determined at a later time. If the perfidious act, such as the use of a protected emblem, takes place during a phase of the cyber operation distinct from the action causing the non-cyber effect, then the perfidious act is unlikely to be the proximate cause of the death or injury. The amount of time between the perfidious act and the resulting death or injury is not determinative of proximate causation, though it may be an element in the analysis. Rather, the key question in determining proximate causation is whether the perfidious act directly produced the injury or death. If an intervening step is required to produce the injury or death, the misuse would not constitute prohibited perfidy. A similar category of indicators subject to IHL that could potentially implicate perfidy are enemy and neutral indicators, which will be discussed below in detail.

3. IMPROPER USE OF ENEMY AND NEUTRAL INDICATORS

A. The Law Governing Use of Enemy and Neutral Indicators

Indicators of national military status (uniforms, flags, and other military emblems) do not inherently indicate a protective status under international law. However, rules regarding the use these indicators are among the earliest found in warfare.[54] In current practice, Hague Regulation 23(f) and Article 39 of AP I are the primary treaty laws governing the use of these indicators. The Hague regulation prohibits the "improper use of a flag of truce, of the national flag or of the military insignia and uniform of the enemy." AP I Article 39(2) defines "improper use" as uses of adverse party indicators "while engaging in attacks or in order to shield, favour, protect or impede military operations." The commentary to AP I notes that this "includes the preparatory stage to the attack."[55]

In addition to treaty law, there is a variety of state opinion regarding enemy indicators. Many states consider the Article 39(2) rule to only apply in instances of attack, with varying reservations to the broader inclusion of those uses "in order to shield, favour, protect or impede military operations."[56] For example, Canada made such a reservation upon ratification of AP I[57], and the US believes that the prohibition extends to combat, but use outside combat would not be improper.[58] The ICRC CIL Study states that "it cannot be concluded, therefore, that the wearing of enemy uniforms outside combat would be improper." It should also be noted that the question as to whether use of enemy uniforms would constitute perfidy is unsettled, as no specific IHL protections attach to enemy uniforms.[59] However, the ICRC CIL Study notes that the wearing of such uniforms may invite the confidence of the enemy.[60]

Article 39(1) of AP I also covers the use of neutral party indicators, prohibiting the "use in an armed conflict of the flags or military emblems, insignia or uniforms of neutral or other States

[53] CEH guide, at 42.
[54] *See* AP I Commentary, ¶ 1526; and CIHL Study, r. 62.
[55] AP I Commentary, ¶ 1586.
[56] For discussion on state interpretations of the rule, *see* CIHL Study, r. 62.
[57] Canada, Reservations and statements of understanding made upon ratification of Additional Protocol I.
[58] *DoD Law of War Manual*, § 5.32.1.
[59] Id. at § 5.32.1.1.
[60] CIHL Study, r. 62.

not Parties to the conflict."[61] The AP I standard only requires connection to an armed conflict. According to the commentary, the prohibition includes the use for espionage purposes, "as this constitutes an intervention of a military nature."[62] The ICRC CIL Study found no contrary state practice or claims to the AP I Article 39(1) rule regarding neutral indicators.[63] The use of neutral military emblems may also be considered perfidious.[64]

Although these prohibitions are relatively well-defined, they do leave open the question of what constitutes an enemy or neutral indicator. Unlike the well-defined emblems of the Red Cross and the UN, the "military emblem" referenced in AP I is undefined in the text of the treaty, although the commentary does provide an in-depth discussion of the matter. There, emblems of nationality are described as "essentially customary in nature," meaning those uses in international society that "constitute a generally recognized language which is accorded the same respect as the spoken or written word in relations between individuals."[65] The commentary also notes that military emblems are "visible signs."[66] Though the description of military emblems is somewhat vague, it is clearly much broader than the definitions given the distinctive emblem of the Red Cross or of the UN. This broad definition is the key to determining the applicability of the law regarding enemy and neutral indicators in cyberspace.

B. Improper Use of Enemy Indicators in Cyberspace

The *Tallinn Manual* rule on enemy indicators holds that "[i]t is prohibited to make use of the flags, military emblems, insignia, or uniforms of the enemy while visible to the enemy during an attack, including a cyber attack."[67] This rule adds the AP I commentary requirement of visibility and drops the phrase "or in order to shield, favour, protect or impede military operations." In the cyber context, the meaning of the visibility requirement is key to determining the rule's applicability.

The Tallinn IGE postulated that "it is unlikely that improper use of the enemy uniforms and other indicators will occur during a remote-access cyber attack, as the cyber operators would not be in visual contact with the adversary."[68] However, under treaty provisions and the AP I commentary explanation of military emblems, there is no requirement that the attacker is physically visible to the adversary for its use to be improper, only that the military emblem is visible.[69] The minority position of the IGE pointed to the fact that such a provision "appears in neither Article 39(2) of Additional Protocol I, nor in the ICRC Customary IHL Study's discussion of that Article."[70]

The majority opinion is supported by historical precedent, and to some extent common sense, where concrete, visible signs attached to military uniforms and equipment allowed distinction between parties to the conflict.[71] However, IHL does not require that an attacker, much less

[61] AP I, art. 39(1).
[62] AP I Commentary, ¶ 1569.
[63] CIHL Study, r. 63.
[64] Id., r. 65.
[65] AP I Commentary, ¶ 1562.
[66] Id. at ¶ 1578.
[67] *Tallinn Manual*, r. 64.
[68] Id., cmt. accompanying r. 64, ¶ 3.
[69] AP I, art. 39(2); AP I Commentary, ¶ 1562-1587.
[70] *Tallinn Manual*, cmt. accompanying r. 64, ¶ 4.
[71] Bothe et al., at 214.

the attacker's uniform or other military indicator, be visible to the intended target. Although cyber operations were not considered when the rule emerged, the development of other types of beyond visual range weaponry, such as artillery, certainly existed. Additionally, the AP I commentary notes that in modern warfare, military equipment is:

> […] supplied in large numbers throughout the world by a few manufacturers. They are all the same model, and very often it is only the emblems of nationality which unequivocally identify to which side they belong.[72]

The same rationale can be applied to the Internet and other networked information systems. National domains and systems are often uniquely identified at both the human and system interface levels.

If military emblems are to be extended to cyberspace, the next issue is what qualifies as a military emblem in cyberspace under IHL. The Tallinn IGE did raise the question of enemy marked computer hardware, finding the "rule has no application with regard to enemy marked computer hardware over which control has been remotely acquired and that is used for conducting attacks against the enemy." However, the Tallinn IGE did not extend the discussion to visible enemy network identifiers such as domain names.

Given the requirements for military emblems to be visible and identifiable in a "generally recognized language," cyber identifiers with no visual component, such as meta-data, and those that may be visible but are not generally recognized, such as an IP address, should not be considered as military emblems. However, commonly understood identifiers that appear at the human interface level, such as web and e-mail domain names, graphical symbols, and formal representations such as signature blocks in e-mails and electronic documents appear to qualify as military emblems. These identifiers provide official representation of a particular organization or individual. In this manner, they act just as a military emblem, designed to provide official notification of a combatant's status as a member of a particular party to the conflict. For example, there is a general trust that a domain name of www.af.mil is operated by, or at least under the control of, the United States Air Force. Similarly, an electronic letter bearing the official seal of the United States Navy and the signature block of the Chief of Naval Operations would indicate that the sender was a member of the United States Navy. Thus, if used in a cyber operation as part of an overall military attack campaign, they initially appear to constitute an improper use of an enemy military emblem. Prior to making this determination, however, the practitioner must analyze the issue of state practice concerning transmission of bogus signals, dispatches, and other communications considered as ruses of war.[74]

Ruses of war are defined in AP I as:

> […] acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in an armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under that law.[75]

[72] AP I Commentary, ¶ 1572.

[73] *Tallinn Manual*, cmt. accompanying r. 64, ¶ 6.

[74] Id., cmt. accompanying r. 64, ¶ 5.

[75] AP I, art. 37(2).

For example, the UK *Manual* includes in its list of permissible ruses:

> […] transmitting bogus signal messages and sending bogus dispatches and newspapers with a view to their being intercepted by the enemy; making use of the enemy's signals, passwords, radio code signs, and words of command.[76]

These types of ruses find application in the cyber context, including most spoofing of enemy e-mail addresses, social messaging identifiers, and graphical symbols included in attachments, as these uses of military emblems are communicative in nature.[77]

Although most uses of enemy cyber identifiers would find function in permissible ruses, some cyber indicators that meet the definition of military emblems have the potential to fall outside this realm. Take, for example, the spoofing of a military web domain identified by a military domain address, such as "www.navy.mil." Domain names serve to identify networks, or portions thereof, as belonging to particular organizations. This is distinct from a communicative purpose such as relaying false information. Any example of an attack using a domain name to conduct an attack might be a weaponized honeypot using an enemy's spoofed domain name against their own forces, as that domain name represents a military emblem that falls outside the category of acceptable ruse. Here, it is not the communicative nature of the domain name that is used to conduct an attack, but rather the enemy's reliance on it as an indicator of their own forces.

Types and functions of cyber identifiers are likely to evolve at a rapid pace and any attempt to create an exclusive list of identifiers held to be military emblems would have limited utility. Fortunately, the definition described in the AP I commentary, describing emblems as "customary nature" allows the understanding of military emblems in cyber to evolve in CIHL as the technology changes. However, this understanding requires close coordination between legal practitioners and technical experts to determine what indicators should be considered as military emblems for IHL purposes. The definition of a military emblem also plays a key role in law regarding improper use of neutral indicators, to be discussed in the following section.

C. Improper Use of Neutral Indicators in Cyberspace

Similar to protected emblems, the prohibition on the improper use of neutral indicators is absolute "in an armed conflict."[78] Thus they cannot be used "for the promotion of the interests of a Party to the conflict in the conduct of that conflict."[79] There are no required elements of use in an attack, visibility, or any required result of the use.[80] Additionally, unlike false representations of enemy communications, those that are represented to come from neutral parties are not considered permissible ruses of war.[81] Therefore improper use of the whole potential range of previously discussed military emblems, such as web and e-mail domain names, graphical symbols, and formal representations such as signature blocks in e-mails and electronic documents are also prohibited under this rule.

[76] United Kingdom Ministry of Defence, *The Manual of the Law of Armed Conflict* ¶ 5.17.2 (2004) (hereinafter UK Manual).

[77] CIHL Study, r. 57; *Tallinn Manual*, cmt. accompanying r. 64, ¶ 5.

[78] AP I, art. 39(1).

[79] AP I Commentary, ¶ 1565.

[80] Id.

[81] CIHL Study, r. 57, 63.

This broad prohibition is reflected in the Tallinn Rule 65, stating that "[i]n cyber operations, it is prohibited to make use of flags, military emblems, insignia, or uniforms of neutral or other States not party to the conflict."[82] Curiously, the Tallinn IGE raise the issue of "employment of other reliable indicators of neutral status," by referencing the discussion of protected indicators in Rule 62 and the UN emblem discussed in Rule 63, rather than the more legally relevant discussion of military emblems in Rule 64.[83] Again, protected emblems and the UN emblem draw on separate law for definition and application than military emblems. Therefore, for example, discussion of the improper use of the ICRC.org domain name must necessarily be distinct from that of the navy.mil domain name, although they may sometimes come to the same conclusion.

Combining the broad definition of military emblems, the absolutist nature of the prohibition, and the lack of customary practice regarding use of neutral indicators in communications by parties to the conflict,[84] there is little room left for the possible legal use of neutral military emblems in cyberspace by parties to the conflict. Still excluded are indicators that do not meet the definition of military emblems, such as IP addresses. The only exception to this rule is matters not related to the conduct of hostilities, such as "matters of police or civil administration."[85]

4. CONCLUSION

The balancing of military necessity and humanity is not an equation that can be definitively solved. As nations, non-state actors, and the methods and means of warfare continue to evolve, the law seeking to provide that balance will need to evolve with it. However, there are some enduring provisions at the core of IHL that states must always defend. One of these is the protection of groups seeking to enable that IHL balance during an armed conflict. The protected indicators of the GC, the recognized emblem of the UN, and the military emblems that identify neutral parties are vital to protecting these categories. With warfare moving into the human-made domain of cyberspace, every effort must be made to identify what these indicators look like in the new domain.

This paper has argued that current cyber indicators such as domain names, e-mail addresses, electronic signature blocks, and graphically displayed emblems can constitute the traditionally understood indicators in warfare. However, there are also many gaps that could be exploited by parties to a conflict. For example, IP addresses cannot be understood as protected indicators because they are not generally understood by participants, but they can be instrumental in providing a trust relationship between information systems. This trust could be exploited by adversaries and therefore undermine the effectiveness of the organizations represented by the indicators. Given the broad acceptance of the importance of protected and recognized indicators, state should seek to close these gaps in the law and defend a core aspect of IHL.

[82] *Tallinn Manual*, r. 65.
[83] Id., cmt. accompanying r. 65, ¶ 4.
[84] CIHL Study, r. 63.
[85] AP I Commentary, ¶ 1565.

REFERENCES

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, June 8, 1977, 1125 UNTS.

Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949 (Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., 1987).

Convention (I) for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field, Aug. 12, 1949, 6 UST 3114, 75 UNTS 31.

Convention for the Protection of Cultural Property in the Event of Armed Conflict, May 14, 1954, 249 UNTS. 240

Customary International Humanitarian Law (Jean-Marie Henckaerts & Louise Doswald-Beck eds., 2005).

Tallinn Manual on the International Law Applicable to Cyber Warfare (Michael N. Schmitt ed., 2013).

Regulations Respecting the Laws and Customs of War on Land, annexed to Convention No. IV Respecting the Laws and Customs of War on Land, Oct. 18, 1907, 36 Stat. 2227, T.S. No. 539.

Commentary to Geneva Convention I for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field (Jean Pictet ed., 1952).

International Committee for the Red Cross, Commentary to Geneva Convention I for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field (2d ed. 2016) [https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=EC42EEC3274A7323C1257F7A0056F100].

International Committee for the Red Cross, Regulations on the Use of the Emblem of the Red Cross or the Red Crescent by the National Societies (1991).

Office of the General Counsel, U.S. Department of Defense, *Law of War Manual* (2016).

Michael Gregg, *Certified Ethical Hacker Cert Guide* (2014).

Ed Skoudis & Tom Liston, *Counter Hack Reloaded* (2006).

Vienna Convention on the Law of Treaties, May 23, 1969, 1155 UNTS 331.

The United Nations Flag Code and Regulations, ST/SGB/132, United Nations, January 1967.

Michael Bothe, Karl Josef Partsch & Waldemar A. Solf, New Rules for Victims of Armed Conflicts, Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949 (1982).

Canada, Reservations and statements of understanding made upon ratification of Additional Protocol I.

United Kingdom Ministry of Defence, *Manual of the Law of Armed Conflict* (2004).

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Update to Revolving Door 2.0: The Extension of the Period for Direct Participation in Hostilities Due to Autonomous Cyber Weapons

Tassilo V. P. Singer
Research Associate
Public and Administrative Law, Public International Law,
European and International Economic Law
University of Passau, Germany
tassilo.singer@gmx.de

Abstract: The rule concerning the direct participation in hostilities (DPH) by civilians is one of the most controversial rules within the law of armed conflict. While civilians are generally guaranteed protection by the principle of distinction, DPH provides for a loss of protection 'for such time as they take a direct part in hostilities'. This temporal component of DPH poses significant challenges in light of the use of autonomous cyber weapon systems (ACWS), as the active part of the civilian can be reduced to the second of activation.

Autonomy in this context means that the ACWS is able to operate without any human control, rendering human influence on the actions impossible. As soon as the ACWS is fully self-operating, the common interpretation based on the original wording of the DPH-rule would suggest that DPH is no longer possible. Consequently, a civilian hacker using ACWS cannot lawfully be attacked after control of the system has been relinquished even if damage occurs or the attack is recognised later. As a result, civilian hackers are legally privileged without any discernible justification. In order to remedy this unsatisfactory situation, this article suggests an extension of the relevant period of time for DPH to the whole period of the operation of the ACWS to solve the 'Revolving Door 2.0' problem. It is further submitted that concerns that such an extension would unduly broaden the scope of the DPH-rule can be met by the requirement that, additionally, the regular cumulative criteria for DPH have to be fulfilled.

Keywords: *direct participation in hostilities, loss of protection as a civilian, timeframe of direct participation in hostilities, autonomous cyber weapon systems*

1. INTRODUCTION

The protection of civilians under IHL ceases following their direct participation in hostilities (DPH), according to Article 51(3) AP I. However, the protection is regained after return or the end of the activity.[1] This transition regularly poses problems in practice, but the revolving door of protection and civilian DPH is even more critical when the use of autonomous cyber weapons (ACWS) is involved.

Imagine critical parts of a state's military networks like unmanned combat aerial vehicle (UCAV) control have been hacked, leading to the UCAVs being turned against the state's own troops. After the type and source of the attack has been identified, it turns out the malware used has been operating for months without any human control. In such a case, the predominant legal interpretation is that the perpetrator must not be attacked as a civilian directly participating in hostilities. According to the common understanding, the loss of protection of a civilian is terminated after the last moment of control, even if the attack continues or its effects occur later.[2] This article will focus on the discrepancy between current interpretation and reality, and argues towards an extension of the time period of DPH as a solution to the dilemma posed by the use of ACWS by civilians.

2. DIRECT PARTICIPATION IN HOSTILITIES

According to the principle of distinction set out in Articles 48 and 51(2) AP I,[3] civilians may not be the object of an attack, and such an attack does not form a part of the military advantage for a conflict party.[4] However, if civilians decide to take up arms and fight against a conflict party, the balance between humanity and military necessity[5] would be negatively affected. It would be unacceptable to grant protection to civilians while they attack combatants, but the latter may not counter the attack due to the general rule of distinction. Therefore, the rule of Article 51(3) AP I provides for a loss of protection for the period of time during which civilians directly take part in hostilities to sanction the participation in hostilities by an originally protected person.[6] However, the interpretation of – and state practice on – the rule for DPH is fairly controversial.[7]

As the meaning of DPH is neither specified in treaty law nor clarified by sufficient state practice or international jurisprudence,[8] the rule is open for interpretation under Article 31 of the Vienna

[1] *ICRC, Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law*, Nils Melzer (ed.) (May 2009), 67.

[2] Marco Roscini, *Cyber Operations and the Use of Force in International Law* (1st edn OUP 2014) 209; Heather Harrison-Dinniss, 'Participants in Conflict – Cyber Warriors, Patriotic Hackers and the Laws of War' in Dan Saxon (ed), *International Humanitarian Law and the Changing Technology of War*, 251-278, 274-276.

[3] *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226, para. 78.

[4] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (2nd edn, 2010) 123-125.

[5] Id., 4-5.

[6] Michael N. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyberwarfare* (Cambridge University Press 2013), 104-105, 118-119; Roscini (n 2), 203.

[7] ICRC (n 1); Program on Humanitarian Policy and Conflict Research at Harvard University, *Commentary on the HPCR Manual on the International Law Applicable to Air and Missile Warfare (hereinafter HPCR Manual)*, Rule 28, 119.

[8] ICRC (n 1), 41.

Convention.[9] Direct participation derives from the notion of 'taking no active part in the hostilities' of Common Article 3 GC I–IV, so active and direct participation share a common background.[10]

Article 51(3) AP I has three elements. First, the civilian has to take a direct part, which is an element of directness and or immediacy.[11] It may be better to choose a narrow interpretation of the directness to increase the protection of civilians.[12] This can be countered, however, by arguing that a wide interpretation of the rules would encourage innocent civilians to stay away from the hostilities as far as possible.[13] The second element is 'in hostilities', and indicates a certain nature or threshold of harm. Finally, 'for such time' implies a time-frame[14] and indicates only a temporary loss of protection. The content of this period is the decisive legal question concerning autonomous cyber programs. If one interprets the wording strictly, it only refers to the actual conduct of the civilian. After this period of direct participation, a civilian generally regains protection.

According to the ICRCs study on DPH, three criteria have to be met cumulatively under Article 51(3) AP I: the threshold of harm; a causal link between the act and the intended or inflicted harm; and the belligerent nexus, meaning the act must directly be related to the hostilities.[15] The threshold of harm means the required 'specific acts of war which by their nature or purpose are likely to adversely affect military operations or capacity'[16] or to cause a certain damage to protected targets.[17] Besides potential attacks on military targets,[18] it remains unclear whether it is necessary that damage to protected civilian targets actually occurs, or if potential danger to civilian targets suffices.[19]

On direct causation, the ICRC study determines that either the act has to be one causal step between act and effect,[20] or the relevant contribution of the civilian must constitute an integral part of a coordinated military operation.[21] However, the direct character is not precluded if the 'effects occur' with a delay, meaning 'some time after the malware is inserted'.[22] Also, due to the 'integral part' requirement, group-based actions which alone would not suffice for DPH can be taken into account if these are connected to a specified operation.[23]

The third criterion, the belligerent nexus, means that the relevant act has to be linked 'in some direct way [...] to the armed conflict'[24] and consequently has to be 'specifically designed to

[9] Vienna Convention on the Law of Treaties (VCLT), 1155 UNTS 331.

[10] *The Prosecutor v. Akayesu*, ICTR, case no. ICTR-96-4-T (1998), paras. 175, 182, 582; *Public Committee Against Torture in Israel et al v The Government of Israel et al*, HCJ 769/02, Supreme Court, 11 December 2005 (hereinafter *Supreme Court*), para 34; ICRC (n 1), 43.

[11] *Supreme Court* (n 10), para. 35.

[12] Id., para. 34.

[13] Ibid.

[14] Michael N. Schmitt, 'Deconstructing Direct Participation in Hostilities: The Constitutive Elements' (2010) International Law and Politics, Vol. 42, 728-729, 738; Supreme Court (n 10), para. 38.

[15] ICRC (n 1), 46; 47 ff, 51 ff, 58 ff.

[16] Id., 47.

[17] Ibid.

[18] Compare Schmitt (n 14), 715-716.

[19] See Roscini (n 2), 204-205.

[20] Id., 206.

[21] ICRC (n 1), 51.

[22] Roscini (n 2), 207.

[23] *HPCR Manual* (n 7), Rule 29, 120, para. 3; id, 207.

[24] Schmitt (n 14), 735.

directly cause the [...] threshold of harm in support of a party and to the detriment of another'.[25] The design of the act does not depend on the subjective intent of the actor.[26] However, the ICRC study suggests that:

> the conduct of a civilian, in conjunction with the circumstances prevailing [...], can reasonably be perceived as an act designed to support one party [...] by directly causing [...] harm to another party.[27]

These findings have been criticized for various reasons.[28] First, the term 'threshold' in the first criterion was considered misleading because the issue should be the nature of harm and not a set threshold.[29] Second, the outcome that one causal step is necessary can also be objected to, as it would be impractical and contradictory to ask for an immediate consequence of the act and allow for a temporal distance.[30] Third, the requirement of direct causation was welcomed, but seen as used too restrictively. Schmitt holds that the integral part criterion[31] should not only be limited to coordinated military operations, but should also extend to individuals.[32] Thus, the act of the civilian must be more generally 'an integral part of the conduct that adversely harms one party and benefits the other party to a conflict'.[33] Concerning the belligerent nexus, one can criticise that the act must be linked directly to the causation of such harm and the causation of the benefit, and Schmitt suggests an alternative approach focusing on a link of the act either to the support or the detriment of one party.[34]

Even if there is disagreement concerning the exact content of the three criteria recommended by the ICRC, the critics agree that at least the three constitutive criteria can be applied[35] and have to prevail independently of the exact legal content in every situation of DPH by a civilian.

For the discussion of the particular problem of ACWS, an international armed conflict has to prevail[36] and a civilian has to act. A negative definition of who should be considered a civilian can be found in Article 50 AP I.[37] Also the discussion regarding organised armed groups (and a continuous combat function in this context)[38] can be disregarded.

[25] ICRC (n 1), 58.

[26] Id., 59; Schmitt (n 14), 735-736.

[27] ICRC, (n 1), 63-64.

[28] Compare as overview of critical views on the ICRC study by Boothby, Schmitt, Watkin and Hays Parks and a response by Melzer: *New York Journal of International Law and Politics* 42 (2009-10).

[29] Schmitt (n 14), 716.

[30] Id., 728.

[31] ICRC (n 1), 51.

[32] Schmitt (n 14), 729 ff.

[33] Id., 739.

[34] Id., 736.

[35] Id., 738; *Tallinn Manual* (n 6), 119.

[36] The rule is viewed as customary international law and can be applied in non-international armed conflicts: Jean-Marie Henckaerts, Louise Doswald-Beck, *Customary International Humanitarian Law*, Vol I, Rule 5, 17; *Supreme Court* (n 10), para. 38.

[37] Referring to Article 4 A (1),(2), (6) of GC III and Article 43 AP I.

[38] Compare Roscini (n 2), 200 f.

3. DIRECT PARTICIPATION IN HOSTILITIES BY THE USE OF AUTONOMOUS CYBER WEAPONS

A. Technical Peculiarities

Under these legal circumstances, a big challenge is posed by new cyber tools like conditioned and delayed code, scripts or malware[39] which have autonomous behaviour. After insertion or after the start of the program's working process, the malware's capabilities may include self-guidance, self-reproduction, and even redefinition and adaptation. Many variations of such software are conceivable.[40] All of these have in common that they require only a very short period of human interaction with the software in order to start process. Human participation can be reduced to pressing 'enter', to sending an email, or to integrating the malware somewhere. From the moment when the malware operates independently and is not controllable by a human any more, a cyber tool can be called autonomous.

Software and programs can calculate extremely quickly and thereby make decisions in less than a fraction of a second. In the last decade, the speed of data transmission via cable or satellite has increased enormously, correlating with a growing interconnectivity worldwide. As the amount of software and code has expanded exponentially, so have their weaknesses, as the potential contact surface for attacks has increased.[41] Tools for encryption of data and communication have also spread out and become more sophisticated, further complicating retracing and attribution.[42] These factors, combined with certain malware which itself can act independently without any human control, pose significant challenges for detection and set the factual framework for the envisaged problem.

Malware can contain multiple tools with sub-functions (as weapons) separately and independently from one another comparable to a weapon system[43] in the original sense. The exact qualification of a cyber tool as a weapon, a weapon system, or more generally as a mean of warfare does not matter for the legal problem discussed here as long as the cyber tool has the potential to cause the required damage to constitute a DPH. As the law of armed conflict applies to every weapon system,[44] this is also true of the DPH-rule in the context of ACWS.

B. Transfer of the Constitutive Criteria to ACWS

The constitutive criteria have to be transferred to the use of ACWS by a civilian. In the aforementioned situation, the manipulation of the UCAV control systems by an ACWS leads to friendly fire by the UCAVs. As military targets have been attacked, the required threshold of harm is met. Arguably, the manipulation of the military control system could be considered sufficient in itself, even if no material damage is caused but military operations have been hampered.[45] The same would apply if an autonomous program was able to penetrate a military

[39] Malicious software, compare *Tallinn Manual* (n 6), Glossary.

[40] Compare CERT-UK, *Common Cyber Attacks: Reducing the Impact*, 2015, last viewed 22.12.2016, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/400106/Common_Cyber_Attacks-Reducing_The_Impact.pdf ; US-CERT, National Cybersecurity and Communications Integration Center, *DDos Quick Guide*, last viewed 22.12.2016, available at: https://www.us-cert.gov/sites/default/files/publications/DDoS Quick Guide.pdf.

[41] William A. Owens, Kenneth W. Dam and Herbert S. Lin, *Technology, Policy, Law and Ethics* (1st edn., National Academies Press 2009), 80-108.

[42] Ibid.

[43] For differentiation compare: *HPCR Manual* (n 7), Means: Rule 1 (t); Weapon: Rule 1 (ff).

[44] *Nuclear Weapons* (n 3), para.86.

[45] *Tallinn Manual* (n 6), 119.

communication network and interrupt the communication links by continuously changing the IP-allocation of the central communication hub.

A civilian starts an ACWS by using a computer and software to send or activate the tool. This act by the civilian has to be viewed as the last causal step before the ACWS causes the damage. The ACWS, if programmed properly, will run continuously until it reaches its goal without any means for the human to control it. An ACWS checking the military network for the UAV control and finally using an exploit in the code to manipulate or take over control is comparable to a fired missile, even if there is a much longer interval upon effect. An additional causal step is not possible or necessary. Even if this view is not shared, the act of activation (and possibly the preparation of the ACWS) by the civilian is at least an integral part of the specified operation and thus constitutes the required direct causation.

Finally, the act of the civilian to start the ACWS to harm the UAV control is at least designed to harm the adversary party in the conflict. Generally, the acts of the ACWS have to be viewed as one complex in relation to the belligerent nexus. If this complex is designed to either benefit one party or directly cause harm to the other, the nexus is given.

C. Acts of Participation in Hostilities

The relevant acts of a civilian directly participating are often the preparation, the act of direct participation itself, and the return.[46] Transferred to autonomous cyber weapon systems[47] the elements could look as follows.[48] The ACWS first has to be prepared, a process which could consist of coding or at least an acquisition process. During the preparation a target or a framework for potential targets has to be set. It is problematic to distinguish between having just to prepare the software without further participation, and the necessary integral part, meaning the preparation of a malware with direct causation.

During the preparation the civilian cannot generally be lawfully attacked.[49] The question then is when exactly the civilian becomes a legitimate target in the latter case. One could draw the line and find a sufficient act of preparation as soon as the software is shaped for implementation or specified[50] in order to cause a certain form of damage (e.g. the malware is shaped to the specifics of a common control-software for system-processes in military networks).[51] Ergo, the software has to have the potential to damage relevant targets.[52] The process has to be developed to the point that the damage can occur even if the code is not yet sophisticated enough to avoid any detection. However, the monitoring of someone preparing an ACWS is hardly ever possible to accomplish given the speed, the internationality, and the interdependence of cyberspace.

[46] ICRC (n 1), 65; *HPCR Manual* (n 7), 118, Rule 28-29; *Tallinn Manual* (n 6), 121.

[47] Compare DPH and the use of unmanned systems: Dorota Banaszewska, 'Kombattanten und Zivilisten weit weg vom Schlachtfeld' in Robert Frau (ed.), *Drohnen und das Recht* (Mohr Siebeck 2014).

[48] See, for a different description for types of actions: Roscini (n 2), 204.

[49] Compare ICRC (n 1), 53, 68; *HPCR Manual* (n 7), 120, para. 3; Bill Boothby, ' "And for such time as": The Time Dimension to direct participation in Hostilities', [2010] *International Law and Politics* 42, 748, 752; Roscini (n 2), 201 (concerning continuous combat function), 207-209.

[50] See ICRC (n 1), 52-54; Roscini (n 2), 208.

[51] It is necessary to differentiate between military targets, where the likelihood definitely suffices and civilian targets which may require a damage to occur. Roscini (n 2), 205-206.

[52] See ICRC (n 1), 47; Roscini (n 2), who refers to 'objective likelihood that the act will result in such harm', 206.

D. The Problem of DPH Using ACWS – Revolving Door 2.0

The act of participation itself might be observable, e.g. a civilian shooting at a military convoy with a gun. Looking at the use of ACWS, the situation is far more difficult. The period wherein the civilian is active could be only a few seconds and the necessity to act is often just the process to let the ACWS start its self-guidance; the effect and damage usually occur later. In practice, the ACWS becomes recognisable to the victim only in this moment due to detection and retraceability problems.[53]

Unfortunately, due to the regular time delay the action of the perpetrator will commonly have been over for weeks or even months. In such a situation, the civilian would not directly participate in hostilities any more, and consequently would enjoy legal protection again. A comparable situation is the revolving door problem known from the war against insurgents in Afghanistan and articulated in the 'farmer by day, fighter by night' problem.[54]

However, in this case the problem is raised to a new level. Even if all available surveillance tools are used, the identification of the origin and the proof of the use of ACWS itself is much more difficult in this short amount of time. Because the circumstances in the context of ACWS are even more challenging and the technical and legal limits even tighter, the problem could be called revolving door 2.0.

The decisive factor is the time-frame 'for such time' which is a prerequisite of the legal exception of DPH. In a case concerning DPH, the Israeli Supreme Court quoted the AP I commentary, proposing that the time-frame should neither be interpreted too narrowly nor too widely.[55] An international group of experts found that DPH contains:

> all actions immediately preceding or subsequent to the qualifying act. In a cyber operation, this period might begin once an individual begins probing the target system for vulnerabilities, extend throughout the duration of activities against the system, and include the period during which damage is assessed to determine whether re-attack is required.[56]

Concerning delayed effects, the majority found:

> that the duration of the individual's direct participation extends from the beginning of his involvement in mission planning to the point when he or she terminates an active role in the operation. [...] Note that the end of the period of direct participation may not necessarily correspond with the point at which the damage occurs.[57]

[53] Compare: Marco Roscini, 'Evidentiary Issues in International Disputes Related to State Responsibility for Cyber Operations', [2015] *Texas International Law Journal*, 234-238; Wissenschaftliche Dienste des Bundestags, *Anwendbarkeit des humanitären Völkerrechts auf Computernetzwerkoperationen und digitale Kriegsführung (Cyber Warfare)*,(2015) WD 2 – 3000 – 038/15, 10-11; Different view: Russell Buchan, Cyber Warfare and the Status of Anonymous under International Humanitarian Law, [2016] *Chinese JIL*, 767.

[54] *Supreme Court* (n 10), para. 40; *HPCR Manual* (n 7), Rule 28, 119, para. 5; ICRC (n 1), 70-71; Boothby (n 49), 753-758.

[55] *Supreme Court* (n 10), para. 34; Jean De Preux, *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*, 516, para. 1679.

[56] *Tallinn Manual* (n 6), 121.

[57] Ibid.

If one sticks closely to the wording and just considers the time during which the civilian is active as the moment of losing protection, the period of DPH ceases after activation[58] and the civilian directly participating by using ACWS will again become legally privileged. The civilian may not be attacked after the action,[59] or at the time when nearly all of these actions will be noticed due to the practical peculiarities of the autonomous behaviour of the ACWS.

By a narrow understanding of the legal criteria, the time-frame in which the person is directly participating and the time when one can recognise the actions differ widely. Therefore, sticking to the original, restrictive understanding[60] would mean creating a legal grey area. A civilian must be afraid of becoming a legitimate target only during preparation and in the short period of activating the ACWS. Thus, a civilian user of ACWS could continue such actions and is encouraged to do so without the threat of retaliation. This could lead to an escalation of the use of ACWS by civilians in future conflicts. Hence, a different legal perception and understanding of the criterion 'for such time' is needed.

4. SOLUTION FOR THE REVOLVING DOOR 2.0

A. Proposal for Solution

For the solution of this dilemma, it is proposed that the legally relevant time-frame be extended from the act to the whole action, meaning the ongoing cyber operation. As long as the ACWS is working, the person responsible for activating the program has to be considered as directly taking part in hostilities.[61] As soon as the damage has occurred, meaning the end of this period, the time-frame to react has to be narrowed down until the moment when an appropriate and proportional reaction is no longer possible depending on the single case. This means that, until the program has reached its final destination and stopped its damaging purpose, plus an appropriate and proportional time to react after recognition, the civilian loses their protection based on Article 51(3) AP I.

In theory, ACWS could operate endlessly and so in theory the protection could never be regained. However, this would be in contrast to the exceptional character of the rule and the wording indicating a temporary period ('for such time'), which implies the end of the loss of protection, too. Therefore, the extension of the time period 'for such time' has to be understood as applying only for the time that the ACWS is actively operating and cumulatively coinciding with the set constitutive criteria for DPH.

B. Possible Critique and Replies to the Critique

Nevertheless this view can be criticized for several reasons. The ICRC states that:

> any extension [...] beyond specific acts would blur the distinction made in IHL between temporary, activity-based loss of protection (due to DPH), and continuous, status or function-based loss of protection [due to combatant status or continuous combat function].[62]

[58] ICRC (n 1), 65, 67-68.
[59] Buchan (n 53), 766-767.
[60] Critic: Boothby (n 49), 743.
[61] Compare: Dinstein (n 4), 148; Ibid., 758; *Tallinn Manual* (n 6), 121.
[62] ICRC (n 1), 44-45.

Concerning remote attacks, the ICRC guidance states that:

> the duration of direct participation in hostilities will be restricted to the immediate execution of the act and preparatory measures [...].[63]

This apparently rejecting statement, however, does not have to mean that the use of an ACWS cannot extend the period of DPH. If the execution phase also encompasses 'the period over which the [malware] is installed or deployed' or 'the period over which [...] the targeted systems are compromised' and 'to the time period over which the victim actually suffers the effects',[64] the use of ACWS leads to a DPH by a civilian.

The arguments on a narrow or wide interpretation of the DPH also apply here: a narrow understanding of 'for such time' increases the protection of civilians using ACWS. The criticism that there is no military necessity to attack a civilian 'who is no longer playing a role in the operation'[65] aims in the same direction. The script or code of the ACWS may still be running while the civilian is attacked.[66]

On the other hand, a narrow interpretation would lead to a privilege for a civilian using such ACWS. These civilians will be encouraged to continue participating in hostilities,[67] when they realise that their actions cannot or do not provoke sanctions. This contravenes the original purpose of the protection of civilians in exchange for refraining from hostilities.

A compromise could be to fall back on criminal law and to try to get hold of the person instead of attacking them.[68] However, this requires territorial control or the cooperation of the host state of the perpetrator, which is hardly likely, and does not counter the threat of a continuing ACWS.

A civilian who wants to desist from attacks by ACWS may have to inform concerned conflict parties about the ACWS's existence in order to regain protection,[69] as the danger posed by such tools can be unlimited in time.

Finally, the military necessity can be found in the balance of humanity and military necessity itself.[70] This 'subtle equilibrium' has to be preserved[71] under all circumstances. Additionally, the prevention of future attacks could be considered necessary, too.

Often the situation is compared to a civilian placing a mine or an IED, who is regarded as not directly participating after its return, completing the action (the revolving door problem),[72] but these two situations are only superficially comparable. A minelayer could in theory be under surveillance during preparation, and a mine is physical and can be found by technical

[63] Id., 68.
[64] Owens, Dam, and Lin (n 41), 90.
[65] Roscini (n 2), 209.
[66] Dinniss (n 2), 274-276.
[67] *Supreme Court* (10), para. 34.
[68] Id., para. 40.
[69] Boothby (n 49), 757.
[70] Compare: Ibid., 767; Buchan (n 53), 768.
[71] Dinstein (n 4), 5.
[72] Dinniss (n 2), 275-276; Buchan (n 53), 767.

means. With ACWS, the fog of war is much denser; ACWS are non-physical and cyberspace works at high speed. The interconnectivity offers endless possible connections and there are many possibilities to hide the origin of an attack in the data transfer chain. If the perpetrator's computer is not linked to the Internet, state hacked or continuously observed, a civilian DPH using ACWS cannot be watched in nearly any case.

C. Arguing in Favour of an Extension of the Time Period

The period during which the civilian loses protection is not unlimited due to the necessary prevalence of the cumulative requirements. In particular, the loss of the belligerent nexus due to a provable withdrawal could prohibit a loss of protection concerning the hit of an arbitrary target.

Another argument in favour of the extension can be based on the wording of the rule. The phrase 'for such time' does not prevent an expansive interpretation if one views the running ACWS as the continuing act of participation. The use of ACWS can be considered as an extended arm of the human acting behind it. The system as a tool fulfils the set duties or frameworks and acts as the human would. Thereby the ACWS substitutes the human element of directness and immediacy.[73]

The same applies to the need for the requirement of only one causal step.[74] As soon as the malware is activated, it works autonomously and no other causal step by a human is needed. Even the ICRC guidance points out that causal proximity and temporal proximity do not have to coincide.[75]

The act of activation of the ACWS is a physical act[76] and an integral part to cause the result, which should not be underestimated. As soon as an ACWS is set free it will act independently and can rarely be stopped, comparable to an artillery shell or an unsophisticated rocket. The activation is the last step that the human has available to refrain from an attack using an ACWS. One could argue that the belligerent nexus is not given, as an arbitrary target is chosen by the ACWS, which was not originally intended by the civilian. Therefore the establishment of the necessary link could also be unintended. But even if the ACWS does not attack a specifically designated target, as long as there is a link to the armed conflict (and some argue that as long as it is either in support of a party or to the detriment of another, not both), the required nexus exists for the civilian. By using an ACWS a danger is created, which damages unforeseeable and unintended targets exactly because of its autonomy.[77]

From an objective perspective, this solution is favourable. The victim recognises that something is harming its systems, but whether this is directly controlled by a human or not is unknown to the victim. Even if the system does not have a set framework which substitutes for a kind of ongoing human control, the perpetrator creates a danger of an unforeseeable ACWS acting with less limits in cyber space. It could theoretically target endlessly and indiscriminately, potentially resulting in civilian casualties.

[73] Compare *Supreme Court* (n 10), para. 35.
[74] Roscini (n 2), 206.
[75] Schmitt (n 14), 728; ICRC (n 1), 55.
[76] Compare Ibid., 56, 57; Schmitt (n 14), 732.
[77] Schmitt (n 14), 735-736.

The bad faith of the civilian to use such cyber tools also militates against a narrow interpretation of the time-frame.[78] Besides the necessary tools and abilities, an in-depth knowledge of the functioning of the ACWS, and possibly of the targeted systems and their security barriers, is also required – possible attack points, the logic of the whole network, and so on. The level of knowledge increases with the sophistication of a cyber operation. If an attack causes not only digital damage but also physical damage as a consequence of a manipulation of system controls, detailed knowledge of these processes is needed to prepare such a tool. A civilian using ACWS has to be considered able to foresee and understand the damage they might cause, and thus the consequences of their actions. For this reason, there is no legal need for a regaining of protection by the law, except where there are contravening acts of withdrawal by the civilian, like providing information about the operating ACWS.

'[G]rey areas should be interpreted in favour of finding direct participation' to enable 'a clear distinction between civilians and combatants'.[79] It would also be inequitable to restrict the range of time concerning a civilian using ACWS. The rule of equity is known especially in the Anglo-American legal system and requires that 'those who seek equity shall come with clean hands'.[80] The civilian makes use of the protection of humanitarian law but nevertheless acts to its detriment by getting involved in hostilities. Therefore, they cannot seek equitable treatment, meaning they cannot be protected by international law, while at the same time violating it by the same acts.

An extension of the time-frame would also provide for legal equality with the perpetrator.[81] Otherwise the civilian would be privileged by law due to using ACWS or delayed or conditioned attack tools. The use of delayed attacks would legally be protected, even if this conduct violates the principle of distinction in its negative understanding (whom not to consider as a protected civilian).

Finally, the requirement that the constitutive criteria have to prevail for implying a direct participation of a civilian restrict an inadmissible or unlawful extension of DPH.

5. CONCLUSION

The proposed solution for an extension of the DPH rule applying to a civilian using ACWS is practical. It restores the balance between the protection of civilians and military necessity by preventing legal privileging of wilful perpetrators. Nevertheless, a lawful attack on a participating civilian has a high prerequisite: an attacking party has to be able to attribute and identify the individual[82] with a high level of certainty and proof[83] as the life of the perpetrator is at risk due to the consequences. This and the determination of the DPH criteria have to

[78] Compare Boothby (n 49), 759-760.

[79] Michael Schmitt 'Direct Participation in Hostilities and 21st Century Armed Conflict' in H. Fischer (ed.) *Crisis Management and Humanitarian Protection: Festschrift für Dieter Fleck* (2004), 505, 509.

[80] *Diversion of Water from the Meuse, Netherlands v Belgium*, Judgment, PCIJ Series A/B No 70, ICGJ 321 (PCIJ 1937) [Individual Opinion of M Hudson] 77; *Military and Paramilitary Activities in and against Nicaragua, Nicaragua v. US of America*, (Dissenting Opinion of Judge Schwebel) [1986] ICJ Rep 259 [268]; compare also the Latin legal quote 'nemo auditur propriam turpitudinem allegans'.

[81] Compare: Boothby (n 49), 757.

[82] Wolff Heintschel von Heinegg, 'Cyberspace- Ein völkerrechtliches Niemandsland' in: Schmidt-Radefeldt/ Meissler (Hrsg.), *Automatisierung und Digitalisierung des Krieges*, (Nomos 2012) 159, 172.

[83] *Supreme Court* (n 10), para. 40; Roscini, *Evidentiary Issues* (n 53), 254; Schmitt (n 14), 736.

be conducted with 'all feasible precautions', according to Article 57(2)(a)(i) AP I. If there is doubt, the person has to be treated and protected like a civilian.[84] In reality, a possible practical consequence could be a targeted cyber counterstrike on a civilian computer or computer network instead of a targeted lethal strike on an identified civilian. All in all, the problem of cyber operations by civilians will increase in the future and more sophisticated and effective cyber means will be available. The suggested solution should be recognised as one answer to react to these significant threats, especially those posed by autonomous abilities of cyber tools.

REFERENCES

Dorota Banaszewska, 'Kombattanten und Zivilisten weit weg vom Schlachtfeld' in Robert Frau (ed.), *Drohnen und das Recht* (Mohr Siebeck 2014).

Bill Boothby, ' "And For Such Time As": The Time Dimension to Direct Participation in Hostilities', [2010] *International Law and Politics* 42, 741-768.

Russell Buchan, 'Cyber Warfare and the Status of Anonymous under International Humanitarian Law', [2016] *Chinese JIL*, 741-772.

Heather Harrison-Dinniss, 'Participants in Conflict – Cyber Warriors, Patriotic Hackers and the Laws of War' in Dan Saxon (ed.), *International Humanitarian Law and the Changing Technology of War* (Martinus Nijhoff Publishers 2013) 251-278.

Jean De Preux, 'Protocol I – Article 43' in: Yves Sandoz, Christophe Swinarski, Bruno Zimmermann (eds.), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Martinus Nijhoff Publishers 1987).

Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (2nd edn., Cambridge University Press 2010).

Wolff Heintschel von Heinegg, 'Cyberspace – Ein völkerrechtliches Niemandsland' in Schmidt-Radefeldt/ Meissler (eds.), *Automatisierung und Digitalisierung des Krieges* (Nomos 2012) 159-174.

Jean-Marie Henckaerts, Louise Doswald-Beck, *Customary International Humanitarian Law*, Vol I, (Cambridge University Press 2005).

ICRC, *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law*, Nils Melzer (ed.) (ICRC 2009).

William A. Owens, Kenneth W. Dam, and Herbert S. Lin, *Technology, Policy, Law and Ethics* (1st edn., National Academies Press 2009).

Program on Humanitarian Policy and Conflict Research at Harvard University, *Commentary on the HPCR Manual on the International Law Applicable to Air and Missile Warfare* (2010), available at www.ihlresearch.org/amw.

Marco Roscini, *Cyber Operations and the Use of Force in International Law* (1st edn. OUP 2014).

Marco Roscini, 'Evidentiary Issues in International Disputes Related to state Responsibility for Cyber Operations', (2015) *Texas International Law Journal*, Vol 50, Symposium Issue 2, 233-273.

Michael N. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyberwarfare* (Cambridge University Press 2013).

[84] Schmitt (n 14), 736.

Michael N. Schmitt, ‘Deconstructing Direct Participation in Hostilities: The Constitutive Elements’, (2010) *International Law and Politics*, Vol. 42, 697-739.

Michael N. Schmitt, ‘Direct Participation in Hostilities and 21st Century Armed Conflict’ in H. Fischer (ed.) *Crisis Management and Humanitarian Protection: Festschrift für Dieter Fleck* (Berliner Wissenschafts-Verlag 2004), 505-529.

Wissenschaftliche Dienste des Bundestags, *Anwendbarkeit des humanitären Völkerrechts auf Computernetzwerkoperationen und digitale Kriegsführung* (Cyber Warfare), (2015) WD 2 – 3000 – 038/15.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

From the Vanishing Point Back to the Core: The Impact of the Development of the Cyber Law of War on General International Law

Kubo Mačák
Law School
University of Exeter
Exeter, United Kingdom
k.macak@exeter.ac.uk

Abstract: The law of war was famously described by Sir Hersch Lauterpacht as being 'at the vanishing point of international law'. However, in a historical twist, international legal scrutiny of cyber operations emerged and developed precisely through the optics of the law of war. This paper analyses the influence that the development of the cyber law of war has had and might yet have on the 'core' of international law, in other words, on general international law. It analyses three key dimensions of the relationship between the law of war and general international law: systemic, conceptual, and teleological. It argues that, firstly, a systemic-level shift has taken place in the discourse, resulting in the academic debate and state focus moving from law-of-war questions to questions of general international law including sovereignty, non-intervention, and state responsibility. A better understanding of this trend should allay the fears of fragmentation of international law and inform the debate about the relationship between the law of war and 'core' international law. Secondly, this development has created fertile grounds for certain concepts to migrate from the law of war, where they had emerged, developed or consolidated, into general international law. A case in point is the functionality test, which originated as a compromise solution to determine whether a cyber operation amounts to an 'attack' under the law of war, but which may offer additional utility in other areas of international law including the law of state sovereignty and the law of arms control and disarmament. Thirdly, however, it is imperative that the unique teleological underpinning of the law of war is taken into consideration before introducing its rules and principles to different normative contexts. Paradoxically, a blanket transplantation of these norms might in practice jeopardise the underlying humanitarian considerations.

Keywords: *cyber attacks, cyber security, functionality test, general international law, law of war*

1. INTRODUCTION

Described in the press as 'the year of the hack',[1] 2016 was anything but short on major cyber security incidents. The technology company Yahoo! revealed that more than one billion of its user accounts had been compromised.[2] Three US intelligence agencies suggested in a joint report that the current Russian leadership had ordered specific cyber operations with the intent to interfere with the 2016 presidential election in favour of Donald Trump.[3] Dwarfing all other incidents in terms of its immediate impact, a DDoS attack against the internet infrastructure provider Dyn made dozens of major internet platforms and services inaccessible across the world.[4]

Although the prominence of these attacks can hardly be disputed, there have been no serious claims that any of them should be viewed as an act of war or perceived through the lens of the regulation of warfare on the international plane. It appears that the 'military paradigm' is now firmly on the decline when it comes to analysing malicious cyber operations from the perspective of international law. The bold prediction from a few years ago that 'Cyber War Will Not Take Place' in an eponymous article by Thomas Rid[5] seems to have held water. But does that mean the law of war has nothing to offer to general international law with respect to the regulation of cyber operations?[6]

[1] Geof Wheelwright, 'How 2016 Became the Year of the Hack – and What it Means for the Future' *The Guardian* (21 December 2016) <https://www.theguardian.com/technology/2016/dec/21/how-2016-became-the-year-of-the-hack-and-what-it-means-for-the-future>.

[2] 'Important Security Information for Yahoo Users' *Business Wire* (14 December 2016) <http://www.businesswire.com/news/home/20161214006239/en/Important-Security-Information-Yahoo-Users>.

[3] United States (US), Office of the Director of National Intelligence, 'Background to "Assessing Russian Activities and Intentions in Recent US Elections": The Analytic Process and Cyber Incident Attribution' (6 January 2017) <https://www.dni.gov/files/documents/ICA_2017_01.pdf>.

[4] Nicky Woolf, 'DDoS attack that disrupted internet was largest of its kind in history, experts say' *The Guardian* (26 October 2016) <https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet>.

[5] Thomas Rid, 'Cyber War Will Not Take Place' (2012) 35 Journal of Strategic Studies 5. An extended version of the argument was published as Thomas Rid, *Cyber War Will Not Take Place* (OUP 2013).

[6] For the purposes of this paper, the term 'law of war' is understood as 'that part of international law that regulates the resort to armed force; the conduct of hostilities and the protection of war victims in both international and non-international armed conflict; belligerent occupation; and the relationships between belligerent, neutral, and non-belligerent States'. US Department of Defense, *Law of War Manual* (2015). The term 'general international law' has been authoritatively defined as 'that which is binding upon a great many states. General international law, such as provisions of certain treaties which are widely, but not universally, binding and which establish rules appropriate for universal application, has a tendency to become universal international law'. Robert Jennings and Arthur Watts, *Oppenheim's International Law* (9th edn., OUP 2008) 4.

This paper challenges the assumption belying that rhetorical question. A point of historical irony ought to be highlighted at the outset. To borrow from Sir Hersch Lauterpacht's famous statement, the law of war has long been seen as confined to 'the vanishing point of international law'.[7] Yet, in an unexpected twist, international legal scrutiny of cyber operations emerged and developed precisely through the optics of the law of war. And that is the background against which the paper argues for particular effects which the law at the 'vanishing point' has had, may yet have, but also ought not to have on the 'core' of international law.

The paper commences by analysing the development of the cyber law of war and the corresponding gradual systemic shift of the discourse towards areas of international law closer to its core (Section 2). It then examines the conceptual dimension of the relationship between the law of war and general international law by focussing on the 'functionality test' as a concept arising in the former area and on its potential and actual impact on the latter (Section 3). Finally, it warns that the unique teleological underpinning of the law of war must be taken into consideration before transplanting its rules and principles to different normative contexts more generally (Section 4).

2. SYSTEMIC DIMENSION

Traditionally, international law maintained a strict division between war and peace. Since Ciceronian times,[8] it has been said that *inter bellum et pacem nihil est medium*: there was no intermediate state between war and peace.[9] This meant that international law was in fact a composite of two disparate bodies of rules. There was one set of norms for the peacetime (the law of peace) and another one that applied in times of war (the law of war). The orthodox view was that there was no status mixtus under international law: each situation was either one of peace or war, and the corresponding body of law would apply to it exclusively.[10]

That is no longer true today.[11] For a number of reasons – which include the existence of a multitude of actors on the international plane, the complexity of relations in the globalised world, as well as the asymmetrical nature of most contemporary armed conflicts – norms that used to be clumped together as 'peacetime law' now do not cease to apply with the outbreak of hostilities. This has been recognised expressly by the International Court of Justice (ICJ) with respect to a central area of the law of peace, namely international human rights law.[12] In a situation of armed conflict, the two originally separate bodies of law now apply in parallel,[13]

[7] Hersch Lauterpacht, 'The Problem of the Revision of the Law of War' (1952) 29 British Year Book of International Law 360, 382.

[8] Cicero, *Philippics 3–9* (Gesine Manuwald tr, Walter de Gruyter 2007) vol I, 260.

[9] Hugo Grotius, *The Law of War and Peace in Three Books* (Francis W Kelsey tr, 1625) book III, ch XXI, s I, §1.

[10] Robert Joseph Phillimore, *International Law* (Butterworths 1879) vol III, 794.

[11] Aoife O'Donoghue, 'Splendid Isolation: International Humanitarian Law, Legal Theory and the International Legal Order' (2011) 14 *Yearbook of International Humanitarian Law* 107, 114; Jann K Kleffner, 'Scope of Application of International Humanitarian Law' in Dieter Fleck (ed.) *The Handbook of International Humanitarian Law* (3rd edn., OUP 2013) 70.

[12] ICJ, Legality of the Threat or Use of Nuclear Weapons Case (Advisory Opinion) [1996] ICJ Rep 226 [24]–[25] ('Nuclear Weapons Advisory Opinion'); ICJ, Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory (Advisory Opinion) [2004] ICJ Rep 136 [106].

[13] Cf. UN, Human Rights Committee, General Comment No. 31, UN Doc CCPR/C/21/Rev.1/Add. 13 (26 May 2004) ('both spheres of law are complementary, not mutually exclusive').

and any norm conflict is resolved (or, as the case may be, avoided) by the applicable maxims of interpretation (including the central principle of *lex specialis derogat legi generali*).[14]

However, not all traces of the historical distinction have disappeared. It is true that general international law and its norms concerning state responsibility, treaty interpretation or identification of custom apply equally during armed hostilities and in times of peace. However, norms governing the use of armed force have maintained their conceptual separation from the rest of international law. These norms comprise mainly the *jus ad bellum* and the *jus in bello*: the former sets out conditions under which states may resort to war, the latter provides rules which the belligerents must abide by once they are engaged in an armed conflict.[15] Crucially, neither of these bodies of law applies to situations not characterised by any use of armed force. Therefore, the radically different optics of rules applicable during peacetime and those applicable to uses of armed force are still very much present in modern international law.

In this sense, international legal scrutiny of cyber operations initially emerged and developed through the optics of the law of war, and not the law of peace. The first analyses came in the form of academic writings in the late 1990s, which considered cyber conflict as a species of armed conflict.[16] The risk of a 'Cyber Pearl Harbor' or a 'Cyber Armageddon', a hypothetical devastating cyber attack against state infrastructure, was envisaged for the first time in the same period.[17] (Today, nearly two decades later, such incidents belong only in the realm of journalists' and novelists' imaginations,[18] and the likelihood of their happening in reality is seen as extremely low.[19]) Against this background, scholars considered what degree of 'computer network attack' would qualify as a use of force under Article 2(4) of the UN Charter.[20]

Early state reactions were similar in their scope and approach. For instance, a prescient 1999 memo by the United States (US) Department of Defense expressly noted that '[t]he law of war is probably the single area of international law in which current legal obligations can be applied with the greatest confidence to information operations'.[21] Another prominent statement by the US, the 2011 *International Strategy for Cyberspace*, went even further and warned that the

[14] See further Matthew Happold, 'International Humanitarian Law and Human Rights Law' in Nigel D White and Christian Henderson (eds.), *Research Handbook on International Conflict and Security Law* (Edward Elgar 2012) 459–463; Marko Milanovic, 'The Lost Origins of *Lex Specialis*: Rethinking the Relationship between Human Rights and International Humanitarian Law' in Jens David Ohlin (ed.), *Theoretical Boundaries of Armed Conflict and Human Rights* (CUP 2016) 103–113.

[15] See, e.g., Christopher Greenwood, 'The Relationship Between *Ius ad Bellum* and *Ius in Bello*' (1983) 9 *Review of International Studies* 221.

[16] See, e.g., Richard W Aldrich, 'The International Legal Implications of Information Warfare' (1996) 10 Airpower Journal 99; Michael N Schmitt, 'Computer Network Attack and the Use of Force in International Law: Thoughts on a Normative Framework' (1999) 37 *Columbia Journal of Transnational Law* 885.

[17] See, e.g., Pierre Thomas, 'Experts Prepare for "An Electronic Pearl Harbor"' CNN (7 November 1997) <https://archive.is/aL98j>.

[18] See, e.g., Mark Russinovich, *Zero Day* (Corsair 2012).

[19] Sean Lawson, 'Does 2016 Mark the End of Cyber Pearl Harbor Hysteria?' *Forbes* (7 December 2016) <http://www.forbes.com/sites/seanlawson/2016/12/07/does-2016-mark-the-end-of-cyber-pearl-harbor-hysteria/>.

[20] See, e.g., Todd A Morth, 'Considering Our Position: Viewing Information Warfare as a Use of Force Prohibited by Article 2(4) of the U.N. Charter' (1998) 30 *Case Western Reserve Journal of International Law* 567; Mark R Jacobson, 'War in the Information Age: International Law, Self-Defense, and the Problem of 'Non-Armed' Attacks' (1998) 21 *Journal of Strategic Studies* 1; Schmitt (n 16).

[21] US, Department of Defense, 'An Assessment of International Legal Issues in Information Operations' (May 1999) 11.

US would respond to hostile acts in cyberspace in line in accordance with its inherent right to self-defence.[22] In the interim, other states either remained silent or endorsed a similar approach.

With hindsight, the 2007 cyber attacks against Estonia stand out as a watershed event for international cyber security law. That year, a decision to remove a Soviet-era sculpture of a Red Army soldier from a central square in the capital city Tallinn[23] led to a concentrated series of cyber operations against a multitude of public and private targets across the country.[24] While the events were unfolding, Estonia even described itself as 'under attack' by Russia.[25] In the aftermath of the incident, NATO established the Cooperative Cyber Defence Centre of Excellence (CCD COE) in Tallinn.[26] Much of the early work of the CCD COE, of which the most prominent was the 2013 *Tallinn Manual on the International Law Applicable to Cyber Warfare*, maintained the law-of-war focus on the regulation of cyberspace.[27]

The *Tallinn Manual* itself dedicated nearly 90% of its rules to the *jus ad bellum* and the *jus in bello*.[28] The remaining ones were reportedly added in the final stages of editing to provide a general context and to underscore the continued application of general international law even in times of (cyber) conflict.[29] Somewhat paradoxically, given the origins of the project, the international group of experts concluded that the attacks against Estonia in 2007 did not reach the threshold of an armed attack or armed conflict and therefore fell outside the scope of the law of war.[30] Nonetheless, the attacks have remained a prime reference point and a trigger for much legal development in the area of cyber security.

To some extent, the predominant law-of-war focus of all of these developments is certainly understandable. Notably, this progression mirrors the development of the Internet, which also started as a military project in the US. It is also undoubtedly related to states' general preoccupation with national security and the high priority that many governments accord to military defence against foreign threats. Last but not least, the military focus has in large part been caused by the fact that a lot of thinking on cyber security within governments was originally done within military and defence circles.[31]

[22] US, The White House, International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World (2011) 9.

[23] Steven Lee Myers, 'Estonia removes Soviet-era war memorial after a night of violence' *The New York Times* (27 April 2007) <http://www.nytimes.com/2007/04/27/world/europe/27iht-estonia.4.5477141.html>.

[24] See further Eneken Tikk, Kadri Kaska, and Liis Vihul, International Cyber Incidents: Legal Considerations (CCD COE 2010) 18–24.

[25] 'Statement by the Foreign Minister Urmas Paet' Eesti Päevaleht (1 May 2007) <http://epl.delfi.ee/news/eesti/statement-by-the-foreign-minister-urmas-paet?id=51085399> ('The European Union is under attack, because Russia is attacking Estonia').

[26] 'NATO Opens New Centre of Excellence on Cyber Defence' *NATO News* (14 May 2008) <http://www.nato.int/docu/update/2008/05-may/e0514a.html>. It should be noted that the concept for a cyber defence centre was submitted by Estonia to NATO already in 2004. It was subsequently approved by the Supreme Allied Commander Transformation in 2006. See NATO CCD COE, 'History' (undated) <https://ccdcoe.org/history.html>.

[27] Michael N Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare* (CUP 2013).

[28] Id., rules 10–95.

[29] Michael J Adams, 'A Warning About Tallinn 2.0 … Whatever It Says' *Lawfare* (4 January 2017) <https://www.lawfareblog.com/warning-about-tallinn-20-%E2%80%A6-whatever-it-says>.

[30] *Tallinn Manual* (n 27) 57–58 and 75.

[31] See further Kenneth Geers, *Strategic Cyber Security* (CCD COE 2011) 19–32.

However, this approach has also led to serious concerns about the downsides of applying the so-called 'military paradigm' to international legal regulation of cyberspace. These concerns have been voiced by several western academics,[32] but – perhaps more importantly – also by non-Western states led by China.[33] According to such critical views, the key risk of the law of war focus on the regulation of cyberspace was that it would 'aggravate the arms race and militarisation in cyberspace'.[34] This fundamental difference in approach to the rule of law in cyberspace between the West and 'the Rest' led by China and Russia has been reflected in commentators' references to two competing 'camps' of states as far as cyber security is concerned.[35]

It would be hasty to equate the question of applicability of the law with the threat of militarisation of cyberspace. As the ICRC stated in a 2015 report to the 32nd International Conference of the Red Cross and Red Crescent:

> [A]sserting that IHL applies to cyber warfare is not an encouragement to militarize cyberspace and should not, in any way, be understood as legitimizing cyber warfare.[36]

This is an obvious truth. Similarly, it would be manifestly wrong to claim that an assertion that Geneva Conventions apply somehow legitimises warfare in general.[37]

Still, the worries about the creeping militarisation of cyberspace cannot easily be brushed away. In the concluding remarks of a 2014 monograph dedicated to a close scrutiny of cyber operations from the perspective of international law, Professor Marco Roscini made the following striking summation: '[t]he militarization of cyberspace is not a risk, it is already a fact'.[38] This state of affairs has some obvious negative consequences. It has been well documented that framing cyber threats as strategic-military concerns contributes to the creation of an 'atmosphere of insecurity and tension in the international system'.[39]

However, a closer look reveals a paradox at the heart of the issue. The originally predominant law-of-war approach based on the viewing of threats to cyber security through the prism of

[32] See, e.g., Mary Ellen O'Connell, 'Cyber Security without Cyber War' (2012) 17 *Journal of Conflict and Security Law* 187; Robin Geiss and Henning Lahmann, 'Freedom and Security in Cyberspace: Shifting the Focus Away from Military Responses Towards Non-Forcible Countermeasures and Collective Threat-Prevention', in Katharina Ziolkowski, *Peacetime Regime for State Activities in Cyberspace: International Law, International Relations and Diplomacy* (NATO CCD COE 2013).

[33] Ma Xinmin, 'Key Issues and Future Development of International Cyberspace Law' (2016) 2 *China Quarterly of International Strategic Studies* 119, 128.

[34] Ibid.

[35] See, e.g., Scott Shackleford and Amanda Craig, 'Beyond the New "Digital Divide": Analyzing the Evolving Role of National Governments in Internet Governance and Enhancing Cybersecurity' (2014) 50 *Stanford Journal of International Law* 119, 135; Kristen Eichensehr, 'The Cyber-Law of Nations' (2015) 103 Georgetown Law Journal 317, 333; Nigel Inkster, *China's Cyber Power* (Routledge 2015) 9.

[36] ICRC, 'International Humanitarian Law and the Challenges of Contemporary Armed Conflicts' (October 2015) <https://www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf > 40.

[37] See Gabor Rona, 'Challenges New Weapons and Humanitarian Assistance Present for International Law' *Just Security* (20 November 2015) <https://www.justsecurity.org/27789/challenges-weapons-humanitarian-assistance-present-ihl/>.

[38] Marco Roscini, *Cyber Operations and the Use of Force in International Law* (OUP 2014) 280.

[39] Myriam Dunn Cavelty, 'The Militarisation of Cyberspace: Why Less May Be Better' in C Czosseck, R Ottis, K Ziolkowski (eds.), *2012 4th International Conference on Cyber Conflict* (NATO CCD COE 2012) 141.

strategic or military discourse does not actually correspond to reality. In fact, most cyber operations either fall below the threshold of armed conflict, or do not occur in the context of an armed conflict, or even if they do, they fail to be attributable to a state or to bring about a sufficiently serious effect.[40] As such, the framework of the law of war is not applicable to the vast majority of the existing cyber operations.

Gradually, key actors have recognised that to resolve this paradox, cyber operations must indeed be analysed outside the 'military paradigm'.[41] This has led to a systemic shift in the discourse, with academic debate and state focus migrating from law-of-war questions to questions of general international law including sovereignty, non-intervention, and state responsibility. Among scholars, this shift was reflected in criticism levied against the method and remit of the *Tallinn Manual* as a perceived leading example of the militarisation trend.[42] To their credit, the international group of experts responded by enlarging the scope of the project to include 'below the threshold' cyber operations.[43] As acknowledged by Professor Michael Schmitt, the project director, 'preoccupation with cyber armed attacks is counter-experiential'.[44] The second edition of the Manual, which has just been published, reflects this shift in the discourse and includes a discussion of state responsibility, the law of the sea, air and space law, and even human rights law.[45]

Similarly, states have moved away from the military paradigm in their recent conduct and official statements. One overarching example is the ongoing work of the United Nations (UN) Group of Governmental Experts (GGE), which has so far resulted in the adoption of three substantive reports.[46] These reports are based on the official submissions of the most cyber-active states and are adopted by consensus of all representatives.[47] To the extent that they have related to international law (as opposed to political norms of conduct in cyberspace), these references have almost exclusively maintained the perspective of peacetime law.[48] A similar trend can be observed at the level of individual states. Even the US, as the supposed principal proponent of the military paradigm, has modified its approach. For example, in late 2016, Brian Egan, the US State Department Legal Adviser, delivered a landmark speech on 'International

[40] See, e.g., Tikk, Kaska, and Vihul (n 24) 82–83 ('it is highly problematic to apply Law of Armed Conflict to the Georgian cyber attacks – the objective evidence of the case is too vague to meet the necessary criteria of both state involvement and gravity of effect').

[41] See, e.g., Oona A Hathaway et al., 'The Law of Cyber-Attack' (2012) 100 *California Law Review* 817, 840; Geiss and Lahmann (n 32) 657; Michael N Schmitt, '"Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law' (2014) 54 *Virginia Journal of International Law* 697, 698.

[42] See, e.g., Dieter Fleck, 'Searching for International Rules Applicable to Cyber Warfare: A Critical First Assessment of the New *Tallinn Manual*' (2013) 18(2) *Journal of Conflict and Security Law* 331, 332–335; Kristen Eichensehr, 'Review of The *Tallinn Manual* on the International Law Applicable to Cyber Warfare (Michael N. Schmitt ed., 2013)' (2014) 108 *American Journal of International Law* 585, 589; Ma Xinmin, 'Letter to the Editors: What Kind of Internet Order Do We Need?' (2015) 14 *Chinese Journal of International Law* 399, 402.

[43] Jill Dougherty, 'NATO Cyberwar Challenge: Establish Rules of Engagement' CNN (7 November 2016) <http://edition.cnn.com/2016/11/07/politics/nato-cyber-centre-international-law/>.

[44] Schmitt (n 41) 698.

[45] Michael N Schmitt (ed.), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd ed., CUP 2017).

[46] UN Doc A/65/201 (2010); UN Doc A/68/98 (2013); UN Doc A/70/174 (2015).

[47] See further UN, 'Developments in the Field of Information and Telecommunications in the Context of International Security' (undated) <https://www.un.org/disarmament/topics/informationsecurity/>.

[48] But see UN Doc A/65/201 (2010) para 7; UN Doc A/70/174 (2015) para 28(d).

Law and Stability in Cyberspace'.[49] Although the speech did contain a brief section on cyber operations in the context of armed conflict,[50] the vast majority of the text was devoted to peacetime aspects of cyberspace regulation.[51]

Therefore, an examination of the systemic dimension of the relationship between the law of war and general international law reveals a clear shift of focus from the former to the latter, as far as the international regulation of cyber operations is concerned. This has had two main consequences. First, the transition of the debate from one specific area of international law – the law of war – to general international law demonstrates that concerns about a supposed fragmentation of international law may in fact be less salient than some have worried.[52] Secondly, this systemic shift has allowed for concepts and approaches to migrate from the law of war to general international law. The next two sections explore the opportunities and limitations posed by this particular migratory pattern.

3. CONCEPTUAL DIMENSION

The conceptual level of the relationship between the law of war and general international law reveals the potential for certain ideas, concepts or approaches to migrate from the former into the latter. Given the limited scope of the paper, this section focusses on one particular issue, namely the conceptualisation of the non-physical effects of cyber operations in international law.[53] Unsurprisingly, this question acquires central importance in relation to cyber security. Unlike ordinary conduct in the physical world, cyber operations result in effects that are normally invisible to the naked eye. At what point do they then become relevant from the perspective of the law?

In the law of war, this question arises in the context of the regulation of targeting. This area of the law is based on the principle of distinction, codified in Article 48 of Additional Protocol I (AP I), described by the ICJ as one of the 'cardinal principles' of the law of war,[54] and generally considered to reflect customary law.[55] This provision mandates that belligerents must at all times distinguish 'between civilian objects and military objectives and accordingly [must] direct their operations only against military objectives'.[56]

[49] Brian J Egan, 'International Law and Stability in Cyberspace' (10 November 2016) <www.justsecurity.org/wp-content/uploads/2016/11/Brian-J.-Egan-International-Law-and-Stability-in-Cyberspace-Berkeley-Nov-2016.pdf>.

[50] Id., 8–10.

[51] Id., 1–8 and 11–26.

[52] See Martti Koskenniemi and Päivi Leino, 'Fragmentation of International Law? Postmodern Anxieties' (2002) 15 *Leiden Journal of International Law* 553.

[53] Other conceptual questions at the intersection between the law of war and general international law include the geographical scope of applicability of the law in relation to cyberspace; the requirement of organisation in online groups; or the problem of calculating proportionality *largo sensu* in the cyber context.

[54] Nuclear Weapons Advisory Opinion (n 12) [78].

[55] Id. [79]; Eritrea-Ethiopia Claims Commission, Partial Award, Western Front, Aerial Bombardment and Related Claims, Eritrea's Claims 1, 3, 5, 9–13, 14, 21, 25 & 26 (2005) 26 RIAA 291, 327; Jean-Marie Henckaerts and Louise Doswald-Beck (eds.), *Customary International Humanitarian Law* (CUP 2005) 3.

[56] Article 48 AP I.

There is some discussion as to the precise meaning of the term 'operations' (or, in full, 'military operations'[57]) in the cited provision for the purposes of the law of war. Some believe that it is practically synonymous with the notion of 'attack' used in almost all specific rules on targeting in the same section of the Protocol.[58] Others consider 'attacks' to be a subset of 'military operations'.[59] According to this latter view, activities such as moving armed forces, gathering military intelligence or providing logistical support qualify as military operations, but not as 'attacks' *stricto sensu*. Nevertheless, for our present purposes, it will suffice to focus on the concept of 'attack' as it is the central threshold notion of the law of targeting.

Accordingly, it is by reference to that notion that we determine whether the relevant targeting rules apply to a particular combat action against the enemy. If the conduct in question is an 'attack', then it must conform to these rules, which include the prohibition of attacks against civilians and civilian objects,[60] the prohibition of indiscriminate attacks,[61] and the prohibition of attacks causing disproportionate 'collateral' damage to civilians or civilian property.[62] This conclusion, however, invites the question how to determine whether a given cyber operation qualifies as an 'attack'.

Although the first Additional Protocol provides a widely accepted definition of that term, its precise application to cyber operations is controversial. According to Article 49(1) AP I, '"[a]ttacks" means acts of violence against the adversary, whether in offence or in defence'. Despite the reference to violence, '[t]he use of physical force is not a *sine qua non* of an attack under the terms of the Protocol'.[63] Yet the spectrum of views concerning which cyber operations would qualify accordingly is very broad.

At the one end of the spectrum, the requirement of violence is interpreted as meaning that 'attack' only includes those operations which result in injury or death to individuals or damage or destruction of physical objects.[64] On this view, an operation aimed at temporarily disabling a device or a network – for instance, by shutting down an electrical distribution grid[65] – would not qualify as an 'attack'.[66] At the other extreme, it was argued that cyber operations with reversible effects should be seen as 'neutralising' an object,[67] and as such they would qualify as

[57] See Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds.) Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949 (ICRC 1987) 600.

[58] See, e.g., Michael N Schmitt, 'Wired Warfare: Computer Network Attack and *Jus in Bello*' (2002) 84 *International Review of the Red Cross* 365, 376; David Turns, 'Cyber War and the Concept of 'Attack' in International Humanitarian Law' in Dan Saxon (ed.) *International Humanitarian Law and the Changing Technology of War* (Brill 2013) 217; Roscini (n 38) 178.

[59] See, e.g., Michael Bothe, Karl Josef Partsch and Waldemar A Solf (eds.), *New Rules for Victims of Armed Conflicts: Commentary on the two 1977 Protocols Additional to the Geneva Conventions of 1949* (Martinus Nijhoff 1982) 408; UK Ministry of Defence, *The Manual of the Law of Armed Conflict* (OUP 2004) 81 fn 187; Heather Harrison Dinniss, *Cyber Warfare and the Laws of War* (CUP 2012) 199.

[60] Article 51(2) AP I.

[61] Article 51(4) AP I.

[62] Article 51(5)(b) AP I.

[63] Kubo Mačák, 'Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law' (2015) 48 *Israel Law Review* 55, 76.

[64] Schmitt (n 58) 374.

[65] Knut Dörmann, 'Applicability of the Additional Protocols to Computer Network Attacks' <https://www.icrc.org/eng/assets/files/other/applicabilityofihltocna.pdf> 4.

[66] Schmitt (n 58) 374.

[67] The 'neutralisation' of an object is one of the modalities of enemy engagement foreseen by Article 52(2) AP I and therefore, the argument goes, a form of 'attack'.

‘attacks’.[68] Consequently, an operation against the electric grid mentioned above would be an ‘attack’ even if its aim was merely to disable it, and not to cause its destruction.

Professor Schmitt documented how the discussions in the context of the *Tallinn Manual* project resulted in the adoption of a new, compromise position between the two extremes in the form of a ‘functionality test’.[69] As expressed in both editions of the *Manual*, this test mandates that ‘interference by cyber means with the functionality of an object constitutes damage or destruction’[70] and as such it amounts to an ‘attack’.[71] Although there were some shades of difference among the experts as to the precise meaning of interference,[72] the majority of them agreed that the criterion would be met if, as a result of a cyber operation, ‘the object in question is unusable for its intended purpose, at least until some form of repair is undertaken’.[73] Hence, the operation against the electric grid would qualify as an attack if it either made the grid permanently inoperable or necessitated some degree of repair.[74]

Crucially, this functionality-oriented approach towards the definition of a cyber attack can have impact on other areas of international law closer to its core. Two specific observations can be made in this regard. Firstly, the functionality test directly influences what is to be considered as a ‘weapon’, which is a term that exceeds the boundaries of the law of war. This is because ‘cyber capabilities that are used, designed, or intended to be used’ for the purposes of ‘attacks’ in the sense of Article 49(1) AP I, are ‘cyber weapons to which the law of weaponry applies’.[75] Accordingly, a cyber tool that may cause loss of functionality of an object is subject to the prescriptions of weapons law, a body of law that includes, in addition to the law of war rules, the law of arms control and disarmament.[76] As it is well established that states must refrain from using ‘weapons’ against civil aircraft in flight,[77] the functionality test may also have further influence even on international civil aviation law.[78] It should be noted that prominent official as well as scholarly definitions of ‘cyber weapons’ align with this functionality-oriented interpretation.[79]

Secondly, the functionality test has proven capable of application in other areas of international law. The discussions in the Tallinn 2.0 process have led to an agreement among the international

[68] ICRC, ‘International Humanitarian Law and the Challenges of Contemporary Armed Conflicts’ (October 2011) <https://app.icrc.org/e-briefing/new-tech-modern-battlefield/media/documents/4-international-humanitarian-law-and-the-challenges-of-contemporary-armed-conflicts.pdf> 37.

[69] Michael N Schmitt, ‘Rewired Warfare: Rethinking the Law of Cyber Attack’ (2014) 96 *International Review of the Red Cross* 189, 198–201.

[70] *Tallinn Manual* (n 27) 108; *Tallinn Manual 2.0* (n 45) 417.

[71] *Tallinn Manual* (n 27) 108–109; *Tallinn Manual 2.0* (n 45) 417.

[72] See *Tallinn Manual* (n 27) 108–109 paras 10 and 11; see further *Tallinn Manual 2.0* (n 45) 417–418 paras 10–12.

[73] Schmitt (n 69) 199.

[74] *Tallinn Manual* (n 27) 108–109; *Tallinn Manual 2.0* (n 45) 417.

[75] William H Boothby, *Weapons and the Law of Armed Conflict* (2nd edn., OUP 2016) 238; see also *Tallinn Manual* (n 27) 141–142; *Tallinn Manual 2.0* (n 45) 452–453.

[76] Boothby (n 75) 2–3.

[77] Convention on International Civil Aviation (signed at Chicago on 7 December 1944) 15 UNTS 295 (‘Chicago Convention’), as amended by the Protocol Relating to an Amendment to the Convention on International Civil Aviation (signed at Montreal on 10 May 1984), Article 3*bis*(a) (‘every State must refrain from resorting to the use of weapons against civil aircraft in flight’).

[78] See further *Tallinn Manual 2.0* (n 45) 268–269.

[79] See, e.g., US, Department of the Air Force Instruction 51-402, Legal Reviews of Weapons and Cyber Capabilities (27 July 2011) 6, incorporated by reference in *US Law of War Manual* (n 6) 999 fn 75; Thomas Rid and Peter McBurney, ‘Cyber Weapons’ (2012) 157 *RUSI Journal* 6, 7.

group of experts that state-sponsored cyber operations that result in the loss of functionality of cyber infrastructure on the territory of another state without its consent may amount to a violation of that state's sovereignty.[80] It also reflects a broader development in the international community and states' general approach to the regulation of cyberspace. For instance, in a submission to the UN GGE, Panama expressly described operations restricted to cyberspace as a 'new form of violence'.[81] More recently, the US State Department Legal Advisor Brian Egan expressly stated that 'one state's non-consensual cyber operation in another state's territory could violate international law, even if it falls below the threshold of a use of force'.[82] The functionality test may provide the much-needed granularity to such generalised and open-ended official proclamations.

4. TELEOLOGICAL DIMENSION

The foregoing discussion might suggest that it is tempting to search for as many notions and approaches at the interface of the law of war and general international law as possible. Admittedly, a legal standard like the functionality test may bring additional granularity to currently less developed areas of international cyber security law. However, we should be wary of transplanting law-of-war notions and approaches without closer scrutiny. Particularly, we must not lose sight of the unique teleological underpinning of the law of war, which sets it apart from other disciplines of international law.

The teleology of the law of war is 'predicated on a subtle equilibrium between the two diametrically opposed stimulants of *military necessity* and *humanitarian considerations*'.[83] This means that most of its rules are based on an inbuilt balance between these two keystone values. For instance, the principle of proportionality in targeting permits attacks against lawful targets (thus allowing conduct that is militarily necessary), but only up to that abstract point at which the expected collateral damage of the attack outweighs the anticipated military advantage (thus disallowing such attacks for humanitarian reasons).[84] It is because of this equilibrium unique to the law of war that notions and approaches developed in its context cannot be automatically transplanted to domains where one or both of the stimulants are absent.

This is an important warning, as calls for the wholesale adoption of law-of-war approaches to decision-making about cyber defence generally have already started appearing in the literature.[85] To some extent, this development is understandable because the law of war may seemingly offer a ready-made detailed legal framework which has been extensively mulled over in relation to cyber operations. The key implication of such proposals is that the permissibility of a specific cyber operation would be determined by a *mutatis mutandis* application of the law-of-war principles of targeting, including distinction, precautions in attack, and proportionality.

[80] *Tallinn Manual 2.0* (n 45) 20–21.
[81] UN Doc A/57/166/Add.1 (29 August 2002) 5.
[82] Egan (n 49) 13.
[83] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (3rd edn., CUP 2016) 9 (emphases added).
[84] Articles 51(5)(b), 57(2)(a)(iii), and 57(2)(b) AP I.
[85] See, e.g., Corey T Holzer and James E Lerums, 'The Ethics of Hacking Back' (2016) <https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2016-01.pdf>.

However, it is submitted that if a cyber operation takes place outside of the context of an armed conflict or even within an armed conflict but does not reach the threshold of an 'attack' under the law of war, it would be misguided to subject such an operation to the strict application of the targeting principles. This argument applies regardless of the specific conceptualisation of the term 'attack'. Even under the loose conception, according to which reversible cyber operations seeking to disable objects qualify as attacks,[86] some cyber operations still manifestly fall outside of the scope of the law of war. This would be the case, for instance, with respect to the dissemination of propaganda by cyber means or to operations tantamount to economic sanctions.[87]

In such circumstances, to insist on the applicability of the law-of-war principles of targeting would be misplaced and could in fact result in a reduction of humanitarian protection. By way of illustration, let us consider the example of interfering with a civilian television broadcast during an armed conflict. There is a general consensus that an operation consisting solely of temporarily blocking a television broadcast would not amount to an 'attack' under the law of war.[88] Yet if principles of targeting were to be applied to a cyber operation aiming to disrupt the transmission of television signal, the operation would highly likely fail to live up to the principle of distinction. This is because its target, specifically the communications system supporting the television broadcast, would normally be civilian in nature and as such it would not meet the criteria of a military objective.[89]

However, the principle of distinction along with other targeting principles flows from the equilibrium between the two opposing considerations of military necessity and humanity. Once these underlying 'driving forces'[90] are taken out of the equation, the reliance on the principles may in fact result in overall detriment. This is apparent when we contrast two recent examples of targeting TV stations, one kinetic, the other one digital.

The 'kinetic' example is the 1999 NATO bombing of the Serbian state-owned TV station in Belgrade, which took the station off the air for about six hours[91] and resulted in the death of sixteen civilians.[92] Kinetic bombing is manifestly an attack under the framework of the law of war. Hence, NATO sought to justify the bombardment by describing the TV station as part of the enemy's Command, Control and Communications (C3) military network as well as of its propaganda machinery used to control the population—and as such a legitimate military objective.[93] In a later report, a committee at the International Criminal Tribunal for the former Yugoslavia (ICTY), expressed some doubts about this justification,[94] but nonetheless

[86] See text corresponding to notes 67–68 above.

[87] See, e.g., Cordula Droege, 'Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians' (2012) 94 *International Review of the Red Cross* 533, 559–560.

[88] Schmitt (n 69) 203 fn 63; Droege (n 87) 560; *Tallinn Manual 2.0* (n 45) 418.

[89] Article 52(2) AP I.

[90] Dinstein (n 83) 8.

[91] 'Bombed Serb TV back on air' *BBC News* (23 April 1999) <http://news.bbc.co.uk/1/hi/world/europe/326339.stm>.

[92] 'NATO challenged over Belgrade bombing' *BBC News* (24 October 2001) <http://news.bbc.co.uk/1/hi/world/europe/1616461.stm>.

[93] NATO, Press Conference by NATO Spokesman, Jamie Shea and Colonel Konrad Freytag, SHAPE (23 April 1999) <http://www.nato.int/kosovo/press/p990423l.htm>.

[94] ICTY, Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia (14 June 2000) <http://www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal> [75]–[76].

recommended against commencing an investigation related to the incident.[95] Be that as it may, from a humanitarian perspective, the civilian deaths pose a grave collateral effect of the operation that cannot be ignored.

In contrast, one may consider the 'digital' example of the 2015 cyber attack against the French network TV5 Monde. (For the purposes of this paper, we put aside the vexing question of attribution of the operation and focus solely on its impact in order to contrast it with the kinetic attack described above.[96]) Essentially, the only direct effect of the operation was to block broadcasting by the TV5 Monde network on 12 channels for approximately 10 hours.[97] If the operation had taken place in the context of an armed conflict, it would not have risen to the level of an 'attack'. However, all targeted systems were civilian in nature; therefore, the operation would by definition fall short of the principle of distinction. Although an operation of this kind does not lead to any civilian casualties, to subject it to the law-of-war targeting principles would thus render it unlawful.

This results in a paradox. A state aiming to disrupt the television signal of a station on enemy territory may be faced with a difficult dilemma. It may either engage in a kinetic attack, which might be justified along the same lines of reasoning advanced by NATO in the example above, but which will almost certainly result in civilian deaths. Alternatively, it may choose to disable the TV broadcast by way of a cyber operation, but the application of targeting principles might mean that such conduct would become internationally unlawful. In this situation, the state may well choose the 'legally safe' option, even if it would entail more human suffering.

The same logic applies to cyber operations that take place outside of any armed conflict. It may well be that an operation against a TV broadcaster in peacetime amounts to a violation of sovereignty of the territorial state, to a prohibited harmful interference, or even to a prohibited form of intervention. But this does not follow from the civilian status (*vel non*) of the target in question. The lawfulness of such a cyber operation must be determined by the application of the relevant rules,[98] but it would be misplaced to draw these directly from principles that have evolved in a different normative context and which reflect the specific aims of the law of war.

5. CONCLUSION

The 'year of the hack' confirmed the ubiquity of cyber security threats posed by malicious actors in cyberspace. In the meantime, states are slowly accepting the need to articulate international regulatory responses.[99] Against this backdrop, three overarching conclusions may be drawn from the preceding analysis. Firstly, a systemic shift has taken place, moving the regulatory focus from the law of war to general international law. A better understanding of this trend should alleviate some of the fears of fragmentation of international law and inform the debate about the relationship between the law at the vanishing point and at the core of international

[95] Id. [79].

[96] On that question, see further Graham Cluley, 'TV5Monde Attack Proves Hacking Attribution Is Very Difficult' (10 June 2015) <https://www.grahamcluley.com/tv5monde-attack-hacking-attribution>.

[97] 'How France's TV5 was almost destroyed by "Russian hackers"' *BBC News* (10 October 2016) <http://www.bbc.co.uk/news/technology-37590375>.

[98] See *Tallinn Manual 2.0* (n 45) 17–27 (violation of sovereignty), 294–298 (prohibition of harmful interference), 312–327 (prohibition of intervention).

[99] See, e.g., Egan (n 49) 7.

law. Secondly, this trend has allowed for specific concepts to migrate from the law of war where they had originated, evolved or consolidated and to influence other areas of international law. An illustrative example is the functionality test, which offers significant utility for the law of state sovereignty as well as the law of arms control and disarmament. Thirdly, however, it is imperative that the unique teleological underpinning of the law of war is taken into account before introducing its rules and principles to different normative contexts. Paradoxically, a blanket transplantation of these norms might in practice jeopardise the underlying humanitarian considerations.

ACKNOWLEDGEMENTS

I am thankful to Rogier Bartels, Ana Beduschi, Charles Garraway, Michael N Schmitt, and the anonymous reviewers for their valuable comments and suggestions on earlier drafts of this paper. Any remaining errors or omissions are my own responsibility.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Control and Capabilities Test: Toward a New *Lex Specialis* Governing State Responsibility for Third Party Cyber Incidents*

Peter Z. Stockburger
Senior Managing Associate
Dentons
San Diego, California, US
peter.stockburger@dentons.com

Abstract: It is well accepted under international law that a State is generally responsible for the internationally wrongful acts of its *de jure* and *de facto* State organs. It is equally well accepted that a State is generally responsible for the internationally wrongful acts of non-State actors who are neither *de jure* nor *de facto* State organs if the State sufficiently directs and controls each element of the internationally wrongful act committed by the non-State actor. This general rule, known as the "effective control" test, is recognized as the *lex generalis* governing imputed State responsibility for the unlawful actions of non-State actors. As the *lex generalis*, this principle does not vary with the nature of the wrongful act in question unless there is a clearly expressed *lex specialis*. Based on a review of State practice since 2014, there is, in fact, a *lex specialis* forming that would allow for imputed State responsibility for the internationally wrongful cyber operations of non-State actors even in the absence of evidence demonstrating "effective control." Specifically, a review of State practice since 2014 reveals that States have attributed the unlawful cyber operations of non-State actors to States, publicly, even in the absence of evidence demonstrating clear State direction and control. States have instead applied what this paper calls the "control and capabilities" test, examining a multitude of factors to determine State responsibility, including: (1) the relationship between the non-State actor and the State, if any; (2) any apparent influence the State exercises over the non-State actor; (3) the methods used by the non-State actor; (4) the motivations of the two parties, if known; (5) whether the two parties use similar code; (6) technical capabilities; and (7) geographic location. This new attribution model, if risen to the level of customary international law as the *lex specialis*, would represent a dramatic shift in the law of State responsibility and would supplant the *lex generalis*

* The views and opinions stated herein belong to the author only, and are not reflective of Dentons or Dentons US LLP.

"effective control" test in the context of imputed State responsibility for the unlawful cyber operations of non-State actors.

Keywords: *state responsibility, lex specialis, effective control, customary international law, cyber attribution*

1. INTRODUCTION

State attribution for the internationally wrongful cyber operations of a non-State actor is an issue that lies at the heart of a complicated and dynamic public debate. As in all areas of State responsibility, attribution in cyberspace is critical when determining the rights and responsibilities of States. Without proper attribution, States are limited in their options to defend against unlawful cyber operations, both within the *jus ad bellum* and *jus in bello*. Attribution in cyberspace is also incredibly difficult to establish factually. Non-State actors often mask their identity, and State actors often hide their true intentions. The degree to which the rules of State responsibility apply in cyberspace is therefore a matter of great public importance.

It is well accepted under the law of State responsibility that a State is generally responsible for the internationally wrongful acts of its *de jure* and *de facto* State organs.[1] It is equally well accepted that a State is generally responsible for the internationally wrongful acts of non-State actors, who are neither *de jure* nor *de facto* State organs, if the non-State actor in question operates on the instructions of, or under the direction or control of the State.[2] This general rule, commonly referred to as the "effective control"[3] test, is recognized as the *lex generalis* governing imputed State responsibility for the internationally wrongful conduct of non-State actors. As the *lex generalis*, this rule does not vary "with the nature of the wrongful act in question" unless there is a "clearly expressed *lex specialis*."[4] Scholars agree this principle, as the *lex generalis*, applies in cyberspace.[5]

[1] *Case Concerning the Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosn. & Herz. v. Serb. & Mont.)*, 2007 I.C.J. Rep 43 at 210 (Feb. 26) *("Bosnia Genocide")*.

[2] See International Law Commission, *Articles on the Responsibility of States for Internationally Wrongful Acts*, Report of the International Law Commission on the Work of its 53rd session, A/56/10, August 2001, UN GAOR, 56th Sess Supp No 10, UN Doc A/56/10(SUPP) (2001), art. 4(1) ("Articles on State Responsibility").

[3] See *Military and Paramilitary Activities in and against Nicaragua (Nicar. v. U.S.)*, 1986 ICJ Rep 14 at 109, 115 (June 27) ("*Nicaragua*")*; Armed Activities on the Territory of the Congo (Dem. Rep. Congo v. Uganda)*, 2005 I.C.J. Rep 168 at 228 (Dec. 19) ("*Congo*")*; Bosnia Genocide*, note 1, at 209. There is a competing test known as the "overall control" test, which is discussed further herein. *Prosecutor v. Tadić*, Case No. IT-94-1-I, Appeal Judgment, ¶ 118 (Jul. 15, 1999) ("*Tadić*").

[4] *Bosnia Genocide*, note 1, at 209; Articles on State Responsibility, note 2, at art. 55.

[5] See *Intl'l Grp. of Experts at the Invitation of the NATO Coop. Cyber Def. Ctr. Of Excellence, Tallinn Manual on the International Law Applicable to Cyber Warfare* at 3 (Michael N. Schmitt ed., 2013) (hereinafter "*Tallinn Manual*") (recognizing it is well accepted that international norms apply in cyberspace); *Int'l Grp. of Experts at the Invitation of the NATO Coop. Cyber Def. Ctr. Of Excellence, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* at 3, 94-100 (Michael N. Schmitt ed., 2017) (hereinafter "*Tallinn Manual 2.0*") (same); NATO Coop. Cyber Def. Ctr. Of Excellence, International Cyber Norms: Legal, Policy & Industry Perspectives at 13-14 (Anna-Maria Osula and Henry Rõigas eds., 2016) (same); United Nations, General Assembly, *Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*, A/68/98 (24 June 2013) (same); Peter Z. Stockburger, *Known Unknowns: State Cyber Operations, Cyber Warfare, and the Jus Ad Bellum*, 31 Am. J. Int'l L. 545, 548-550 (2016) (same).

This paper posits that a new *lex specialis* is forming, which would, if risen to the level of customary international law, supplant the *lex generalis* "effective control" test and allow for imputed State responsibility for the internationally wrongful cyber operations of non-State actors even in the absence of evidence demonstrating express direction or control. Specifically, a review of the general practice of interested States since 2014 reveals that the internationally wrongful cyber operations of non-State actors have been attributed to States in the absence of evidence of State direction or control where a number of factors are considered, including: (1) the relationship between the non-State actor and the State, if any; (2) any apparent influence the State exercises over the non-State actor; (3) the methods used by the non-State actor; (4) the motivations of the two parties, if known; (5) whether the two parties use similar code and/or technology; (6) technical capabilities; and (7) geographic location. This developing *lex specialis*, referred to herein as the "control and capabilities" test, would, if elevated to the level of customary international law, supplant the *lex generalis* rule of "effective control" and mark a significant shift in the law relating to imputed State responsibility in cyberspace.[6]

2. STATE RESPONSBILITY STANDARDS

State responsibility is generally premised on two elements: (1) the act or omission that breaches the international obligation; and (2) attribution of that act or omission to the responsible State.[7] As a general rule, the acts or omissions of a private person or group are not attributable to the State.[8] There are, however, exceptions.

A. Direct State Responsibility – De Jure and De Facto State Organs

The first exception relates to the acts and/or omissions of *de jure* or *de facto* State organs. A "State is responsible only for its own conduct, that is to say the conduct of persons acting, on whatever basis, on its behalf".[9] This includes acts:

> carried out by [State] official organs, and also by persons or entities which are not formally recognized as official organs under internal law but which must nevertheless be equated with State organs because they are in a relationship of complete dependence on the State.[10]

These types of individuals and groups are commonly referred to as *de jure* and *de facto* State organs. Both are recognized under Articles 4-6 of the Articles on State Responsibility,[11] and it is widely accepted that the conduct of *de jure* and *de facto* State organs is generally attributable to the State.[12]

B. Imputed State Responsibility – Control and Direction

The conduct of a non-State actor that is neither a *de jure* nor *de facto* State organ may also be attributable to the State "if the [non-State actor] is in fact acting on the instructions of, or

[6] *Nicaragua*, note 3, at 109, 115; *Congo*, note 3, at 228; *Bosnia Genocide*, note 1, at 168-211.

[7] Articles on State Responsibility, note 2, at art. 2(a)-(b); Michael N. Schmitt & Liis Vihul, *Proxy Wars in Cyberspace: The Evolving International Law of Attribution*, Fletcher Security Review, Vol. I, Issue II at 57 (2014) ("Schmitt & Vihul").

[8] Articles on State Responsibility, note 2, at art. 8, commentary (1).

[9] *Bosnia Genocide*, note 1, at 210.

[10] *Ibid.*; Articles on State Responsibility, note 2, at arts. 4-6.

[11] Articles on State Responsibility, note 2, at arts. 4-6.

[12] *Ibid.*; see *Bosnia Genocide*, note 1, at 210.

under the direction or control of, [the] State in carrying out the conduct[.]"[13] This direction and control test, known as the "effective control" test, reflects the *lex generalis* as it pertains to imputed State responsibility for the conduct of non-State actors.

1) Effective Control Test Background

The "effective control" test was first outlined by the International Court of Justice (ICJ) in its 1986 case concerning *Military and Paramilitary Activities in and against Nicaragua*. There, the Court examined whether US control over Nicaraguan *Contras* was sufficient to impute the actions of the *Contras* to the US under international law. In so doing, the Court made clear that although the actions of *de facto* State organs may be attributable to the State, the actions of non-State actors not totally dependent on the State, but who are nonetheless paid, financed and equipped by the State, would be attributed to the State only if it were established that the State "directed or enforced the perpetration" of the internationally wrongful act in question.[14] Under this "effective control" test, the Court determined that although the US was responsible for the general "planning, direction and support" given to the *Contra*s, the US was not internationally responsible for the internationally wrongful actions of the *Contras* because "there [was] no clear evidence of the US having actually exercised such a degree of control in all fields as to justify treating the [*Contras*] as acting on its behalf."[15] This "effective control" test, reflected in Article 8 of the Articles on State Responsibility,[16] was expressly endorsed by the International Group of Experts (IGE) in the recently published Tallinn Manual 2.0 as reflective of customary international law.[17]

2) Overall Control Test Introduced

Thirteen years later, in 1999, a competing test known as the "overall control" test was introduced by the Appeals Chamber of the United Nations International Criminal Tribunal for the Former Yugoslavia (ICTY) in its influential *Tadić* opinion. There, the tribunal rejected the "effective control" test, and instead applied an "overall control" test to determine the attribution of acts of hierarchically structured groups, such as a military unit or armed bands of irregulars or rebels under the *jus in bello*.[18] According to the tribunal, in such circumstances, the State will be internationally responsible for the wrongful acts of the non-State actor where the State exercises "overall control" over the non-State actor, and not the higher standard of "effective control." The tribunal adopted this approach with hierarchically structured groups because such groups are less likely to receive express direction and control from the State due to their "structure, a chain of command and a set of rules as well as the outward symbols of authority."[19] Instead, the State is more likely to exercise "overall control" over the unit. The tribunal's decision was also limited to the application of the doctrine within the *jus in bello* framework as it was determining the existence of an international armed conflict under the Fourth Geneva Convention of August

[13] Articles on State Responsibility, note 2, at art. 8.

[14] *Nicaragua*, note 3, at 61-64; Antonio Cassese, *The Nicaragua and Tadic Tests Revisited in Light of the ICJ Judgment on Genocide in Bosnia*, 18 Eurp. J. Int'l L. 649, 652 (2007) ("Cassese").

[15] *Nicaragua*, note 3, at 51, 62, 64-65.

[16] Articles on State Responsibility, note 2, at art. 8, commentary (7).

[17] *Tallinn Manual 2.0*, note 5, at 97, Rule 17, commentary (5) ("The International Group of Experts agreed that the phrase 'effective control' employed by the International Court of Justice in the *Nicaragua* and *Genocide* judgments captures the scope of the concept" under Article 8 of the Articles on State Responsibility).

[18] *Tadić*, note 3, at ¶ 120.

[19] *Ibid*.

12, 1949.[20] The “overall control” test is discussed in the commentaries to Article 8 of the Articles on State Responsibility, and is generally seen as a lower standard of attribution than the “effective control” test.[21] Under the “overall control” test, the State need only have control over the group generally, and not have given specific direction for each alleged internationally wrongful act in order for there to be imputed State responsibility.

3) Effective Control Test Revisited

The ICJ revisited the “effective control” test in 2007 in the case concerning the *Application of the Convention on the Prevention and Punishment of the Crime of Genocide*. There, the Court criticized the ICTY’s “overall control” test as going beyond the ICTY’s jurisdiction, and being unsupported in State practice. The Court reaffirmed the customary status of the “effective control” test, and announced that the actions of the Republika Srpska and certain paramilitary groups known as the Scorpions, Red Berets, Tigers and White Eagles were not attributable to the Federal Republic of Yugoslavia (FRY) because there was insufficient evidence demonstrating that State instruction and direction was given with regard to each operation in which the alleged violations occurred, and not generally in respect of the overall actions taken by the persons or groups of persons having committed the violations.[22] Consequently, and controversially, the Court determined that the FRY could not be internationally responsible for the acts committed by the non-State actors in question, most notably the massacres at Srebrenica.[23]

4) Safe Harbor / Duty to Prevent

There has been additional State practice endorsing a theory for imputed State responsibility wherein the State in question harbors and provides material support to those who cause harm in another State. After the September 11, 2001 attacks, the US invoked its right to self-defense pursuant to Article 51 of the United Nations Charter on the premise that the September 11 attacks constituted an “armed attack.”[24] The US attributed those attacks to the Taliban regime in Afghanistan because they were:

> made possible by the decision of the Taliban regime to allow the parts of Afghanistan that it controls to be used by this organization as a base of operation. Despite every effort by the United States and the international community, the Taliban regime has refused to change its policy. From the territory of Afghanistan, the Al-Qaeda organization continues to train and support agents of terror who attack innocent people throughout the world and target United States nationals and interest in the United States and abroad.[25]

[20] Art. 2 of the Statute of the International Tribunal for the Prosecution of Persons Responsible for Serious Violations of International Humanitarian Law Committed in the Territory of the Former Yugoslavia since 1991, S/RES/827 (1993) of 25 May 1993, Annex.

[21] Articles on State Responsibility, note 2, at art. 8, commentary (5) (noting it is a “matter for appreciation” in each case whether particular conduct was or was not carried out under the control of a State, to such an extent that the conduct controlled should be attributed to it).

[22] *Bosnia Genocide*, note 1, at 208.

[23] *Id*. at 206-08.

[24] UN Security Council, Letter Dated 7 October 2001 From the Permanent Representative of the United States of America to the United Nations Addressed to the President of the Security Council, UN Doc No S/2001/946 (2001).

[25] *Ibid*.

The global community generally accepted the legality of this action.[26]

This "safe harbor" principle is distinct from the "effective" or "overall" control tests because it is premised upon a separate doctrine of international law – namely, the duty to prevent trans-boundary harm and the "due diligence" principle as articulated in the 1941 *Trial Smelter* arbitration,[27] the ICJ's *Corfu Channel* judgment,[28] the ICJ's Advisory Opinion on the *Legality of the Threat or Use of Nuclear Weapons*,[29] and the ICJ's *Case Concerning Gabčíkovo-Nagymaros Project* judgment.[30] The "due diligence" principle has been applied to many areas of international law, including international environmental law,[31] human rights law,[32] and State responsibility.[33] It is distinct from imputed State responsibility because due diligence is an "obligation of conduct rather than of result[.]"[34] The doctrine therefore imposes responsibility directly on the State for its own actions, and is not on the basis of imputed State responsibility. The degree to which a State must exercise "due diligence" under international law remains highly contextual.[35]

5) Articles on State Responsibility

The "effective control" test is articulated in Article 8 of the Articles on State Responsibility, which provides that the:

> conduct of a person or group of persons shall be considered an act of a State under international law if the person or group of persons is in fact acting on the instructions of, or under the direction or control of, that State in carrying out that conduct.[36]

The "overall" control test is discussed in the commentary to Article 8,[37] which also provides that it is:

[26] See G.A. Res. 56/220, U.N. GAOR, 56th Sess., 91st mtg., U.N. Doc. A/RES/56/220 A-B (2001) (affirming the condemnation of the use of Afghan territory for terrorist activities); G.A. Res. 56/1, U.N. GAOR, 56th Sess., 1st mtg., Agenda Item 8, U.N. Doc. A/RES/56/1 (2001) (noting "those responsible for aiding, supporting or harbouring the perpetrators, organizers and sponsors of such acts will be held accountable"); S.C. Res. 1378, U.N. SCOR, 56th Sess., 4415th mtg., U.N. Doc. S/RES/1378 (2001) (condemning the Taliban "for allowing Afghanistan to be used as a base for Al-Quaida [sic]"); S.C. Res. 1373, U.N. SCOR, 56th Sess., 4385th mtg., U.N. Doc. S/RES/1373 (2001) (reaffirming principle that every State has "the duty to refrain from organizing, instigating, assisting or participating in terrorist acts in another State or acquiescing in organized activities within its territory directed towards the commission of such acts [.]").

[27] *Trail Smelter (U.S. v. Can.)*, 3 R.I.A.A. 1905, 1965 (1941).

[28] *Corfu Channel Case (U.K. v. Alb.)*, 1949 I.C.J. 4, 35 (Apr. 3).

[29] *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, 1996 I.C.J. 226, 29 (July 8)

[30] *Gabčíkovo-Nagymaros Project (Hung. V. Slovk.)*, 1997 I.C.J. 7, 53 (Sept. 25).

[31] Report of the United Nations Conference on the Human Environment Held at Stockholm, 5-16 June 1971, Principle 21, at 7, U.N. Doc. A/CONF.48/14 (1972), reprinted in 11 I.L.M. 1416, 1420 (1972); *Trail Smelter*, note 27, at 1965.

[32] See, e.g., Special Rapporteur on Violence against Women, its Causes and Consequences, *Report of the Special Rapporteur on violence against women, its causes and consequences*, Comm'n on Human Rights, U.N. Doc. A/HRC/23/49 (May 14, 2013) (by Rashida Majoo); *Velasquez Rodriguez Case*, 1988 Inter-Am. Ct. H.R. (ser. C) No. 4 (July 29) at ¶ 166; Compilation of General Comments and General Recommendations Adopted by Human Rights Treaty Bodies, U.N. Human Rights Comm'n, 44th Sess., Gen. Cmt. 20 art. 7 para. 13, at 32, U.N. Doc. HRI/GEN/1/Rev.1 (1994).

[33] See Ian Brownlie, *Principles of Public International Law*, 7th ed., 2008, pp. 275-285 ("Brownlie").

[34] David Freestone, *Advisory Opinion on the Seabed Disputes Chamber of the International Tribunal for the Law of the Sea on 'Responsibilities and Obligations of States Sponsoring Persons and Entities with Respect to Activities in the Area*, 15 Amer. Soc. Int. L. (March 2011).

[35] See *Makaratziz v. Greece*, (No. 50385/99), 2004 Eur. Ct. H.R. 694.

[36] Articles on State Responsibility, note 2, at art. 8.

[37] *Id*. at art. 8 commentary, (4).

> a matter of appreciation in each case whether particular conduct was or was not carried out under the control of a State, to such an extent that the conduct controlled should be attributed to it.[38]

Article 55 of the Articles of State Responsibility is entitled "Lex specialis" and provides that the Articles of State Responsibility "do not apply where and to the extent that the conditions for the existence of an internationally wrongful act or the content or implementation of the international responsibility of a State are governed by special rules of international law."

6) Tallinn Manual Recognition

The authors of the Tallinn Manual have also concluded that the "effective control" test is the *lex generalis* controlling imputed State responsibility for the conduct of non-State actors. Although the IGE recognized the tension between the "effective" and "overall" control tests in the 2013 Tallinn Manual,[39] they appear to have discarded the "overall control" test in the revised 2017 Tallinn Manual 2.0 and instead focus on the "effective control" test as the test that applies in cyberspace.[40] Specifically, Rule 17 of the Tallinn Manual 2.0, reflecting Article 8 of the Articles on State Responsibility, provides that cyber operations conducted by a non-State actor are attributable to a State when "engaged in pursuant to its instruction or under its direction or control[.]"[41] By this standard, the IGE notes that a State may, either by specific directions or by exercising control over a group, in effect assume responsibility for their conduct, with each case dependent on its own facts.[42] Instructions within this context "refers most typically to situations in which a non-State actor functions as a State's auxiliary."[43] And a State is in "effective control" of a particular cyber operation by a non-State actor whenever it is the State that "determines the execution and course of the specific operation and the cyber activity engaged in by the non-State actor is an 'integral part of that operation.' Effective control includes both the ability to cause constituent activities of the operation to occur, as well as the ability to order the cessation of those that are underway."[44]

3. TOWARD A NEW *LEX SPECIALIS* – THE "CONTROL AND CAPABILITIES" TEST

Based on the foregoing, it is well accepted that the "effective control" test is the *lex generalis* governing the imputation of State responsibility for the internationally wrongful conduct of non-State actors. As the *lex generalis*, this test applies in all situations unless there is an express *lex specialis* providing otherwise.[45] As explained below, a review of State practice since 2014 reveals that a new *lex specialis* has, in fact, begun to form that would, if risen to the

[38] *Id*. at art. 8 commentary, (5), citing the Iran-United States Claims Tribunal and the European Court of Human Rights as additional authority for the proposition that institutions have wrestled with the "problem of the degree of State control necessary for the purposes of attribution of conduct to the State[.]"

[39] See *Tallinn Manual*, note 5, at 32 (suggesting that a State's responsibility for cyber attacks may become rather common under the "effective control" standard).

[40] *Tallinn Manual 2.0*, note 5, at 4 (the rules adopted reflect customary international law as applied in the cyber context), 94-100 (applying "effective control" standard to imputed State responsibility for cyber operations of non-State actors).

[41] *Id*. at 94, Rule 17.

[42] *Id*. at 95, Rule 17, commentary 4.

[43] *Ibid*.

[44] *Id*. at 96, Rule 17, commentary 6, citing Articles on State Responsibility, art. 8, para. 3, commentary.

[45] *Bosnia Genocide*, note 1, at 209; Articles on State Responsibility, note 2, at art. 55.

level of customary international law, supplant the *lex generalis* "effective control" test for the imputation of State responsibility for the internationally wrongful cyber operations of non-State actors. Specifically, States that have attributed the internationally wrongful cyber operations of non-State actors to States since 2014 have done so on the basis of a multitude of factors, including but not limited to geographic location, methods and motivations, capabilities and technical indicators. This State practice appears to deviate from the rigid focus on control and direction as outlined by the "effective control" test, and instead focuses on a multi-factored analysis to determine State responsibility for cyber operations perpetrated by non-State actors. This State practice is creating a new test under customary international law referred to herein as the "control and capabilities" test.

A. Development of Custom

Customary international law is defined as the general practice of States accepted as law.[46] This definition comes from Article 38(1)(b) of the ICJ's Statute, and encompasses two elements: (1) long-term, widespread practice by interested States;[47] and (2) *opinio juris*, or the requirement that "[S]tates must believe that conformance with the practice is not merely designed, but mandatory and required by international law."[48]

For State practice to become a binding norm of customary international law, it must be "extensive and representative."[49] It does not, however, need to be universal.[50] There is no "precise number or percentage" of States participating required for the formation of custom, because the question is not "how many States participate in the practice" but instead which States participate.[51] Where States with influence in a particular area impacted by the normative development adopt a practice, it is given more weight in the analysis of whether the particular State practice has risen to the level of customary international law. Therefore, whether a particular State practice has achieved a level of compliance necessary for normative effect is a question of fact, involving an analysis of both physical and verbal acts of States.[52] The requirement of *opinio juris* refers to the legal conviction that a particular practice is carried out as required by law.[53] It is usually not necessary to demonstrate separately the existence of an *opinio juris* because it is generally contained within a particular dense practice. That said, proving *opinio juris* is still critical when proving the establishment of custom.

B. General Overview of "Control and Capabilities" Test

A survey of State practice since 2014 reveals that States generally do not adhere strictly to the "effective control" test set forth under Article 8 of the Articles on State Responsibility or Rule 17 of the Tallinn Manual 2.0 when attributing the internationally wrongful cyber operations of non-State actors to the State. They instead apply the "control and capabilities" test, examining the methods and motivations of the non-State actor, their geographic location, and whether,

[46] Statute of the International Court of Justice, art. 38(1)(b).

[47] Lynn Loschin, *The Persistent Objector and Customary Human Rights Law: A Proposed Analytical Framework*, 2 U.C. Davis J. Int'l L. & Pol'y 147, 148 (1996) (quoting Ian Brownlie, Principles of Public International Law 6-7 (2d ed. 1973) who noted that elements of this part of custom are duration, uniformity and consistency of practice, and generality of practice) ("Loschin").

[48] *Ibid*.

[49] Loschin, note 47, at 148.

[50] Stockburger, note 5, at 564.

[51] *Ibid*.

[52] *Ibid*.

[53] See generally Lori F. Damrosch et al., *International Law Cases and Materials* 3 (4th ed. 2001); *Nicaragua*, note 3, at 126.

if at all, the State had similar technical capabilities. Below is a survey of that State practice, which this paper argues reflects the development of a *lex specialis* for the imputation of State responsibility for the unlawful cyber operations of non-State actors.

C. State Practice

1) Pre-2014 Private Attribution Based on Control and Capabilities

Prior to 2014, although public attribution of internationally wrongful cyber operations was virtually non-existent, private attribution began to follow the control and capabilities model. In 2007, after Estonia was hit with a wave of distributed denial of service ("DDoS") attacks after deciding to remove a Soviet-era bronze soldier monument from its location in central Tallinn, Estonia, a number of scholars and jurists privately attributed the attacks to Russia.[54] Evidence showed that the "hackers claimed to be Russian, the tools to hack and deface were contained in Russian websites and chatrooms, and the attacks picked a day of" significance to most Russians.[55] Moreover, although the botnets used included computers from different countries, at least some of the attacks "originated from Russian IP (internet protocol) addresses, including those of state institutions."[56] In 2008, a similar scenario played out with Russia being privately blamed for carrying out a DDoS attack during its 2008 conflict with Georgia[57] wherein fifty-four "web sites in Georgia related to communications, finance, and the government" were attacked[58] "immediately before and continu[ing] throughout the armed conflict between" the two States.[59] All signs pointed to a Russian hacker community as the responsible perpetrator,[60] including the fact that coordination for the attacks took place in the Russian language, and in Russian or Russia-related "fora."[61] Likewise, despite confirmation from State actors regarding the 2010 Stuxnet virus,[62] and although it has been reported that the Stuxnet virus was formally developed under former US administrations,[63] no formal attribution has been declared. And in 2013, the controversial Mandiant Report attributed APT1 attacks to the Chinese State based on the geographic location of the bad actors, the methods and capabilities of the actors in question, and the motivations of the Chinese State. The report followed the US Department of Defense's 2013 Report to Congress that indicated some of the 2012 cyber intrusions into US government computers appeared to be attributable directly to the Chinese State (without providing detail as to why that attribution was provided).

2) Sony - 2014

It was not until 2014 that a State first publicly attributed what appeared to be a non-State actor's unlawful cyber operations to a State, and stated the reasons publicly for the attribution. In 2014, Sony Pictures was hit with a highly publicized DDoS attack of unknown proportions after its

[54] See Ian Traynor, "Russia accused of unleashing cyberwar to disable Estonia", *The Guardian*, May 16, 2007, https://www.theguardian.com/world/2007/may17/topstories3.russia.

[55] *Ibid*.

[56] Marco Roscini, "Evidentiary Issues in International Disputes Related to State Responsibility for Cyber Operations", in *Cyber War: Law and Ethics for Virtual Conflicts* 216 (Jens David Ohlin et al. eds., 2015) ("Roscini").

[57] David Hollis, "Cyberwar Case Study: Georgia 2008", Small Wars J., Jan. 6, 2011, at 1.

[58] *Ibid*.

[59] Roscini, note 56, at 216.

[60] *Ibid*.

[61] *Ibid*.

[62] Katharina Ziolkowski, Stuxnet - *Legal Considerations* 3 (2012).

[63] David P. Fidler, "Recent Developments and Revelations Concerning Cybersecurity and Cyberspace: Implications for International Law", Am. Soc'y Int'l L. Insights (June 20, 2012), available at https://www.asil.org/insights/volume/16/issue/22/recent-developments-and-revelations-concerning-cybersecurity-and.

computer systems were compromised by suspected North Korean tied hackers.[64] The attack surrounded the release of the movie "The Interview" about the fictional assassination of the North Korean leader, Kim Jong-un. Prior to the movie's release, the spokesperson for North Korea's Ministry of Foreign Affairs said in a statement "that the country would take 'a decisive and merciless countermeasure' if the United States' government permitted Sony to make its planned Christmas release of the comedy."[65] In November, Sony's computer systems were compromised, and embarrassing e-mail communications from its CEO were leaked online. A group calling itself the "Guardians of Peace" claimed responsibility for the attack.[66]

Unbeknownst to North Korea, years earlier US officials gained access to the North Korean cyber infrastructure and implanted malicious code to track North Korean operations, allowing them to identify certain IP addresses that were being used to send spear phishing e-mails from North Korea.[67] This capability allowed US officials to trace the origins of the Sony attack[68] and shortly thereafter, in December 2014, publicly attribute the attack to the "North Korean government[.]"[69]

The US Federal Bureau of Investigation (FBI) attributed the attack to the North Korean State not on the basis of direction or control, but instead on the methods and motivations of the attackers and the North Korean government, including: (1) data deletion malware used in the attack revealing links to other malware that the FBI knew North Korean actors previously developed, including "similarities" in specific lines of code, encryption algorithms, data deletion methods, and compromised networks; (2) a significant overlap between the "infrastructure" used in the attack and other malicious cyber activity the US government had previously linked directly to North Korea; (3) IP addresses associated with known North Korean infrastructure; and (4) tools used in the attack that had "similarities" to a cyber attack in March 2013 against South Korean banks and media outlets, "which was carried out by North Korea".[70] This information about methods and capabilities led the US to publicly attribute the attack to North Korea without mention of any direction or control by the North Korean government. In response, the US imposed economic sanctions on North Korea, further reflecting the attribution of the attack to the North Korean State. North Korea also reportedly suffered widespread Internet outages in December 2014, raising the possibility that countermeasures were taken. These countermeasures were likely taken on the basis that the US viewed North Korea's actions, through the non-State actor's cyber operations, as an internationally wrongful act under the law of State responsibility, thereby justifying the imposition of proportional countermeasures.

3) Iran - 2016

In 2016, the US Department of Justice (DOJ) publicly attributed the cyber operations of private

[64] Michael Cieply & Brooks Barnes, "Sony Cyberattack, First a Nuisance, Swiftly Grew Into a Firestorm", *N.Y. Times* (Dec. 30, 2014), available at http://www.nytimes.com/2014/12/31/business/media/sony-attack-first-a-nuisance-swiftly-grew-into-a-firestorm-.html (asserting that Sony was slow to realize the magnitude of the public relations complexities, financial loss, and uniqueness of the cyber attack).

[65] *Ibid*.

[66] FBI, "Update on Sony Investigation", *Press Release* (Dec. 19, 2014), available at https://www.fbi.gov/news/pressrel/press-releases/update-on-sony-investigation.

[67] David E. Sanger and Martin Fackler, "N.S.A. Breached North Korean Networks Before Sony Attack, Officials Say", *N.Y. Times* (Jan. 18, 2015), available at http://www.nytimes.com/2015/01/19/world/asia/nsa-tapped-into-north-korean-networks-before-sony-attack-officials-say.html?_r=0.

[68] *Ibid*.

[69] FBI, note 66.

[70] *Ibid*.

non-State actors to Iran based on a control and capabilities analysis.[71] In March 2016, a grand jury in the Southern District of New York criminally indicted seven Iranian individuals who were employed by two Iran-based computer companies, ITSecTeam and Mersad Company that "performed work on behalf of the Iranian Government, including the Islamic Revolutionary Guard Corps, on computer hacking charges related to their involvement in an extensive campaign of over 176 days of distributed denial of service (DDoS) attacks".[72] The seven individuals were alleged to have "launched DDoS attacks against 46 victims, primarily in the U.S. financial sector, between late 2011 and mid-2013."[73] These attacks purportedly disabled victim bank websites, prevented customers from accessing their accounts online, and cost tens of millions of dollars in damage.[74] One of the defendants was also charged with obtaining unauthorized access into the Supervisory Control and Data Acquisition systems of the Bowman Dam located in upstate New York in August and September 2013.[75] In releasing the indictment, the US stated:

> Like past nation state-sponsored hackers, these defendants and their backers believed that they could attack our critical infrastructure without consequence, from behind a veil of cyber anonymity[.] This indictment once against shows there is no such veil - we can and will expose malicious cyber hackers engaging in unlawful acts that threaten our public safety and national security.[76]

In the indictment, the US attributed the actions of the seven private individuals to the State of Iran because the individuals purportedly "performed work on behalf of" the Iranian Government, as evidenced by the scope and capabilities of their cyber operations.[77] In the press release accompanying the indictment, the US Department of Justice noted these individuals had "ties" to Iran's Islamic Revolutionary Guard.[78] And without explaining the instruction from the State involved, the indictment alleged Ahmad Fathi, as the leader of the defendants, was "responsible for managing computer intrusion and cyber projects being conducted on behalf of the Government of Iran."[79] The indictment also alleged that defendant Amin Shokohi received credit for his work from the Iranian Government towards completion of his mandatory military service in Iran.[80] These allegations did not discuss direction or control, and instead focused on means and methods, capabilities, and motivations of the perpetrators in effectuating State attribution.

4) Russia - 2016 / 2017

a) DNC / US Election

In June 2016, the private security firm CrowdStrike issued a report entitled "Bears in the Midst:

[71] US Dep't of Just., "Manhattan U.S. Attorney Announces Charges Against Seven Iranians For Conducting Coordinated Campaign of Cyber Attacks Against U.S. Financial Sector On Behalf Of Islamic Revolutionary Guard Corps - Sponsored Facilities", *Press Release* (Mar. 24, 2016), available at https://www.justice.gov/usao-sdny/pr/manhattan-us-attorney-announces-charges-against-seven-iranians-conducting-coordinated ("Iran Press Release"); *U.S. v. Ahmad Fathi, et al., Case* No. 16 Cr. 48 (S.D.N.Y. Mar. 24. 2016) ("Iran Indictment").

[72] Iran Press Release, note 71.

[73] *Ibid*.

[74] *Ibid*.

[75] *Ibid*.

[76] *Ibid*.

[77] Iran Indictment, note 71, at ¶ 1.

[78] Iran Press Release, note 71.

[79] Iran Indictment, note 71, at ¶ 11.

[80] *Id*. at ¶ 13.

Intrusion into the Democratic National Committee",[81] which described an investigation into the 2015 and 2016 cyber breaches of the Democratic National Committee's (DNC) computer systems.[82] In the report, CrowdStrike identified two "sophisticated adversaries on the network - COZY BEAR and FANCY BEAR."[83] CrowdStrike concluded these two adversaries were linked to the Russian State because of their "advanced methods consistent with nation-state level capabilities including deliberate targeting and 'access management' tradecraft[,]" and because both adversaries "engage in extensive political and economic espionage for the benefit of the government of the Russian Federation and are believed to be closely linked to the Russian government's powerful and highly capable intelligence services".[84] CrowdStrike determined that COZY BEAR had infiltrated the DNC's computer systems in the summer of 2015, and FANCY BEAR had breached the network in April 2016.[85]

In October 2016, the US Director of National Intelligence (DNI) issued a joint statement on behalf of the DNI and the US Department of Homeland Security (DHS) stating that the US Intelligence Community was "confident" that the "Russian Government directed the recent compromises of e-mails from US persons and institutions, including from US political organizations."[86] The evidence to support this conclusion, however, was that the subsequent disclosures of "hacked e-mails" that followed the DNC intrusion were "consistent with the methods and motivations of Russian-directed efforts[.]"[87] Specifically, the DNI and DHS stated that such thefts and disclosures are reflective of past Russian efforts "across Europe and Eurasia" to "influence public opinion there."[88] Based on the "scope and sensitivity of these efforts," the DNI and DHS concluded that "only Russia's senior-most officials could have authorized these activities."[89]

Several months later, in January 2017, the DHS and the FBI issued a Joint Analysis Report entitled "GRIZZLY STEPPE - Russian Malicious Cyber Activity." In the report, the DHS and FBI expanded on the October 2016 Joint Statement issued by the DNI and DHS, and publicly attributed the DNC cyber intrusion to the Russian State based on a series of "technical indicators":

> Previous JARs have not attributed malicious cyber activity to specific countries or threat actors. However, public attribution of these activities to [Russian civilian and military intelligence Services] is supported by technical indicators from the U.S. Intelligence Community, DHS, FBI, the private sector, and other entities.[90]

[81] Dmitri Alperovitch, "Bears in the Midst: Intrusion into the Democratic National Committee," *CrowdStrike Blog* (June 15, 2016), available at https://www.crowdstrike.com/blog/bears-midst-intrusion-democratic-national-committee/.

[82] *Ibid*.

[83] *Ibid*.

[84] *Ibid*.

[85] *Ibid*.

[86] Director of National Intelligence, "Joint Statement from the Department of Homeland Security and Office of the Director of National Intelligence on Election Security," *Joint Statement* (Oct. 7, 2016), available here https://www.dni.gov/index.php/newsroom/press-releases/215-press-releases-2016/1423-joint-dhs-odni-election-security-statement.

[87] *Ibid*.

[88] *Ibid*.

[89] *Ibid*.

[90] DHS & FBI, "GRIZZLY STEPPE - Russian Malicious Cyber Activity", *Joint Analysis Report* (Dec. 29, 2016), available here https://www.us-cert.gov/sites/default/files/publications/JAR_16-20296A_GRIZZLY%20STEPPE-2016-1229.pdf.

According to the Joint Report, those "technical indicators" prove the threat actors are "likely associated" with the Russian State.[91]

The Joint Report was widely panned by industry experts for its purported failure to provide a "smoking gun" that showed Russian control or direction over the DNC intrusion, and for using false technical indicators linking the purported threat actors to the Russian State.[92] But the focus of the Joint Report was not on express direction or control. It instead focused on capabilities, methods, motivations and technical indicators, further reflecting the development of the control and capabilities test as the growing *lex specialis* for imputed State attribution for the unlawful cyber operations of non-State actors.

b) Yahoo Breach

In March 2017, the US DOJ announced the indictments of five individuals, including two Russian officials, for "computer hacking, economic espionage and other criminal offenses in connection with a conspiracy, beginning in January 2014, to access Yahoo's network and the contents of webmail accounts."[93] In the indictment, the DOJ alleged that "officers of the Russian Federal Security Service" and "intelligence and law enforcement agency of the Russian Federation" "conspired together and with each other to protect, direct, facilitate, and pay criminal hackers to collect information through computer intrusions in the United States and elsewhere."[94] The evidence cited to link the Russian officials to the "criminal hackers", however, was less than express direction and control. Instead, the evidence included: (1) that the criminal hackers obtained evidence of "information of predictable interest to the FSB", including access to "Russian journalists and politicians critical of the Russian government" (i.e., motivations); (2) the geographic location of the criminal hackers; (3) the ability of the Russian State to "arrest and prosecute" the criminal hackers and its failure to do so; and (4) threadbare allegations that the Russian officials provided direction to the criminal hackers.[95] The focus therefore was on motivations, capabilities and geographic proximity, in combination with conclusory allegations of State direction. The test applied was not the "effective control" test. The DOJ instead focused on a control and capabilities analysis.

[91] *Ibid*.

[92] See, e.g., Kelly Jackson Higgins, "DHS-FBI Report Shows Russian Attribution's A Bear", *Dark Reading* (Jan. 4, 2017), available at http://www.darkreading.com/threat-intelligence/dhs-fbi-report-shows-russian-attributions-a-bear/d/d-id/1327828; Justin Raimodo, "The Evidence That Russia Hacked The DNC Is Collapsing", *Zero Hedge* (Mar. 26, 2017), available at http://www.zerohedge.com/news/2017-03-25/; Shaun Waterman, "DHS slammed for report on Russian hackers", *Cyber Scoop* (Jan. 6, 2017), available at https://www.cyberscoop.com/dhs-election-hacking-grizzly-steppe-iocs/.

[93] US Dep't of Justice, "U.S. Charges Russian FSB Officers and Their Criminal Conspirators for Hacking Yahoo and Millions of Email Accounts: FSB Officers Protected, Directed, Facilitated and Paid Criminal Hackers", *Press Release* (Mar. 15, 2017), available at https://www.justice.gov/opa/pr/us-charges-russian-fsb-officers-and-their-criminal-conspirators-hacking-yahoo-and-millions; Ellen Nakashima, "Justice Department charges Russian spies and criminal hackers in Yahoo intrusion", *The Washington Post* (Mar. 15, 2017) available at https://www.washingtonpost.com/world/national-security/justice-department-charging-russian-spies-and-criminal-hackers-for-yahoo-intrusion/2017/03/15/64b98e32-0911-11e7-93dc-00f9bdd74ed1_story.html?utm_term=.d0a7e78e2d2d.

[94] *United States of America v. Dmitry Dokuchaev, et al*., No. CR17-103, Indictment at ¶ 1(N.D. Cal. Feb. 28, 2017), available at https://www.justice.gov/opa/press-release/file/948201/download.

[95] *Id*. at ¶¶ 1-6, 34.

4. CONCLUSION

The foregoing examples of State practice support the conclusion that imputed State responsibility for the unlawful cyber operations of non-State actors who are neither *de jure* nor *de facto* State organs is being assigned without rigid adherence to the "effective control" test. State attribution is instead being assigned based on a control and capabilities test, examining motivations, geographic location, technical indicators, and relationship between the non-State actor and the State. In 2014, the US publicly attributed the Sony attack to North Korea based on similarities between the code and infrastructure used by the malicious actor and the North Korean State. In 2016, the US publicly attributed certain cyber attacks to Iran based on the relationship between the purported bad actors and the State. In 2016 and 2017, the US publicly attributed the cyber intrusion of the DNC computer system to the Russian State based on technical indicators and similarities in motivations between the malicious actors and the Russian State. And most recently, the US publicly linked the 2014 intrusion of Yahoo to Russian State officers based on the geographic and motivational similarities between the criminal hackers and the Russian intelligence officers involved. In none of these cases did the State apply a rigid effective control test to determine attribution.

These State examples, of course, are not conclusive. This paper does not argue that such limited examples of State practice, alone, constitute a binding principle of customary international law. This State practice does, however, indicate that a *lex specialis* is forming that, if risen to the level of customary international law, would supplant the *lex generalis* "effective control" test, endorsed in Article 8 of the Articles on State Responsibility and Rule 17 of the Tallinn Manual 2.0.

The "control and capabilities" test, as it is developing, is not without its drawbacks. Relying solely on digital forensics to establish attribution is rife with risk. Digital evidence is volatile and has a short life span.[96] And although digital evidence may lead to the identification of the computer or computer systems from which the cyber event was triggered, "it does not necessarily identify the individual(s) responsible for the cyber operation (as the computer may have been hijacked, or the IP spoofed)."[97] These are difficult policy discussions that go beyond the scope of this paper.

The purpose of this paper is to highlight the growing trend that State attribution for the unlawful cyber operations of non-State actors who are neither *de jure* nor *de facto* State organs is deviating from the "effective control" test, and is instead focusing on a multitude of factors, including control and capabilities. This shift in State practice is reflective of a developing *lex specialis*. With more State practice, this new *lex specialis* will help shape State response to cyber operations, and will generate additional, and hopefully positive attribution analysis regimes for operational use.

[96] Marco Roscini, "Evidentiary Issues in International Disputes Related to State Responsibility for Cyber Operations", 50 Tex. Int'l L.J. 233, 264 (2015).

[97] *Ibid*.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Defending the Grid: Back-fitting Non-Expandable Control Systems

Robert Koch
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
robert.koch@UniBw.de

Teo Kühn
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany

Abstract: Network security has been a lively research area for more than 35 years and numerous products are available nowadays. In contrast to business networks, which were interconnected from the beginning by design, Industrial Control Systems (ICSs) have always been self-contained networks. Because their key features are real-time capability and their operational constraint to function as specified under maximum load (Carlson 1998), security has played only a subordinate role. Nowadays these systems are increasingly connected to the Internet; for example, wind power is more frequently used and generators are installed in remote and scattered regions that are difficult to access, so remote administration based on mobile communications is required, often using the Internet.

While numerous papers on securing ICSs have been published, interest rose after the incidents in Iran's enrichment plant in Natanz where the SCADA system controlling the centrifuges was attacked by the Stuxnet worm. Even with these intensified efforts, the current security situation is insufficient as numerous security systems perform inadequately in real-world environments. Elderly ICSs are also still in use which cannot be retrofitted easily or at all, and modern systems are often still not developed with 'security by design' in mind. In contrast to general purpose systems, a relatively limited number of processes are executed within ICSs. This enables the use of detection mechanisms based on voltage levels and current drain to build lightweight detection systems without huge databases by measuring the current drain during normal system operation.

Our concept combines the advantages of different detection principles and enhances them to build an Intrusion Detection System usable within ICSs. It is implemented based on low-priced components and can be integrated even in older, originally non-expandable systems.

Keywords: *power-based intrusion detection, ICS and SCADA security, tamper-resistant intrusion detection, anomaly-based IDS, retrofitting IDS, lightweight IDS*

1. INTRODUCTION

Intrusion Detection Systems (IDSs) have been under intense research for more than 35 years. Monitoring system behaviour to learn patterns and detect abnormal behaviour was the first detection technique in 1980. Within this group, anomaly-based detection is the predominant detection principle: benign behaviour is observed over a period of time, and afterwards a model is built based on the observations. During operation, the state of the system is measured and compared to the expectations of the model. If there is a significant deviation above a defined threshold, an alarm is raised. While this approach is able to detect new and unknown threats, it suffers from a high number of false alarms.

More recently, knowledge-based detection techniques have been developed. Here, mainly acquirement of malicious activities is used to realise misuse detection based on attack descriptions (signatures). This technique is able to lower false positive rates, but only known attacks can be detected. In the 1990s, this was a beneficial approach because a limited number of new malicious codes were published repeatedly. Anyway, the rapidly increasing professionalisation of the cybercrime market and the exploding numbers of malicious programs are forcing huge signature databases and time-consuming scans, therefore renewing the need for behaviour-based techniques.

But even with these extensive efforts, successful cyber attacks happen every day with an increasing amount of damage and physical effect. An extensive study of the annual costs to the global economy by McAfee (2014) gives an estimate of more than $400 billion in losses, and an estimate made by Juniper (2015) projects the loss at over $2 trillion by 2019.

ICSs and Supervisory Control and Data Acquisition (SCADA) systems are also under continuous attack. Often designed years or decades ago and originally conceived as isolated networks and systems, nowadays more ICSs are connected to the Internet. For example, wind power is increasingly used in the energy sector and generators are installed in remote regions difficult to access. This requires remote administration which is realised using mobile communications. The interconnection of plants can also be required to open up new business models. Therefore, factories and power plants are no longer conceivable without the use of ICSs.

Even though the security of those systems is obviously very important, this is currently not reflected in the real world. Analysis by Andreeva and colleagues (2016) for Kaspersky Lab concluded that:

> [a]lthough they are designed for critical infrastructures, industrial-sector devices are not secure by default; they contain the same type of vulnerabilities as any other system: including buffer overflows, hardcoded credentials, authentication bypass, cross-site scripting, and many others.

The situation is even more alarming, as the SANS 2016 State of ICS Security Survey (Harp and Gregory-Brown 2016) discloses that 67% of all participants 'perceived severe or high levels of threat to control systems, up from 43% in 2015', but 'security for ICSes has not improved in many areas and that many problems identified as high-priority concerns in our past surveys remain as prevalent as ever'. The recent report of the US Department of Energy (2017, p. 18) states:

> In the current environment, the U.S. grid faces imminent danger from cyber attacks. Widespread disruption of electric service because of a transmission failure initiated by a cyber attack at various points of entry could undermine U.S. lifeline networks, critical defense infrastructure, and much of the economy; it could also endanger the health and safety of millions of citizens.

As the required processing power in ICSs is often precisely defined, these systems routinely lack adequate intrusion detection components and cannot be upgraded. This essay will explore how real-world-usable intrusion detection with high detection and low false alarm rates can be realised for ICSs. Section 2 will discuss the particularities of ICSs and also identify available research and its shortcomings. Unlike general purpose systems, a relatively limited number of processes are executed within ICSs. These processes also remain unchanged for a long time. This enables the use of power-based detection mechanisms to build lightweight detection systems without huge databases, by the creation of an extensive comparative dataset and measuring the current drain during normal system operation, which will be presented in Section 3 A. Based on this, different possibilities of distributed intrusion detection for ICS and SCADA are presented in Section 3 B. Finally, Section 4 summarises core aspects and presents next steps.

2. SCADA & ICS CHALLENGES

A. Particularities of ICSs

ICSs control our entire modern everyday life. Not only do factories and power plants rely on them, critical infrastructure like water supply and transportation are completely IT-based. Some challenges for ICSs directly emerge from their regular use:

> Some SCADA systems are placed in remote locations […] and are designed to run nonstop for months or years. Through Internet connections to SCADA systems, managers can have precise and remote control of their infrastructure machinery. This arrangement also reduces the required number of workers in the field. Industrial control systems were originally designed to operate in isolation, without connection to other networks. As a result, cyber security controls were not built in (Wilson 2012, p. 4).

Therefore, due to the originally unplanned interconnection with non-trustworthy networks such as the Internet, there are often no protective measures available: the challenge of the *unavailability* of security components. Neither is air-gapping an adequate protection mechanism nowadays (Andreeva et al. 2016), as not only was bridging the airgap demonstrated impressively by Stuxnet, but numerous new concepts have also been demonstrated (Guri et al. 2016).

Secondly, ICSs are specifically designed for their respective applications: '[a]n important operational constraint of a SCADA network is that it functions as specified under maximum load. Security cannot hinder such operation' (Carlson 1998, p. 6). This highlights a burning issue for securing ICSs which are already in use, as they generally can neither be updated with additional software nor easily retrofitted with additional hardware; this is the challenge of *non-expandability*. Certification can also be a further hurdle in some areas like medical equipment, preventing a change of hardware or software.

Today's development cycles of commercial off-the-shelf (COTS) products are often of less than one year, and products typically have a limited support technology lifetime of only a few years. In contrast, control systems are used for a long time – often several decades. This presents challenges, like the supply of spare parts or fixing bugs in outdated and no longer supported proprietary software.

Even if software support is still available, patching can be challenging, even for general purpose systems. While there is work in progress on improving and automating program repair (Le 2016), patching is still complicated and the complexity of applying patches should not be underestimated (Cavusoglu, Cavusoglu and Zhang 2008). For example, issues can arise based on bad patch quality, being unable to fix the focused software flaws, interrupting software functionality or introducing new vulnerabilities (Mimoso 2015). Childs (2015) highlighted that:

> [i]n the last half of 2014 alone, users incurred major disruptions after installing patches from Microsoft, Apple, Adobe, and Oracle. There are also times when a security patch itself introduces a security problem. In other cases, the patches do not work as advertised.

While this is challenging for all IT systems, applying patches to ICSs is even more difficult because of 24/7 operation, the lack of testing possibilities before applying a patch, or technical limitations of the target systems. This is the challenge of being *unmaintainable*.

Meanwhile, the increasing quality of attacks is endangering the livelihood of today's societies. For example, the service outages of the Ukrainian power distribution company Kyivoblenergo on 23 December 2015 affected seven 110 kV and 23 35 kV substations, resulting in several outages that caused approximately 225,000 customers to lose power across various areas (Lee et al. 2016). The subsequent investigation of the incident showed that the attackers were able to perform long-term reconnaissance and to use and exploit different attack vectors including spear phishing emails, harvesting system and network credentials, operating SCADA systems, writing and distributing malicious firmware and rendering devices inoperable and unrecoverable (Ibid.). This is the challenge of *attack sophistication*.

Currently, Cyber Commands are being built up in nearly every nation worldwide. The significance of the cyber space for military operations is undisputed:

> In July 2016, Allies reaffirmed NATO's defensive mandate and recognised cyberspace as a domain of operations in which NATO must defend itself as effectively as it does in the air, on land and at sea (NATO 2016).

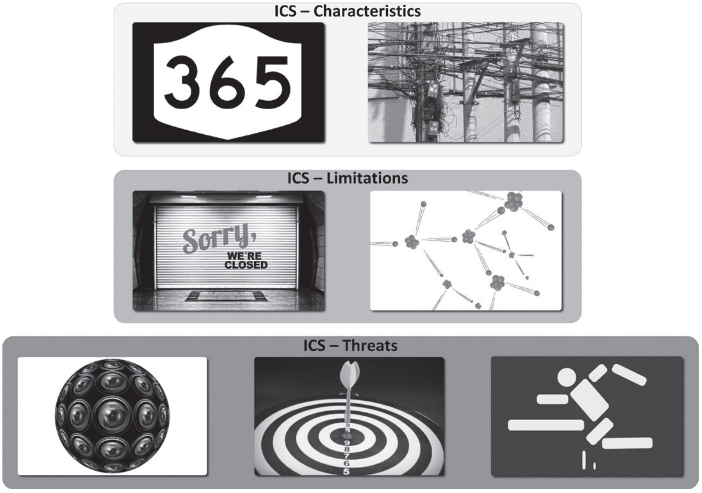
A massive cyberattack could even trigger a collective response by NATO (Reuters 2016). From a military point of view, targeting critical infrastructure can be advantageous, such as disabling the adversary's power grid accurately and promptly without endangering own ground troops, and even being able to make further use of it in contrast to physical destruction (Saglam 2014). Therefore, critical infrastructures are increasingly *eyeballed* as attractive targets.

As the Western critical infrastructures have been scanned systematically for years (Paganini 2014), attack preparation is greatly facilitated. New search engines such as Shodan, scanning all devices connected to the Internet and reading out banner massages, exacerbate the situation even further:

> originally intended to improve security and discover information about machines linked to the Internet, [Shodan] revealed that many SCADA computers that automate water plants and power grids were wide open to exploitation by hackers. The Shodan search engine has reportedly revealed water-treatment facilities, power plants, particle accelerators and other industrial control systems that may have security vulnerabilities (Wilson 2012, p. 4).

While a significant increase in the number of attacks based on Shodan's search results was *not* identified by some research (Bodenheim 2014), at least the reconnaissance is greatly simplified. The situation is getting worse as low-priced Zero Day vulnerabilities for SCADA systems can now be bought easily. For example, the Russian company GLEG Ltd. sells the exploit packages SCADA+ and MedPack, providing hundreds of modules and regularly adding new Zero Day vulnerabilities. For example, SCADA+ 1.5 contained a Zero Day for a vulnerability in 'Carel Plant Visor Pro', which is 'used on nuclear plants e.g. in Canada, [and states that the] exploit allows credentials steal' (GLEG Ltd. 2015). While the cheap prices allow companies to buy products for identifying vulnerabilities in their products and fix them, ICSs are often barely patchable. This opens up the challenge of a *falling attack threshold (easiness of attack)*.

FIGURE 1. PECULIARITIES AND ENDANGERMENTS OF ICSS



Having a look at defence mechanisms,

> some traditional cybersecurity practices and procedures that are standard for office IT systems may not work as well for SCADA systems. For example, because industrial SCADA equipment must send monitoring signals to other industrial controller equipment within milliseconds, traditional antivirus software or network intrusion detection devices will not fit very well (Wilson 2012, p. 7).

One must also not forget that security systems are just program code, introducing an additional number of programming errors (Panko 2008; Baishakhi et al. 2014). Tavis Ormandy (2016) demonstrates that such vulnerabilities are not rare in security programs, such as the remote code execution flaws in CVE-2016-2208. Also, the recommended best practice of building up a defence in depth can be more complicated than expected, as negative effects can occur if the measures are not coordinated in detail (Wolff 2016). This is the challenge of *interference* of traditional security systems.

Taking into consideration an evaluation of control systems cybersecurity made by Idaho National Laboratory (2008), Table 1 summarises and opposes the identified security challenges for general purpose and ICSs.

TABLE 1. COMPARISON OF SECURITY CHALLENGES

Security Challenge	General Purpose	Control Systems
Non-expandable hard-/software	Expandable, interchangeable	Non-expandable
Maintenance	Regular scheduled	Limited, system restrictions
System interference	Generally accepted	Unacceptable
Security systems	Common, widely used	Unavailable, system restrictions
Attack demand	Cybercrime, espionage	Cyber Commands, Terrorism
Attack sophistication	Rising	Rapidly rising
Attack simplification	Countermeasures	Low attack hurdle

B. Related Work

A comprehensive overview of SCADA-specific intrusion detection systems was given by Zhu and Sastry (2010). They analysed and compared different behaviour- and knowledge-based as well as hybrid systems for SCADA (PVAEB, IBM NADS, SRI Modbus, WFBNI, SHARP, IDEM, AAKR-SPRT, EMISDS and MAAC-UFE) and concluded that 'barely any of these systems has a performance evaluation on the false alarms that it generates' (Ibid., p. 13). This will also be a challenge when comparing our evaluation results with other works (see Section 3A.) Mitchell and Chen (2014) surveyed intrusion detection techniques for cyber-physical systems by analysing 28 IDSs. Open research leads in areas such as network-based approaches and the use of behavior-based detection techniques were identified.

Yang et al. (2006) analysed the application of anomaly-based intrusion detection for SCADA systems. They used an auto associative kernel regression model and statistical probability ratio test, applied to a simulated SCADA system. Their results showed that anomaly-based methods can be generally used to detect a variety of common attacks also within SCADA systems. Yang et al. (2014) proposed a multi-attribute SCADA-specific IDS for power networks. Their system consists of three attributes: access control whitelists, protocol-based whitelists and behaviour-based rules, where normal and correct behaviour are found by deep packet inspection. The focus of their evaluation is on the maximum execution time, showing that the standard communication delivery time performance requirements for electric power substation automation (IEEE Standard 1646-2004) are fulfilled. A performance evaluation of the intrusion detection is not given. Also, the system has to be integrated into the target system, cannot cope with encrypted connections, and is limited to power networks.

An integrated OCSVM mechanism for intrusion detection in SCADA systems was proposed by Maglaras et al. (2014). They use a distributed class support vector machine to generate information about the origin and time of an intrusion by reading network traffic and evaluating clusters based on the source of the network packets. Sayegh et al. (2014) proposed a SCADA-specific IDS for detecting attacks based on network traffic behaviour by evaluating frequent patterns of SCADA protocols.

All these publications have in common that the respective systems are implemented in an immersive way as they have to be integrated into the target environment. This violates the identified requirements of *non-expandability* of ICSs and enables unacceptable system *interference*. Recent patents only use trivial approaches, such as creating a whitelist of all connected devices and afterwards creating alerts based on configuration changes like unseen new IP addresses (Mcquillan and Lloyd 2016). Again, the proposed IDS must be integrated into the SCADA system, which prevents the retrofitting of existing systems.

As our proposed concept is based on the evaluation of current drain to respect these requirements with a more tamper-proof system and using better ground truth, respective publications within the area of power-related intrusion detection will be discussed as follows. The need for identifying and evaluating abnormal electric power consumption arose back in the late 90s, when Stajano (1999) was one of the first researchers describing the problem of battery exhaustion attacks.

Nash et al. (2005) proposed an IDS specialised on the detection of battery exhaustion attacks. Their system evaluates parameters like CPU load and disk access of mobile computing devices. The power consumption is estimated using a linear regression model on a per process basis. Based on these evaluations, potential battery exhaustion attacks are identified. In contrast to our approach, the system cannot provide general intrusion and attack detection.

Jacoby and Davis (2007) proposed a battery-based intrusion detection system (B-BID) for mobile handheld devices. Their system consists of three parts: a host intrusion detection engine provides rule-based detection of battery behaviour anomalies based on static threshold levels; a ‘source port intrusion engine’ is for capturing network packets during suspected attacks and a ‘host analysis signature trace engine’ which is used to correlate signature patterns in the frequency domain. Its shortcomings are the restriction to mobile systems and the requirements of specific preconditions like the evaluation of busy, idle and suspend states. Our concept is not restricted to battery-powered systems nor does it require knowledge about process states.

Srinivasan et al. (2006) proposed a self-organised agent-based architecture for power-aware intrusion detection (SAPID). It uses a power level and a hybrid metric to determine traffic, and a self-organising map to recognise anomalies in the network. In contrast to our approach, SAPID is not generally applicable, focusing on ad-hoc wireless networks.

Buennemeyer et al. (2006) proposed a battery-sensing intrusion detection system (B-SIPS). B-SIPS senses anomalous patterns in the battery current to identify possible exhaustion attacks and malicious activities. A server-based correlation intrusion detection engine is used to correlate possible attacks with a network-based IDS. While this system improves capabilities of battery-based intrusion detection and lowers false alarm rates, it focuses on attacks on Bluetooth and WiFi. Also, the system depends on smart battery monitoring capabilities, prohibiting its general application. In contrast, our concept is not restricted to specific network interfaces and does not require smart battery capabilities.

Stepanova et al. (2010) made a homogeneity analysis of power consumption for information security purposes. Their system has to be trained with multiple battery-lifetime periods, limiting its applicability to battery-powered devices. Also, it is focused on the detection of malicious SMS Trojans and MMS-transmitting net worms. Our concept is not limited to battery-powered systems nor restricted to specific malicious software.

While different battery-based systems have been proposed and traditional IDS had been extended for the evaluation of power-based features, their capabilities and applicability are still limited. A recent approach called power fingerprinting (PFP) is more promising. It extracts 'references from the execution of trusted software and use[s] them to compare captured traces to determine whether the same code is executing' (Reed and Gonzalez 2012). While this can be used to identify malign behaviour using a difficult-to-manipulate measuring base, the requirements for PFP cannot always be satisfied. Particularly in general-purpose systems, getting references from trusted software for all possible programs which may be executed is typically not possible, quickly resulting in high false alarm rates. The situation will also become more challenging for ICS when an increasing amount of individualised mass production is introduced in the context of 'Industry 4.0'. Learning phases can also be used to identify the normal behaviour of the network, but they often cannot see all benign behaviour, and the target environment already can be compromised, resulting in learning malicious behaviour as benign (Koch 2009).

In contrast to the current approaches with limited detection capabilities, we propose a new concept of a current-sensing based, lightweight IDS using cheap single-board computers which can be deployed within an existing network structure of, for example, a SCADA system. It extends PFP mechanisms by introducing specific power patterns, lowering the risk of learning malicious behaviour when no reference from trusted software is available, as well as reducing false alarm rates.

3. LIGHTWEIGHT INTRUSION DETECTION FOR ICSs

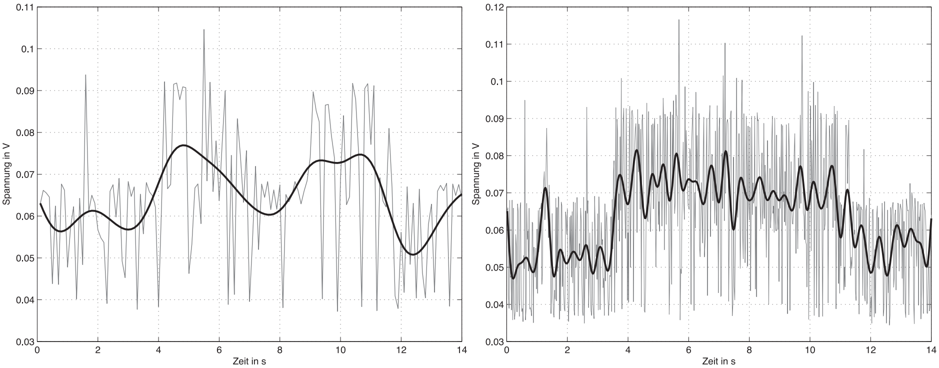
Next, the current-based intrusion detection realised by *Dr.WATTson* is presented and distributed approaches to realise real-world-usable IDSs for ICSs are discussed. We used a design-oriented approach for the development of the concept and the architecture.

A. Current-Based Intrusion Detection by Dr.WATTson

For the design and development of a lightweight current-based IDS, a low-priced but accurate system with respective IO ports is required for measuring and processing. Therefore, available hardware components and their capabilities were analysed to determine their suitability. First, measurements to identify a tamper-resistant detection scheme which can be generally applied by using low-cost equipment were executed. An ODROID-U3 mini-computer was combined with the ODROID Smart Power which collects voltage and current based on a sampling rate of 10 Hz. The ODROID-U3 is now discontinued, but the available ODROID-C2 is even better and will be used for our upcoming setup (see Section 4). By executing different current and voltage measurements and analysing the accuracy of the Smart Power in comparison with the results of a highly accurate oscilloscope, the accuracy of the low-cost setup was verified: all measured values were within the announced deviation of 2%. Next, the required sampling rate was analysed.

A comparative voltage metering was done by using a Keithley Series 2700 (440 Hz) and a NI PCI-2651 high-speed card (up to 2.8 MHz). For the latter, corresponding log files had been generated by an additional PC, as the ODROID does not provide a PCIe interface. The firmware of the Smart Power was adapted to provide headless operation, as the graphical output was not needed but raw measurement values for the calculations was. A resulting effect was a more accurate measurement of the Smart Power, as the GUI is implementing some smoothing and filtering out single peaks, while this data remains available by the recorded logs.

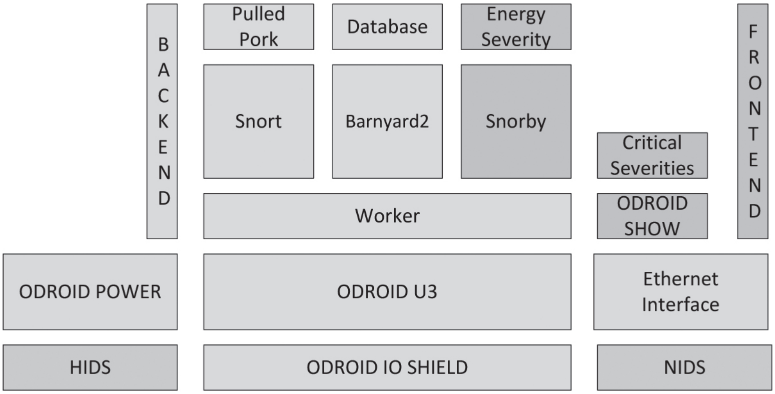
FIGURE 2. VOLTAGE CHARACTERISTICS DURING A PORTSCAN SEEN BY 10 HZ AND 50 HZ SAMPLING RATE. THE RED CURVE IS AFTER FILTERING.



To investigate the sampling frequency for the pattern recognition, multiple test runs were evaluated. Figure 2 shows the voltage characteristics during a port scan seen by sampling rates of 10 and 50Hz. Using a frequency of 10Hz, a clear voltage profile can be recognised. The voltage differed between 60 and 75mV and the end of the scan can be seen around second 11. Having a look at the 50Hz sampling rate, more details can be found with peaks going up to 82mV. A further increase in the sampling frequency sharpens the voltage edges but does not provide new valuable information; starting with a sampling rate of 500kHz, all processes of the switching power supply are recorded together with additional noise: this also hampers the evaluation of patterns. As a result, the 10Hz sampling frequency is enough to recognise attack patterns in the power consumption.

Next, the architecture for current-based intrusion detection was developed as depicted in Figure 3. The *ODROID-U3* is the low-cost hardware platform. The processing of the collected measurement values is done by the *Worker*. There, the cross correlations are calculated, the database containing current flow patterns is administrated, and different modules are controlled. The collection of current measurement values is done by *ODROID Power*; multiple Smart Powers can be connected by USB, or IO SHIELD can be used for the data-collection using 36 additional GPIO ports. This also enables the connection of multiple sensors to one mini-computer. *Snort* is integrated as a traditional NIDS, enabling optional alert correlation (e.g., when the monitored system already has a rule-based IDS and specialised rulesets, e.g., from Digital Bond), while *barnyard2* is used for converting Snorts' output. *Snorby* is implemented as a graphical frontend and *ODROID-SHOW* is used for displaying essential operating parameters and current-based alerts directly at the mini-computer.

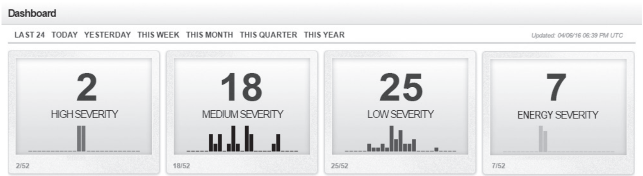
FIGURE 3. ARCHITECTURE OF DR.WATTSON



The described architecture can also be operated as HIDS to perform current-based intrusion detection for itself: the main operational mode is the application as NIDS, providing current-based intrusion detection for one or multiple systems by using current-sensing information, but HIDS and NIDS can be operated simultaneously for a current-sensing IDS with permanent self-monitoring.

The proposed architecture was implemented by the PoC Dr.WATTson, using Ubuntu 14.04.02 LTS for ARM architecture with kernel 3.8.13.30 configured to use *Performance* as CPU governor and Ruby on Rails for the provision of required libraries and the implementation of the Worker. Snort was implemented to verify the compatibility with this widely used IDS, and to provide an additional source for correlation. For storing and visualising of alerts, the *Unified2Binary* data generated by Snort is taken by the open source interpreter barnyard2 and written to disk for further parsing. For visualisation, Snorby was integrated and extended with the new field *Energy Severity*. This additional display represents only alerts which are generated based on current-based detection. Figure 3 shows a screenshot, presenting the newly integrated display.

FIGURE 4. EXTENDED SNORBY INTERFACE WITH NEW ALERT CLASS FOR POWER-BASED ALERTS



Updated: 04/06/16 06:39 PM UTC

2 HIGH SEVERITY 2/52

18 MEDIUM SEVERITY 18/52

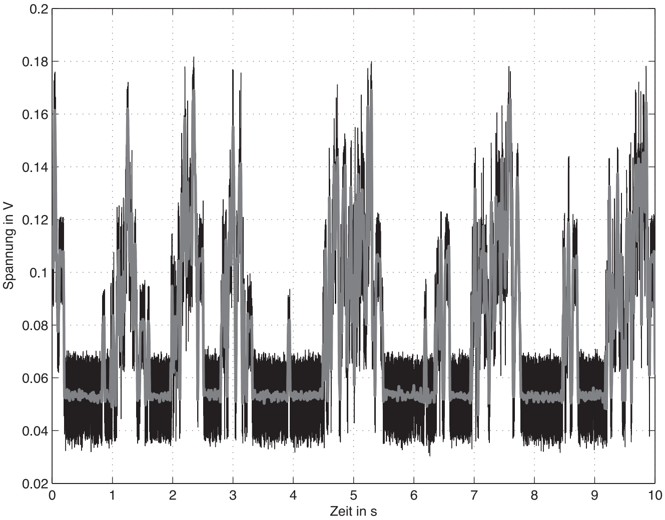
25 LOW SEVERITY 25/52

7 ENERGY SEVERITY 7/52

To exploit that power consumption can be a quite tamper-resistant information source and baseline, first the regular power consumption of the device was evaluated. Reference values were measured using a variety of scenarios to generate comparative values based on this baseline. For systems like ICSs with clearly defined processes, trustworthy comparative data can be generated quite easily by executing multiple measurement cycles. During later operation, measurements have to be compared with the earlier generated current drain values. For the calculation of the similarity, cross correlation functions can be used: a sum function of the cross correlation was implemented to enable a calculation for a discrete base, the recorded measurement values in the logs. Based on these similarity calculations, current-based alerts which are called *Energy Severity Events* are generated.

After the baseline had been established, attack scenarios like DoS, portscans and bruteforce were executed. From the variety of attacks, an SSH bruteforce attack detected by Dr.WATTson is shown in Figure 5. The time frame of ten seconds presents a triangular voltage pattern with steep flanks, which is a typical and explicit power pattern identified for SSH attacks. The measured value series was correlated by the Worker with comparative data from the database to identify possible attacks. With this, a reliable detection of attacks was possible.

FIGURE 5. MEASUREMENT OF DR.WATTSON DURING SSH BRUTEFORCE ATTACK



For the evaluation of Dr.WATTson, the ground truth was generated as previously discussed based on a run of 72 hours. After that, different attack scenarios were executed, each lasting 14 hours and containing 11 discrete attacks per hour. The attack runs were repeated five times to calculate the detection results. The system was able to detect 100% of the executed attacks while delivering a false alarm rate of 0.13%, surpassing the results of other systems. The classification of the resp. attack type, which is a new feature other systems are not able to provide, was correct in 45.5% of cases. The final detection results are summarised in Table 2. Usual research in the area of power-based intrusion detection typically focuses on specific wireless networks, hampering a comparison of detection results (see, for example, Jacoby and Davis 2007), and does not provide detection and false alarm rates because of the focus on battery exhaustion. Even SAPID is limited to wireless ad-hoc networks, in contrast to the generally applicable Dr.WATTson, while OCSVM and Sayegh cannot be used for retrofitting ICSs.

Also note that the 100% detection rate is not based on an overfitting of the system, but on the clear distinction between normal and malign current patterns. Such detection rates are only reachable in such ICS scenarios, having a limited number of well-defined processes, but not in general purpose systems like a desktop PC browsing in the Internet. Real-world applications are more challenging and may generate more noise and therefore higher false alarm rates. The results achieved were even better than hoped for, providing a promising base for the distributed application of Dr.WATTson in a real SCADA system.

Having a look at the security of the architecture itself, it does not introduce new attack vectors as it is only using voltage levels and current drains collected by sensory which is not intervening into the monitored system. The self-monitoring capability of Dr.WATTson also hampers physical manipulation of the system itself.

TABLE 2. EVALUATION RESULTS OF DR.WATTSON AND COMPARISON WITH OTHER SYSTEMS

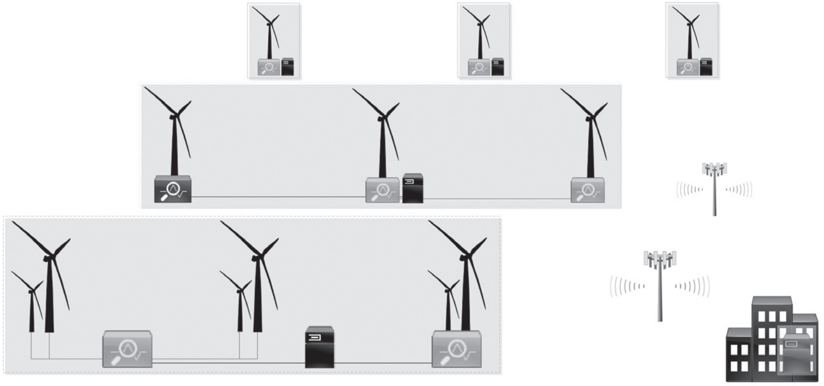
System	Detection Rate	False Alarm Rate	Attack Classification
Dr.WATTson	100	0.13	45.5
SAPID (Srinivasan et al.)	98.0	3.0	-
OCSVM (Maglaras et al, 2014)	96.3	2.5	-
Anomaly-based Intrusion Detection System (Sayegh et al. 2014)	89.9	1.3	-

B. Distributed Intrusion Detection by Dr.WATTson

Based on the detection capabilities of Dr.WATTson, multiple distributed setups for ICSs can be designed. Figure 6 presents a respective SCADA scenario of a wind energy park, consisting of multiple, dislodged generators. The possible visualised measurement setups are:

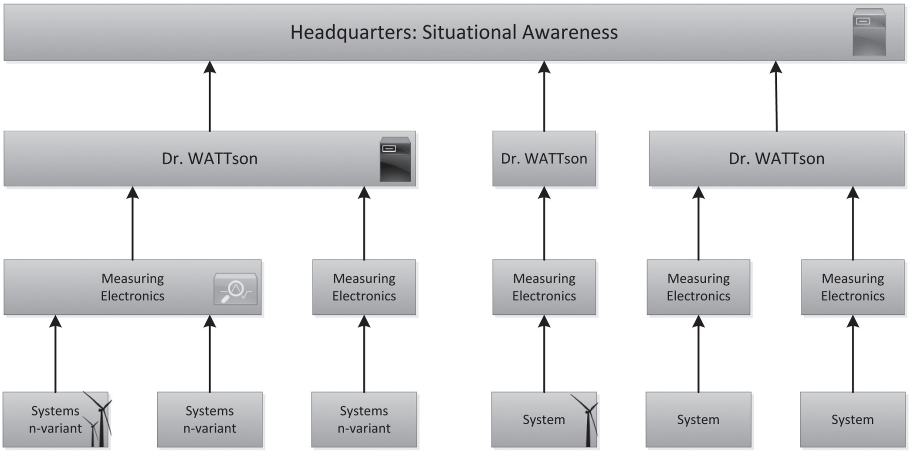
- N-variant systems, centrally measured (first generator line in Figure 6);
- Single systems, decentralised measured (second and third generator line); and, for both cases the possibilities:
- Centrally evaluated (first and second generator line); or
- Decentralised evaluated (third generator line).

FIGURE 6. SCENARIO OF A WIND PARK WITH THREE LINES OF GENERATORS



Note that for a wind park, as each wind generator is a high-value asset, the setup 'single systems, decentralised measured, decentralised evaluated' with providing a consolidate picture for generating the situational awareness at the headquarters would be preferable, while the ICS of a factory with a multitude of components can be likely monitored by a centrally measured, n-variant setup. While centralised measurements of multiple systems saves hardware, the disadvantage is that the 1:1 correlation gets lost: an alert can only be assigned to the measuring instance, not the precise end system. By realising decentralised measurements, the quantity of required hardware is the most, but malign behaviour can be detected and assigned rapidly. The selection of centralised or decentralised evaluation depends on the size of the system to be monitored and the company structure. In all configurations, it is possible to generate a situational awareness picture based on the individual Dr.WATTson instances (see Figure 7).

FIGURE 7. SETUPS FOR DISTRIBUTED INTRUSION DETECTION BY DR.WATTSON



4. OUTLOOK

Intrusion Detection in ICSs and complex SCADA systems is challenging, as several particularities apply. While SCADA systems are increasingly endangered, their security remains inadequate and reports are alarming. Major challenges arise as state-of-the-art IDSs are not able to cope with the special requirements of ICSs and control systems often cannot be retrofitted.

To overcome these shortcomings, we present a new intrusion detection architecture based on low-cost mini-computers which evaluate current drain measurements to achieve intrusion detection. In contrast to other approaches, our system can be used for retrofitting even non-expandable control systems.

Based on the promising results of the prototype Dr.WATTson, we designed a distributed IDS for SCADA systems. Next, a comprehensive test and an evaluation within a productive environment will be done, where we deploy the distributed Dr.WATTson, using an ODROID-C2 hardware base and a new GPU-based pattern evaluation. As the C2 supports GPIOs without additional IO-Shield, system performance is increased but low-priced components are still used. At the moment, we are talking with an energy provider about realising this evaluation of Dr.WATTson within a live SCADA system.

ACKNOWLEDGMENT

This work was partly funded by FLAMINGO, a Network of Excellence project (ICT-318488) supported by the European Commission under its Seventh Framework Programme.

REFERENCES

Andreeva, O., Gordeychik, S., Gritsai, G., Kochetova, O., Potseluevskaya, E., Sidorov, S. I. & Timorin, A. A. (2016). *Industrial Control Systems Vulnerabilities Statistics*. Kaspersky Lab.

Baishakhi, R., Posnett, D., Filkov, V. & Devanbu, P. (2014). A large-scale study of programming languages and code quality in github. *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM New York.

Bodenheim, R. C. (2014). *Impact of the Shodan Computer Search Engine on Internet-facing Industrial Control System Devices*. Thesis, AFIT-ENG-14-M-14. Department of the Air Force Air University. Air Force Institute of Technology.

Buennemeyer, T., Jacoby, G., Chiang, W., Marchany, R. & Tront, J. (2006). Battery-sensing intrusion protection system. *Information Assurance Workshop*, pp. 176-183. IEEE.

Carlson, R. (1998). *Towards a Standard for Highly Secure SCADA Systems. Report SAND98-2220C, Sandia National Laboratories, Albuquerque, NM, and Livermore, CA*. Sandia Corporation.

Cavusoglu, H., Cavusoglu, H. & Zhang, J. (2008). ‘Security patch management: share the burden or share the damage?’ Management Science, 4, pp. 657-670.

Childs, D. (2015) *HP Security Briefing, Episode 22: The hidden dangers of inadequate patching strategies*. April 06. Retrieved January 5, 2017 from https://community.hpe.com/t5/Security-Research/HP-Security-Briefing-Episode-22-The-hidden-dangers-of-inadequate/ba-p/6752022#.WG-1be17a3B.

GLEG Ltd. (2015). *GLEG - information security company. Agora exploit pack developer*. Retrieved January 5, 2017 from http://gleg.net/agora_scada_upd.shtml/.

Guri, M., Solewicz, Y. A., Daidakulov, A. & Elovici, Y. (2016). *DiskFiltration: Data Exfiltration from Speakerless Air-Gapped Computers via Covert Hard Drive Noise.* Ben-Gurion University of the Negev.

Harp, D., & Gregory-Brown, B. (2016). *SANS 2016 State of ICS Security Survey*. SANS Institute.

Idaho National Laboratory. (2008). *Control Systems Cyber Security: Defense in Depth Strategies*. Department of Homeland Security.

Jacoby, G. & Davis, N. (2007). Mobile host-based intrusion detection and attack identification. *IEEE Wireless Communications*, vol. 14, no. 4, pp. 53-60.

Juniper Research. (2015, May 12). 'Cybercrime will Cost Business Over $2 Trillion by 2019'. Retrieved July 6, 2016, from http://www.juniperresearch.com/press/press-releases/cybercrime-cost-businesses-over-2trillion.

Koch, R. (2009). 'Changing Network Behavior'. *Third International Conference on Network and System Security*. IEEE.

Le, X.-B. D. (2016). *Towards Efficient and Effective Automatic Program Repair*. Singapore Management University, School of Information Systems.

Lee, R. M., Assante, M. J. & Conway, T. (2016). *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Electricity Information Sharing and Analysis Center. Washington DC: SANS Institute.

Maglaras, L. A., Jiang, J. & Cruz, T. (2014). Integrated OCSVM mechanism for intrusion detection in SCADA systems. *Electronics Letters* vol. 50 no. 25, pp. 1935-1936.

McAfee. (2014). *Net Losses: Estimating the Global Cost of Cybercrime*. Santa Clara: Intel Security.

Mcquillan, J. L. & Lloyd, C. A. (2016). SCADA Intrusion Detection Systems, *Publication Number US20160094578 A1, PAT. Application Number US 14/501,672*. Schneider Electric USA, Inc.

Mimoso, M. (2015). 'Creaking Patch Tuesday's Viability Rests with Quality, Speed'. Retrieved January 5, 2017 from threatpost.com: https://threatpost.com/creaking-patch-tuesdays-viability-rests-with-quality-speed/110941/.

Mitchell, R. & Chen, I. (2014). 'A survey of intrusion detection techniques for cyber-physical systems'. *ACM Computing Surveys (CSUR)*, vol. 46, no. 4.

Nash, D., Martin, T., Ha, D. & Hsiao, M. (2005).' Towards an intrusion detection system for battery exhaustion attacks on mobile computing devices'. *Third IEEE International Conference on Pervasive Computing and Communications Workshops*, pp. 141-145.

NATO. (2016). 'NATO: Cyber defence'. Retrieved January 8, 2017 from http://www.nato.int/cps/en/natohq/topics_78170.htm.

Ormandy, T. (28. June 2016). 'How to Compromise the Enterprise Endpoint'. Retrieved July 14, 2016 from http://googleprojectzero.blogspot.de/2016/06/how-to-compromise-enterprise-endpoint.html.

Paganini, P. (2014). 'InfoSec Resources - Foreign Hackers Constantly Target US Critical Infrastructure'. 24 November. Retrieved January 7, 2017 from http://resources.infosecinstitute.com/foreign-hackers-constantly-target-us-critical-infrastructure/#gref.

Panko, R. (2008). ‘Error Rates in Programming’. Retrieved July 14, 2016 from http://panko.shidler.hawaii.edu/HumanErr/Index.htm.

Reed, J. H. & Gonzalez, C. R. (2012). ‘Enhancing Smart Grid Cyber Security using Power Fingerprinting’. *Future of Instrumentation International Workshop (FIIW)*. IEEE.

REUTERS. (2016). ‘Massive cyber attack could trigger NATO response: Stoltenberg’. June 16. Retrieved 7 January 2017, from http://www.reuters.com/article/us-cyber-nato-idUSKCN0Z12NE.

Saglam, M. (2014). *A Military Planning Methodology for Conducting Cyber Attacks on Power Grid*. Master Thesis, Virginia Polytechnic Institute and State University, Falls Church, Virginia.

Sayegh, N., Elhajj, I. H., Kayssi, A. & Chehab, A. (2014). ‘SCADA Intrusion Detection System based on temporal behavior of frequent patterns’. *2014 17th IEEE Mediterranean Electrotechnical Conference (MELECON 2014)*, pp. 432-438.

Srinivasan, T., Vijaykumar, V. & Chandrasekar, R. (2006). ‘A self-organized agent-based architecture for power-aware intrusion detection in wireless’. *International Conference on Computing & Informatics*, pp. 1-6.

Stajano, F. (1999). ‘The resurrecting duckling’. *Security Protocols*, pp. 183-194.

Stepanova, T., Kalinin, M., Baranov, P. & Zegzhda, D. (2010). ‘Homogeneity analysis of power consumption for information security purposes’. *Proceedings of the 3rd International Conference on Security of Information and Networks, ser. SIN ‘10*, pp. 113-117.

Wilson, C. (2012). *Industrial and SCADA Systems May Be Increasingly Targeted for Cyberattack*. University of Maryland University College.

Wolff, J. (2016). ‘Perverse Effects in Defense of Computer Systems: When More is Less’. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 4823-4831.

Yang, D., Usynin, A. & Hines, J. W. (2006). ‘Anomaly-based intrusion detection for SCADA systems’. *5th intl. Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies*, pp. 12-16.

Yang, Y., McLaughlin, K., Sezer, S., Littler, T., Im, E., Pranggono, B. & Wang, H. F. (2014). ‘Multiattribute SCADA-Specific Intrusion Detection System for Power Networks’. *IEEE Transactions on Power Delivery*, vol. 29, no. 3, pp. 1092-1102.

Zhu, B. & Sastry, S. (2010). ‘SCADA-specific intrusion detection/prevention systems: a survey and taxonomy’. *Proceedings of the 1st Workshop on Secure Control Systems (SCS)*.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Crowdsourcing Security for Wireless Air Traffic Communications

Martin Strohmeier
Department of Computer Science
University of Oxford
Oxford, United Kingdom
martin.strohmeier@cs.ox.ac.uk

Matt Smith
Department of Computer Science
University of Oxford
Oxford, United Kingdom
matthew.smith@cs.ox.ac.uk

Matthias Schäfer
Department of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
schaefer@cs.uni-kl.de

Vincent Lenders
Cyberspace and Information
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Ivan Martinovic
Department of Computer Science
University of Oxford
Oxford, United Kingdom
ivan.martinovic@cs.ox.ac.uk

Abstract: Protecting the security of the cyber-physical systems that make up the world's critical infrastructures has been a recent hotly debated topic. Legacy wireless communication infrastructure is often an impediment to quickly improving these crucial systems, as cryptographic solutions prove impossible to deploy. In this article, we propose the establishment of a separate verification layer for sensitive wireless data powered by crowdsourced sensors connected to the Internet and apply it to the aviation domain.

We first validate the need for independent data verification in air traffic control networks, where all wireless communication is conducted in the clear and thus subject to manipulation. To counter this threat, we develop a comprehensive model for the verification of wireless communication based on massively distributed data collection and outline how it can be used to immediately improve the security of unprotected air traffic control networks.

By combining several different methods based on the content and the physical characteristics of aircraft signals, our system is able to detect typical injection, modification and jamming attacks. We further develop a trust model to defend against potential insider threats based on compromised sensors.

We illustrate our approach using the crowdsourced sensor network OpenSky, which captures large parts of civil air traffic communication around the globe. We analyse the security of our approach and show that it can quickly, cheaply, and effectively defend against even sophisticated attacks.

Keywords: *critical infrastructures, wireless security, air traffic control, crowdsourcing, sensor networks, participatory sensing*

1. INTRODUCTION

As modern computer networks become increasingly important for the safe facilitation of travel on the ground, sea and in the air, their cyber security must be ensured. In recent years both the academic community and hacker circles have shown that many vulnerabilities exist in the cyber-physical networks used by core critical infrastructures such as air traffic control (ATC) or vessel traffic services [1]–[3]. Numerous reasons prevent the quick and effective replacement of existing technologies with new and secure ones; because of this, novel methods of protection are required.

In this paper, we propose the crowdsourcing of security to protect core critical infrastructures without the lengthy and costly deployment of new technologies. We present the crowdsourced sensor network OpenSky and outline how it can be used to immediately improve the security of ATC networks.

The idea of crowdsourcing has previously been applied to attempt to solve many large-scale scientific problems such as protein folding [4] and classification of galaxies [5]. Recently, numerous private companies have started to use crowdsourced networks to track the locations of ships and aircraft around the globe. As the networks' data collection infrastructure grows, their services are increasingly requested by large industry players and authorities [6].

Thus, it naturally follows to exploit the data and intelligence gathered by such crowdsourced networks to secure the cyber-physical systems of critical infrastructures. In this work, we present an approach to delivering secure civil ATC using OpenSky, a sensor network for ATC communication consisting of more than 300 crowdsourced sensors, which receive over 100,000 transponder signals per second providing coverage across all continents [7], [8].

Our contributions are:

- We explain how crowdsourced networks can be exploited to act as independent witnesses of attacks such as jamming and spoofing, and to provide a trusted third party opinion.
- We present a system that is able to detect data maliciously injected into the wireless channels used for ATC. By employing several independent verification approaches, we can detect even sophisticated attackers and present a constantly validated picture of the airspace.
- We further analyse the security challenges that arise if malicious sensor owners attempt to compromise the system. We describe the trust models used for newly-connected sensors and how to prevent insider attacks.

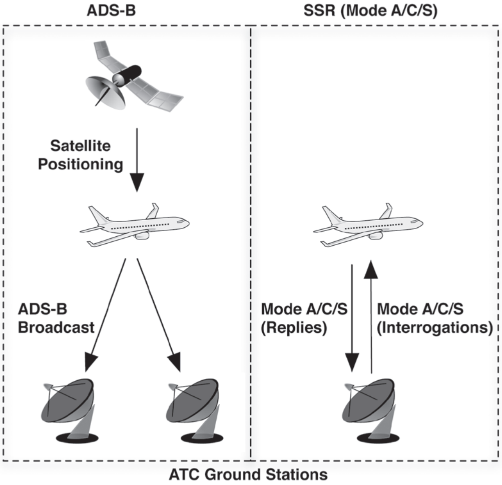
In the remainder of this work, we briefly describe the ATC technologies and their vulnerabilities in Section 2, before considering the related work in Section 3. Section 4 describes the crowdsourced system and the algorithms to detect potential attacks. Section 5 introduces our threat model, while Section 6 explains our countermeasures to wireless ATC threats. Section 7 analyses the security of our system. Section 8 discusses wider application, extensions, and cost. Finally, Section 9 concludes this paper.

2. WIRELESS THREATS TO ATC

With the rise of terrorist and organised crime groups interested in cyber attacks on ATC targets, a wide range of vulnerabilities and cyber attack vectors threaten the aviation infrastructure. Such vectors include malware in ATC networks and computers [9] and direct attacks on aircraft onboard networks [10].

In this work, we focus on the attack vectors provided by wireless ATC technologies, which are all inherently insecure by design [1], [2], [11], [12]. We briefly explain the function of two such technologies and the vulnerabilities that we seek to protect against by using a crowdsourced security model. While they act as technical proof-of-concept, the concepts in this paper can be adapted to other wireless technologies used in ATC, such as those used for data links, navigation, or collision avoidance.

FIGURE 1. REPRESENTATION OF ADS-B AND MODE S SYSTEMS



A. Modern ATC Technologies

Secondary Surveillance Radar (SSR) is a cooperative ATC technology currently based on the so-called transponder Modes A, C, and S, which provide digital target information compared to traditional analogue primary radar (PSR) [13]. Aircraft transponders are interrogated on the 1030MHz frequency and reply with the desired information on the 1090MHz channel, as shown in Figure 1.

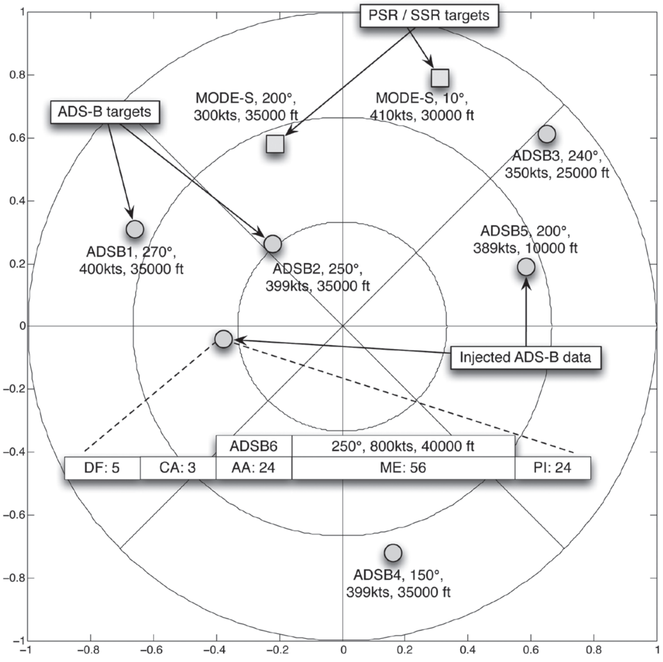
With the newer Automatic Dependent Surveillance-Broadcast (ADS-B) protocol (see Figure 1), aircraft regularly broadcast their own identity, position, velocity and additional information such as intent, status, or emergency codes. These broadcasts do not require interrogation: position and velocity are automatically transmitted twice a second [14].

Unfortunately, security was not part of the design of these systems; neither includes cryptographic protection, which could provide confidentiality, integrity, or authentication for ATC users. Consequently, it is entirely possible to modify any airspace picture based on these technologies, as elaborated in the next section.

B. Wireless Attack Vectors against ATC Infrastructure

It has been shown that commodity hardware can receive and transmit on the frequencies used by ADS-B and SSR, making them accessible even for unsophisticated threat actors [11]. We consider the following common wireless attack vectors against an ATC receiver system providing data to a radar screen:

FIGURE 2. ILLUSTRATION OF AN ATC SCREEN WITH INJECTED ADS-B DATA



1) Injections of Ghost Aircraft

The first attack injects a new ghost aircraft created from scratch, by creating correctly formatted SSR/ADS-B. We assume that the attacker crafts transponder signals with a legitimate identifier and reasonable flight parameters (e.g., plausible altitude and speed) to create an aircraft that is indistinguishable from a legitimate one for radars relying solely on these technologies (see Figure 2 for a high-level illustration). The requirement to handle this aircraft – or many such aircraft – creates significant additional workload for the controller, which can lead to a loss of situational awareness and decrease the safety of the airspace [12]. Consequences may include the requirement to go procedural and use analogue voice technology for communication and aircraft separation, causing significant decrease in air traffic throughput, or even accidents in the worst case [12].

2) Modification of Labels

More subtle is the modification of data sent by real aircraft and seen on the labels on controllers' radar screens. Among other things, these labels include identity, altitude, velocity, heading, and the flight level the aircraft intends to use. It is obvious that any modification of the wireless messages that provide the data for the display can have severe consequences for the safe control of the airspace [12].

3) Jamming or Denial of Service

Lastly, we consider jamming attacks, both targeted and unspecific. In an unspecific attack, the frequency is jammed with noise at the receiver, effectively causing a denial of service (DoS). Targeted attacks destroy only specific packets (e.g. for a single aircraft), which can be done even reactively as the packet is on the wireless channel [15]. Jamming ATC technologies can have severe consequences for the situational awareness of pilots and controllers [12].

3. RELATED WORK

We briefly analyse the state-of-the-art security for wireless ATC technologies before discussing the related work on crowdsourced security in other fields.

A. Wireless Security Approaches for ATC

Several academic works and presentations at hacker conferences in recent years have highlighted ATC vulnerabilities and inspired urgently needed research on potential security measures. Generally, the most promising approaches can be divided into two different avenues, depending on whether or not they use cryptography.

Using cryptography enables authentication, integrity and confidentiality for all ATC messages and is thus the preferable option to defeat message injections and modifications. However, due to real-world constraints, both technological and regulatory, any advances in this direction will not see widespread deployment for many years [11].

As a result, much research conducted to secure ADS-B and SSR today has considered the application of *transparent* security measures, which do not require changes to the protocols or certified transponder hardware installed in aircraft and ground stations around the world. These typically focus on non-cryptographic schemes that exploit physical characteristics of the wireless signals broadcast by the aircraft, including ADS-B/SSR. Examples of such characteristics include the time differences of arrival [16], [17], Doppler shift [18], or strength [19] of the received transponder signals.

B. Crowdsourcing Security using Participatory Sensing

Governments and other institutions are no strangers to the idea of exploiting crowdsourced sensor data to improve the security of various systems. In particular, the ubiquity of smartphones as sensing devices has inspired the popular research area of participatory surveillance [20]. For example, the authors in [21] explore the Department of Homeland Security's Cell-All project, which seeks to equip mobile phones with chemical-agent detectors to detect potentially dangerous volatile chemical compounds.

Similar concepts have been developed for many other security-related domains, such as nuclear disaster recovery or emergency evacuations [22]. However, to the best of our knowledge the present work is the first to propose crowdsourced sensing of unprotected critical infrastructure communication in order to develop an attack detection system for civil ATC.

4. CROWDSOURCING IN AVIATION

In this section, we discuss the motivations and advantages of a separate crowdsourced sensor network used to verify wireless information in critical infrastructures. As a proof-of-concept we introduce the sensor network OpenSky, its existing capabilities, and how we can exploit them to provide a secure radar picture of the civil airspace.

A. Motivation

Crowdsourced networks provide many benefits to aviation users, from airlines and airports to passengers. Airlines and airports use the services offered by several large companies such as Flightradar24 [23] or FlightAware [24], which exploit crowdsourcing to track the movements of all transponder-equipped aircraft around the globe, both in the past and live. Similarly, private end users use the displayed data to track, for example, the punctuality of particular flights relevant to them or simply enjoy following flights in their region for a host of other reasons [25].

Flightradar24 states that their network alone includes more than 10,000 receivers [26]; as the number of such tracking services based on freely available, volunteer-provided ATC communication grows, they illustrate the potential of crowdsourced sensing. For our work, we focus on the utility of these networks for security purposes.

Our goal is to provide a cheap yet reliable and powerful second view of the airspace, acting as a witness and detecting potential attacks on wireless ATC protocols. In a second step, if possible, false data injected by an attacker is filtered out before it is displayed, providing a clean backup view for verification and validation purposes. Finally, it is desirable to narrow down the origin of any such attack to be able to initiate physical protection measures.

B. Advantages of Crowdsourced Networks

In our proposal, we first consider the viewpoint of air navigation service providers who use potentially vulnerable SSR/ADS-B systems. As safety demands in aviation require long, expensive certification processes, such providers cannot quickly upgrade their infrastructure, which explains why ADS-B has been in development since the late 1990s [27]. Implementing new security features into already deployed technologies is also impractical, as it requires an overhaul of all airspace participants [28].

Consequently, we propose an uncertified system based on crowdsourcing to independently verify received ATC signals and validate the information displayed on certified radar systems. Analogous to intrusion detection systems (IDS) in classical computer networks, it collects the available air traffic communication in its coverage range in a distributed fashion, analysing it for potential wireless attacks.

A crowdsourced network of cheap ATC receivers is low-cost, agile and easily scalable. Many compatible receivers already exist around the world, powered by volunteers who provide the view of their local airspace to several crowdsourced networks simultaneously. While it would be feasible to exclusively use self-deployed ATC sensors for a smaller airspace, the use of crowdsourcing offers low barriers to entry, much lower cost and superior scalability for potentially global reach. Using this wealth of data, it is possible to integrate various different data verification methods; and we can directly deploy security software, which does not need to fulfil the strict safety requirements applied in aviation.

In terms of security properties, the massively redundant, distributed and fluid nature of crowdsourced networks provides the key advantage over certified systems with comparatively few expensive sensors in fixed, long-term locations. For example, a single SSR ground sensor is typically responsible for a radius of 200NM [29]. It is thus comparatively easy to mount a jamming attack on such a sensor, which would cause a denial of service in its airspace.

In contrast, the exact number and location of the many active sensors in a crowdsourced network is unknown at any given time. The high sensor density directly improves system resilience, as there are other receivers with overlapping coverage not impacted by an attack.

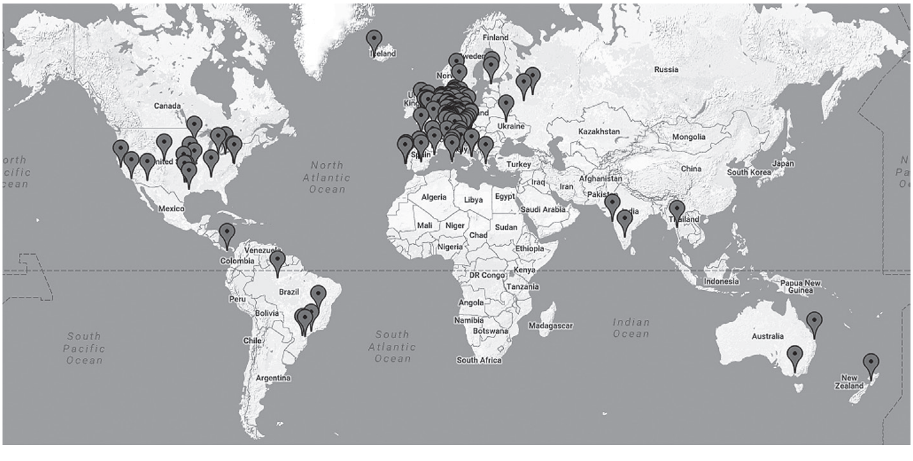
C. The OpenSky Network

OpenSky is a crowdsourced network used as proof-of-concept for our security approach. In the following we discuss its current deployment and trust model for the integration of new sensors.

1) Current Deployment

As of April 2017, the OpenSky Network consists of 260 registered and 300-450 anonymous sensors streaming data to its servers. Registered sensors are those operated by active members of the OpenSky Network community. Their operators are usually known to the administrators, either because the sensor was provided by the network itself or through personal contact. In contrast, the operators of anonymous sensors are unknown. While the exact locations of anonymous sensors are also unknown, they can be approximated based on their coverage and IP address. The locations of registered sensors are known with an accuracy of 10 meters.

FIGURE 3. A MAP OF SENSORS REGISTERED TO THE OPENSKY NETWORK (MARCH 2017)



2) Trust Model

In general, we consider data from a particular sensor trustworthy if the operator is considered trustworthy. However, the crowdsourcing paradigm prevents the use of classical identity verification methods to build trust in operators. Encryption and authentication are not currently implemented by any of the popular ATC receivers that are used for feeding data flows to the OpenSky Network (e.g., dump1090 [30]). Consequently, securing these flows directly would require the use of a non-standard way of feeding, which would ultimately severely hamper the growth of the network.

The success of a crowdsourced network greatly depends on the simplicity of joining the crowd. Complex approaches such as mandatory passport verification of the operator's identity or setting up PGP keys would discourage operators from feeding data. For this reason, building trust in registered operators is achieved through personal long-term relationships and constant communications with other trusted operators and network administrators.

This method, however, cannot be applied to anonymous sensors, where, except for the exchange of sensor data, there is no communication between the network and the operator. Our method to build trust in these sensors is therefore data-based and trust is constantly re-evaluated. Specifically, we consider an anonymous sensor trustworthy if a considerable fraction of its data can continuously be confirmed by an existing trusted sensor with overlapping coverage. If the number of mismatches in the data between a trusted and an anonymous sensor exceeds a certain threshold, the anonymous sensor is considered untrustworthy and its data is ignored by the network. Anonymous sensors with no common coverage with trusted sensors cannot be verified and their data is therefore considered untrustworthy. However, as Figure 3 shows, there are registered sensors in several parts of the world and the trust of most anonymous sensors in the network can be assessed through the transitivity of our approach.

5. THREAT MODEL

We consider two active attackers from a hobbyist to cyber-criminal level as described in [11]. One is an outsider threat, and has the ability to inject, modify or jam transponder signals. The other is an insider threat, and joins the crowdsourcing scheme and attempts to attack it under the guise of being a participant.

We consider the outsider to be moderately-resourced, and capable of performing primarily ground-based attacks. We assume they use off-the-shelf SDRs with readily available transmission equipment. They are capable of attacking from several different positions at once, but without perfect synchronisation.

The insider attacker is also moderately-resourced, and capable of running a number of data feeds at once into the crowdsourcing system. Given the amount of data generated by Mode S, this would be achievable with off-the-shelf computing equipment, though specialist knowledge would be required to synthesise data in such a way that it appears realistic.

The reliable and consistent injection of several matching ATC data feeds with correct timestamps over a long time span (to obtain the necessary trust) exceeds the typical level of sophistication of hobbyists, but it is possible for well-resourced cyber-criminal operations. State actors, which are able to attack network points and modify traffic on the fly, are considered out of our scope.

6. CROWDSOURCED SECURITY METHODS FOR ATC

This section presents a crowdsourced security system based on OpenSky. We discuss several independent approaches based on data obtained by a crowdsourced network to detect attacks on ATC protocols, as considered in Section 2.B. We also describe the procedures taken after an attack is detected.

A. Crowdsourced Attack Detection

We use four different detection methods, with varying levels of complexity and sensor requirements summarised in Table 1. The employed methods range from simple plausibility checks based on the content of the received surveillance data to more complex statistical and cyber-physical analysis.

TABLE 1: OVERVIEW OF CROWDSOURCED ATTACK DETECTION METHODS

Method	Number of Receivers	Complexity	Attack Localisation
Plausibility Checks	1	Low	No
Cross Referencing	2 or more	Low	No
Multilateration	3 or more	High	Yes
Statistical Analysis	2 or more	Moderate	No

1) Plausibility Checks

Many attacks can be detected by running comparatively simple checks on the plausibility of the inputs, a method also used by professional radar systems [31]. As a first line of defence, the network uses rules to detect modified flight data. The first set of rules pertains to the communication of an aircraft:

- The position claimed by an aircraft is outside the known recorded range of the receiving sensor, with a safety margin of 50 km;
- An aircraft suddenly appears well within the communication range of a receiver; or
- The required message types (in particular, position, velocity and identity) are not following the technical standards in terms of frequency and order.

The second set of rules is concerned with the technical capabilities of an aircraft, for example:

- The velocity or altitude of an aircraft is outside the possible parameters for the particular class and model;

- The reported velocity of an aircraft and the same aircraft's velocity as derived from its positional data are different (outside of a safety margin of 50 km/h); or
- An aircraft reports an aircraft class/model or capabilities different from its official registration of the network's aircraft database.

2) Cross Referencing of Data between Sensors

If we have a redundant coverage of sensors in a given region, we can cross-reference their knowledge about the received aircraft messages. This concerns both the content and the physical characteristics of the messages.

For example, if an aircraft claims to be in an area that is covered by three sensors, when only one of these sensors receives the aircraft's messages, there would be cause for concern. As there is significant frequency overuse on the 1090MHz channel, which causes message loss of 50% or more [32], a single message failing such a cross-reference check is too common to be noteworthy. A sequence of several missed messages which, according to the positional claims are well within a sensor's coverage area, should however be treated as highly suspicious. Similarly, different sensors each receiving different message content from the same aircraft should raise immediate concerns, barring any transponder or decoding errors. Any such instance may indicate a typical message injection, where the threat agent is not close to the claimed position but instead attacks a single ATC receiver location, for example at an airport.

3) Multilateration and Aircraft Localisation

Independent localisation of aircraft using the physical characteristics of their communication signals is a popular approach in civil aviation. In areas with sufficient sensor coverage, we can calculate the origin of a signal based on the time differences of arrival at three or more receivers and thus verify an aircraft's location claim. This technique, called multilateration, is used in many modern airspace surveillance systems, but comes at significant costs in installation and maintenance. OpenSky is capable of conducting independent localisation based on the time difference of arrival (TDOA) even with cheap off-the-shelf sensors [33]. While certified systems guarantee higher accuracy and reliability, our crowdsourced approach is sufficient to detect the origins of SSR/ADS-B signals and can easily verify the positions of all civil aircraft. While the requirements for the successful localisation of aircraft signals are high compared to other methods, there is another significant advantage: localising the actual origin of an ATC message will pinpoint the location of the adversary in the event of an attack.

4) Statistical Analysis

For areas covered by fewer than three sensors, we can still exploit the time differences of arrival to build a statistical attack detection system based on hard-to-forge physical layer characteristics. By applying hypothesis testing, we can detect potential attackers quickly, even with only two sensors.

We first learn the distribution of errors between the expected and actual time differences between our sensors. In the attack detection phase, we use the non-parametric Wilcoxon rank-sum test to check if the received sample distribution matches the expected distribution. By establishing the proximity to the expected data distribution, we can validate the sender.

B. Attack Handling with OpenSky

If an aircraft track violates one or more of the expectations set out by the detection methods, it is a strong indication of an anomaly, whether deliberately induced or not. To avoid false positives, which strongly detract from the practicality of the system in real-world environments, we combine the knowledge of all available methods, depending on the quality of the sensor data.

If the occurrence is indeed deemed an attack, there are several potential consequences:

- The concerned aircraft tracks are flagged as unreliable, requiring separate handling from the affected controllers;
- The concerned messages are dropped and their content disregarded for the aircraft display on OpenSky's radar view, which is preferable in case of label manipulation or of denial-of-service attacks, whereby the radar screen is flooded with ghost aircraft beyond the controller's handling ability; or
- In cases where sufficient receiver data is available, we can use localisation techniques to narrow down the origin of the attack and follow up with physical containment procedures.

7. SECURITY ANALYSIS

In this section, we analyse the potential attacks on a crowdsourced system. We divide these into two categories: insider and outsider attacks. Outsider attacks address issues presented in Section 2, namely the main wireless attack avenues for ATC systems. Insider attacks consider adversaries who attempt to attack the system by subverting the crowdsourced network, here OpenSky. This may provide cover or diversion for other attacks.

A. Outsider Attacks on Real-World ATC

We consider the detection of jamming, injection, and modification attacks separately. Depending on the sophistication of the attacker, all attacks can take different forms [11], as explained in our threat model, we consider attackers below the nation state level, which are moderately resourced and are able to operate from a single location or multiple locations at the same time.

Throughout this section, we consider n sensors (collectively referred to as S), which at least partially cover some area X. This area is primarily covered by sensor S_A, a sensor under attack. G is the set of aircraft currently observable over area X for sensors $S\backslash S_A$, with G_{S_A} being the set of aircraft observable by S_A.

1) Jamming Attacks

We distinguish two different cases of jamming attacks: indiscriminate broadband jamming of the 1090MHz channel leading to the disappearance of all aircraft from radar screens, and targeted jamming of a particular aircraft.

Detecting an indiscriminate jam on a sensor S_A is relatively straightforward as the affected sensor will remain online but cease to report aircraft, which other overlapping sensors may be reporting. Similarly, detection of a targeted jam relies on other sensors with overlapping coverage receiving the data correctly. By cross-referencing the set G with G_{S_A}, we can identify any aircraft not detected by S_A.

Ultimately, a geographically distributed crowdsourced sensor network limits the potential impact of both. An attacker would have to be sufficiently close to several sensors in order to jam each of their signals. In the case of jamming particular aircraft, time synchronicity of jamming signals at separate sites would have to be tight and require significant resources.

2) Injection Attacks

This section considers an attacker who constructs SSR messages in order to either create a non-existent aircraft or impersonate one not currently in reception range, referred to as a ghost aircraft in [1], [34], [35]. We analyse two cases attempting to inject aircraft c: single and multiple location attackers. A single location attacker transmits at a power which either only reaches one sensor or reaches several. If it only reaches a single sensor, $\nexists c : c \in G \cap G_{S_A}$ thus indicating the anomalous aircraft. If messages reach some s in S but not all, comparing each G_S will identify anomalous aircraft. Further steps to identify it as such may be needed though if the majority of S report it to exist. A localisation technique such as TDOA will reveal the true location Pos_{TDOA}, for comparison to the claim Pos_{claim}. If these differ significantly, an attack is likely. If messages reach all $s \in S$ sensors, the anomalous aircraft will exist in $\exists c : c \in G \cap G_{S_A}$ thus a comparison will not reveal said aircraft; this is the worst case scenario and would rely on localisation such as TDOA to identify an anomalous aircraft.

An attacker using multiple locations to attack multiple sensors will make some positional claim Pos_{claim} using each location to ensure that a number of crowdsourced sensors receive it. This would cause $\exists c : c \in G \cap G_{S_A}$. However, to defeat TDOA, the time synchronisation required is beyond the capability of our attacker model.

Considering that a crowdsourced network such as OpenSky has significant coverage redundancy, an attacker would need to fool many sensors in order to successfully inject messages, even in the unlikely case that they know the exact location of all active sensors. Each additional sensor requires significant further complexity from an attack.

3) Modification Attacks

Modification attacks occur when the attacker changes data transmitted by an aircraft before it is received at a ground station. Based on [2], there are two different approaches to modify a target message; overshadowing and bit-flipping. Overshadowing transmits an entire message at a higher power than the message from the aircraft, whereas bit-flipping transmits a specific part of a message over the aircraft transmission at a higher power. As overshadowing is much easier to perform than bit-flipping, we assume it here.

Again, we consider the cases of an attacker transmitting from a single or multiple locations. We denote the aircraft under attack as and consider the possible attack intentions to either modify the position/status or identifiers. We first consider each of these for the single-location attacker.

For attackers modifying position, if a single sensor receives modified messages, comparing the position claims of $c \in G$ to $c_{attack} \in G_{S_A}$ will identify the anomalous position, allowing the aircraft in question to be highlighted as anomalous. If modified messages are received by multiple sensors in S then the position claims for c in $\forall\ s \in S : G_s$ must be compared. At this point we cannot know how many members of S are under attack. We can construct $S_{similar}$ by finding $s \in S$ where c in each is making similar position claims. A significant majority of S would need to agree before it could be taken as consensus – otherwise, a multilateration approach would be needed to check the claims for each group of similar sensors. The inaccurate sensors can then be deemed under attack.

Where attackers modify identity, if modified messages are reaching a single sensor, $G \setminus G_{S_A}$ will identify the aircraft c_{true} being attacked, with $G_{S_A} \setminus G$ identifying the modified aircraft c_{attack}. Using this, c_{attack} can be highlighted as anomalous. Furthermore, the positions of c_{true} and c_{attack} should be the same. If modified messages are reaching multiple sensors, we compare c in $\forall\ s \in S : G_s$ by positional claims. If sensors $s_1, \ldots, s_n$ report multiple of different identities aircraft $c_1, \ldots, c_m$ with the same position, this indicates an attack. Then, we must perform multilateration on messages from each $c_1, \ldots, c_m$ to identify which aircraft is in the claimed position and is not attacker-generated. From this we can establish which sensors are not under attack and are thus receiving legitimate messages.

In the case of an attacker using multiple locations, this becomes an extended case of a single location attacker transmitting to multiple sensors. We can detect anomalous aircraft by comparing the output of sensors, assuming the attacker is not simultaneously attacking all sensors. Since we assume our attacker does not have the required infrastructure to defeat TDOA, we rely on it to identify aircraft which are claiming to be in positions which they are not, and in the case of attacks on identity, establish which of the claimed aircraft identities for a given position are actually in the air.

B. Insider Attacks on OpenSky

Besides direct attacks on ATC technologies, it is imperative to consider the integral security of our crowdsourced system. Since we do not directly control all sensors used for attack detection, we require a separate trust-based security layer to detect and defeat internal attacks. We consider the manipulation of data by a single rogue sensor and the stronger Sybil attack.

1) Single Sensor Data Manipulation

First, we consider the case where an attacker manipulates a single sensor, which provides incorrect data to the network. In particular, an attacker may either act benignly for some time to gain trust or hijack a network connection of a trusted sensor. Given the range of a typical SSR sensor, this can result in a significant coverage area being affected.

Sanity checks will defeat more simplistic attacks; for example, should a sensor suddenly change its coverage area or begin to report significantly different numbers of aircraft then this is enough to identify a possible attack and request feedback from the sensor’s operator.

Analogous to outsider attacks, coverage redundancy can also be exploited to identify data manipulation. By identifying areas that overlap with other sensors and crosschecking aircraft appearance and properties, rogue sensors can be identified.

2) Sybil Attacks

Lastly, we consider Sybil attacks, the most sophisticated threat to a crowdsourced sensor network. In a Sybil attack, attackers attempt to integrate multiple sensors under their control (covering their targeted airspace) into the network. In order to hide a real-world ATC attack, these sensors feed false data to OpenSky, similar to the single sensor data manipulation, but this time in a coordinated fashion to outvote legitimate sensors covering the area.

To defend against such attacks, the system monitors new sensors for irregularities. For example, several new sensors in the same region in a short time will require review. However, as [36] proves, despite such defences, Sybil attacks are always possible as long as there is no trusted agency that certifies the identities of entities in a trust network. Other security-related networks such as Tor suffer from the same problem: new actors can behave well for any necessary time period to obtain the desired trust level before attacking.

Hence, the future aim of OpenSky is to individually certify and secure the connections to all sensors. While none of these defences are adequate to defend against the most powerful military or state actors, they are sufficient for our attacker model, from script kiddies to cyber criminals and cyber terrorists.

8. DISCUSSION

Finally, we discuss possible applications to other critical infrastructures, future extensions, and the costs of our system.

A. Application to other Critical Infrastructures

The principles introduced in this paper can also be applied to similar wireless systems. The closest example is provided by the automatic identification system (AIS), an automatic tracking system used on ships and other marine traffic [37]. Similar to ADS-B, its use cases include the identification and localisation of vessels through exchanging messages with nearby ships, base stations, and satellites. Like ATC, AIS also foregoes cryptographic measures, providing a large attack surface [38]. While the maritime scenario adds further difficulties such as shortened signal propagation and aggravated deployment of ground stations, these could be overcome, at least near busy traffic hotspots such as shores and straits.

B. Possible Future Extensions

Two natural extensions come to mind to improve the security of OpenSky in the future: the option of cryptographic certification of individual sensors (as discussed above) and the integration of non-stationary receivers.

By integrating non-stationary, i.e. mobile, receivers such as drones, we can make it harder for an attacker to obtain the exact position of all receivers, which in turn further improves the security of the discussed countermeasures. Related work has shown that a randomly moving mobile substantially increases the difficulty of making false messages appear genuine [39].

C. Cost

The International Civil Aviation Organization (ICAO) specifies the technological cost of using SSR to monitor an en-route airspace (200 NM radius) at $6 million, while a certified ADS-B system is estimated to be significantly cheaper at $380,000 [40]. This estimate assumes more expensive, certified sensors with extremely high availability and technical capabilities (in addition to the cost for the backend system).

As capable commercial off-the-shelf ADS-B receivers priced at around $100 already provide the basis of a highly available and redundant crowdsourced network in most Western airspaces, our system comes in at a fraction of these costs even when taking into account OpenSky's central processing unit and limited maintenance cost. Our best-placed receivers offer a reception radius of up to 600km, thus surveillance of an en-route airspace can be provided even by a single receiver. In practice, 10-50 sensors are sufficient for coverage with high accuracy, redundancy and reliability.

9. CONCLUSION

In this paper, we have presented an approach that exploits the concept of crowdsourcing to build a sensor network which acts as a defensive layer for the wireless interfaces of critical infrastructures. Using ATC as a case study, we introduced the crowdsourced sensor network OpenSky and laid out the potential for employing several independent countermeasures against common wireless attacks. Our security analysis shows that we can reliably detect even more sophisticated attackers and present a constantly validated picture of the airspace.

Based on these results, it is our strong belief that crowdsourcing can facilitate a transparent layer of security for many legacy wireless technologies that cannot be upgraded with cryptographic primitives in the short or medium term. The use of physical security primitives and a massively-distributed infrastructure provide a good defence against all but the most sophisticated military and nation-state attackers. Combined with attractive cost characteristics and agile deployment, crowdsourced networks pose a serious solution for the wireless security vulnerabilities that threaten many critical infrastructures today.

REFERENCES

[1] A. Costin and A. Francillon, 'Ghost in the Air (Traffic): On insecurity of ADS-B protocol and practical attacks on ADS-B devices,' in *Black Hat USA*, 2012, pp. 1-10.

[2] M. Schäfer, V. Lenders, and I. Martinovic, 'Experimental analysis of attacks on next generation air traffic communication,' *Lect. Notes Comput. Sci*., 2013, vol. 7954 LNCS, pp. 253-271.

[3] M. Balduzzi, A. Pasta, and K. Wilhoit, 'A security evaluation of AIS automated identification system,' in *Proceedings of the 30th Annual Computer Security Applications Conference*, 2014, pp. 436-455.

[4] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and others, ‘Predicting protein structures with a multiplayer online game,’ *Nature*, 2010, vol. 466, no. 7307, pp. 756-760.

[5] D. Clery, ‘Galaxy Zoo volunteers share pain and glory of research,’ *Science* (80-.), 2011, vol. 333, no. 6039, pp. 173-175.

[6] Flightradar24.com, ‘Successfully Testing Satellite-based ADS-B Tracking,’ Jul. 2016.

[7] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, ‘Bringing up OpenSky: A large-scale ADS-B sensor network for research,’ *IPSN 2014 - Proc. 13th Int. Symp. Inf. Process. Sens. Networks (Part CPS Week)*, 2014, pp. 83-94.

[8] M. Schäfer, M. Strohmeier, M. Smith, M. Fuchs, R. Pinheiro, V. Lenders, and I. Martinovic, ‘OpenSky’s Report 2016: Facts, Figures and Trends in Wireless ATC Communication Systems,’ in *35th Digital Avionics Systems Conference - Proceedings*, 2016.

[9] C. W. Johnson, ‘Cyber security and the future of safety-critical air traffic management: identifying the challenges under NextGen and SESAR,’ in *IET Conference Proceedings*, 2015.

[10] K. Zetter, ‘Feds Say That Banned Researcher Commandeered a Plane,’ *Wired*, May 2015.

[11] M. Strohmeier, M. Schäfer, M. Smith, V. Lenders, and I. Martinovic, ‘Assessing the impact of aviation security on cyber power,’ in *Cyber Conflict (CyCon), 2016 8th International Conference on*, 2016, pp. 223-241.

[12] M. Strohmeier, M. Schäfer, R. Pinheiro, V. Lenders, and I. Martinovic, ‘On Perception and Reality in Wireless Air Traffic Communications Security,’ *IEEE Transactions on Intelligent Transportation Systems*, October 2016.

[13] C. R. Spitzer, U. Ferrell, and T. Ferrell, *Digital Avionics Handbook*, 3rd ed. CRC Press, 2014.

[14] RTCA Inc., ‘Minimum Aviation System Performance Standards for Automatic Dependent Surveillance Broadcast (ADS-B).’ Dec-2006.

[15] M. Wilhelm, I. Martinovic, J. B. Schmitt, and V. Lenders, ‘Short Paper : Reactive Jamming in Wireless Networks – How Realistic is the Threat,’ *Proc. fourth ACM Conf. Wirel. Netw. Secur. (WiSec ’11)*, 2011, pp. 47-52.

[16] M. Schäfer, V. Lenders, and J. Schmitt, ‘Secure Track Verification,’ in *IEEE Symposium on Security and Privacy*, 2015, pp. 199-213.

[17] N. Xu, R. Cassell, and C. Evers, ‘Performance assessment of multilateration systems - a solution to NextGen surveillance,’ in *Integrated Communications Navigation and Surveillance Conference (ICNS)*, 2010, pp. 2-9.

[18] M. Schäfer, P. Leu, V. Lenders, and J. Schmitt, ‘Secure Motion Verification using the Doppler Effect,’ in *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 2016, pp. 135-145.

[19] M. Strohmeier, V. Lenders, and I. Martinovic, ‘Intrusion Detection for Airborne Communication using PHY-Layer Information,’ in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2015, pp. 67-77.

[20] A. Malatras and L. Beslay, ‘A generic framework to support participatory surveillance through crowdsensing,’ in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2015, pp. 1135-1146.

[21] T. Monahan and J. T. Mokos, ‘Crowdsourcing urban surveillance: The development of homeland security markets for environmental sensor networks,’ *Geoforum*, 2013, vol. 49, pp. 279-288.

[22] T. Ludwig, C. Reuter, T. Siebigteroth, and V. Pipek, ‘Crowdmonitor: mobile crowd sensing for assessing physical and digital activities of citizens during emergencies,’ in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 4083-4092.

[23] Flightradar24 AB, ‘Flightradar24,’ 2017. [Online]. Available: https://www.flightradar24.com. [Accessed: 06-Mar-2017].

[24] FlightAware, ‘FlightAware,’ 2017. [Online]. Available: https://www.flightaware.com/. [Accessed: 06-Mar-2017].

[25] S. Edwards, ‘Inside the World of People Who Track Flights for No Reason,’ *Vice*, Oct. 2015.

[26] Flightradar24 AB, ‘About Flightradar24,’ 2017.

[27] J. Scardina, ‘Overview of the FAA ADS-B link decision,’ Jun. 2002.

[28] K. D. Wesson, T. E. Humphreys, and B. L. Evans, ‘Can cryptography secure next generation air traffic surveillance?’ *IEEE Secur. Priv. Mag.*, 2014.

[29] M. Schäfer, M. Strohmeier, M. Smith, M. Fuchs, R. Pinheiro, V. Lenders, and I. Martinovic, ‘OpenSky’s Report 2016: Facts, Figures and Trends in Wireless ATC Communication Systems,’ in *35th Digital Avionics Systems Conference*, 2016.

[30] M. Robb, ‘Dump1090,’ *GitHub*, 2017. [Online]. Available: https://github.com/MalcolmRobb/dump1090. [Accessed: 12-Feb-2017].

[31] K. Pourvoyeur and R. Heidger, ‘Secure ADS-B usage in ATC tracking,’ in *2014 Tyrrhenian International Workshop on Digital Communications-Enhanced Surveillance of Aircraft and Vehicles (TIWDC/ESAV)*, 2014, pp. 35-40.

[32] M. Strohmeier, M. Schäfer, V. Lenders, and I. Martinovic, ‘Realities and challenges of NextGen air traffic management: the case of ADS-B,’ *IEEE Commun. Mag.*, 2014, vol. 52, no. 5.

[33] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, ‘Bringing up OpenSky: A large-scale ADS-B sensor network for research,’ *Proc. 13th Int. Symp. Inf. Process. Sens. Networks*, 2014, pp. 83-94.

[34] D. McCallie, J. Butts, and R. Mills, ‘Security analysis of the ADS-B implementation in the next generation air transportation system,’ *Int. J. Crit. Infrastruct. Prot.*, 2011, vol. 4, no. 2, pp. 78-87.

[35] M. Schäfer, V. Lenders, and I. Martinovic, ‘Experimental analysis of attacks on next generation air traffic communication,’ in *International Conference on Applied Cryptography and Network Security (ACNS)*, 2013, pp. 253-71.

[36] J. R. Douceur, ‘The sybil attack,’ in *International Workshop on Peer-to-Peer Systems*, 2002, pp. 251-260.

[37] B. Tetreault, ‘Automatic Identification System,’ *Proc. Mar. Saf. Secur. Counc.*, vol. 63, no. 3, 2006.

[38] M. Balduzzi, A. Pasta, and K. Wilhoit, ‘A security evaluation of AIS automated identification system,’ in *Proceedings of the 30th Annual Computer Security Applications Conference*, 2014, pp. 436-445.

[39] R. Baker and I. Martinovic, ‘Secure Location Verification with a Mobile Receiver,’ in *Proceedings of the 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy*, 2016, pp. 35-46.

[40] International Civil Aviation Organization (ICAO), ‘Guidance Material: Security issues associated with ADS-B,’ Montreal, QC, Canada, 2014.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

Scalable Architecture for Online Prioritisation of Cyber Threats

Fabio Pierazzi
Department of Engineering 'Enzo Ferrari'
University of Modena and Reggio Emilia
Modena, Italy
fabio.pierazzi@unimore.it

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

Alessandro Guido
Department of Engineering 'Enzo Ferrari'
University of Modena and Reggio Emilia
Modena, Italy
alessandro.guido@unimore.it

Mirco Marchetti
Department of Engineering 'Enzo Ferrari'
University of Modena and Reggio Emilia
Modena, Italy
mirco.marchetti@unimore.it

Abstract: Detecting advanced attacks is increasingly complex and no single solution can work. Defenders can leverage logs and alarms produced by network and security devices, but big data analytics solutions are necessary to transform huge volumes of raw data into useful information. Existing anomaly detection frameworks either work offline or aim to mark a host as compromised, with high risk of false alarms. We propose a novel online approach that monitors the behaviour of each internal host, detects suspicious activities possibly related to advanced attacks, and correlates these anomaly indicators to produce a list of the most likely compromised hosts. Due to the huge number of devices and traffic logs, we make scalability one of our top priorities. Therefore, most computations are independent of the number of hosts and can be naively parallelised. A large set of experiments demonstrates that our proposal can pave the way to novel forms of detection of advanced malware.

Keywords: *autonomous triage, early prioritisation, security analytics, scalability*

1. INTRODUCTION

The information systems of modern organisations are subject to a multitude of cyber attacks conceived by a wide range of attackers with different goals, capabilities and motivations. Despite all the efforts spent on preventive defences, the reality is that attacks occur every day and no organisation can consider itself secure. This paper shifts the focus from the *prevention* to the *detection* phase.

Existing proposals in academic literature detect specific attacks through heuristics and statistical analysis (e.g., [1, 2, 3, 4]). Most approaches (e.g., [2, 5]) rely on offline post-event analysis. Other online anomaly detectors assume that statistically detectable changes involve huge numbers of hosts (e.g., worm propagation in [6, 7]) or that compromised hosts share similar behaviours (e.g., botnet detection in [8, 9, 10, 11]). However, these assumptions are no longer true in modern human-driven advanced cyber attacks [12], hence existing proposals can be affected by many false positive and false negative alarms.

As no security operator wants to be annoyed by hundreds of alarms notified at the same priority level, we take a different direction and focus on ranking suspicious hosts instead of detecting compromised hosts. To this end, our online analysis begins by monitoring the behaviour of individual hosts over time and by identifying suspicious events involving even single or few hosts. These indicators are aggregated to produce a ranking of the most suspicious hosts, which are then provided to the security operator in a timely fashion.

Due to the massive amount of data to be managed online, we propose a scalable design and implementation of our approach. All initial phases before the final aggregation scale linearly with the number of hosts and can be parallelised. The proposed approach is general enough to be adopted with different types of data, including internal traffic, external traffic, alarms coming from IDS and SIEM, yet the goal of this paper is not to present a complete framework, but rather to propose the idea that the combination of autonomous triage with manual inspection increases the probability of detecting even advanced attacks. For these reasons, we present our approach relying only on network flows of internal corporate traffic, whose effectiveness is shown through experiments applied to networks of more than 1,000 hosts. We consider five main attack scenarios, representative of the activities that an attacker will likely perform from a compromised internal host: reconnaissance; data transfer to a dropzone; man in the middle; watering hole through DNS spoofing; and lateral movement through pivoting.

Section 2 of this paper presents related work. Section 3 outlines the main components and functions of the proposed approach. Section 4 describes the analytics core that extracts useful information and builds layer models from raw network data. Section 5 presents five examples of prioritisation algorithms that leverage outputs produced by the analytics core, along with results from real testbed networks. Section 6 concludes the paper with some final remarks and suggestions for future work.

2. RELATED WORK

Detecting advanced cyberattacks is increasingly difficult, as attackers have several ways to penetrate a network and hide their activities. The huge volume of logs generated by the multitude of servers, firewalls and devices are useful only when integrated with security analytics systems for automatic detection and triage. Considering the attacker's ability and the difficulty of signalling an infected host without causing false alarms, in the area of security analytics we propose an innovative approach. Instead of signalling an impossible 'guaranteed' detection, our system ranks the most suspicious hosts and leaves to the security analyst the task of inspecting a manageable number. Additional features include online processing for early prioritisation and scalability over thousands of hosts, as most analyses can be carried out independently for each host. The proposed approach can be applied on alerts and logs derived by IDS [13, 14], SIEM and other security appliances, and can be integrated with external traffic analyses; but in this paper we present a brand-agnostic approach based exclusively on flows of internal network traffic.

We identify three main areas of related works: offline forensics analysis, advanced malware detection, and online traffic monitoring.

The large majority of related proposals in the literature concern offline analysis for forensics purposes that differ from our online approach. Just to give some representative examples, we can cite [2] on heterogeneous logs analyses, [5] for its original graph-based approach for forensics, *BeeHive* [3], which correlates logs through histogram analysis to identify suspicious activities and corporate policy violations, [15] on forensics for cloud environments, and [16] on mobile forensics. Literature on advanced malware detection focuses on specific attack sequence patterns based on past APT campaigns [17, 18, 19, 20, 21], instead of detecting suspicious activities in each possible phase of an attack. Other more general solutions [17, 18] share our idea of prioritising suspicious hosts, but they are designed for offline or batch analysis.

The proposals based on online analyses focus on detection of DDoS [24, 25, 26], worm propagation and botnets, and these last two are the most related to our work. In worm propagation detection [7, 27], the internal network is usually modelled as a graph, where huge changes in the overall structure are identified as possible infection propagations. These works differ from our proposal because they focus on a specific threat, and their analyses look for huge changes in traffic volumes and patterns, whereas we prioritise signals of malicious activities related to behavioural changes of individual hosts. Our solution is scalable with respect to the number of monitored hosts, while worm propagation analysis depends on the size of the network graph. Botnet detection proposals [8, 9, 10, 11] are based on online scalable solutions for finding hosts that are possibly compromised. However, their underlying assumption is that a large number of hosts are compromised and share similar network behaviour, which is not true in the case of advanced cyberattacks where only a few hosts may be compromised and malicious actions are often human-driven.

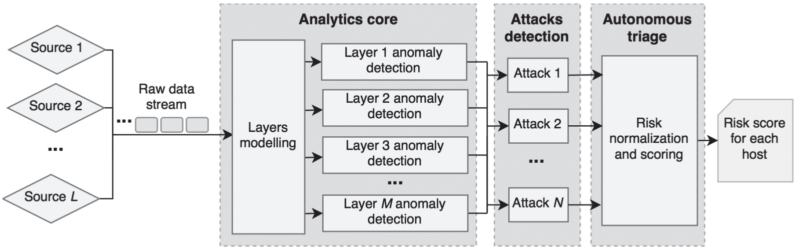
In summary, we can outline the major contributions that differentiate our work with respect to the state of the art:

- *ranking* of suspicious activities instead of specific detection(s);
- *online* analysis instead of offline post-mortem analysis;
- analysis based on *individual host* behaviour that guarantees parallel analyses and scalability; and
- the possibility of capturing suspicious actions involving even *a few hosts*.

3. FRAMEWORK OVERVIEW

We aim to detect anomalous network activities concerning each host of a corporation, and to use this information to rank the most suspicious hosts. In this section, we outline the proposed architecture and design choices for achieving scalability. Figure 1 emphasises that the input is represented by raw network data gathered by internal probes. Without loss of generality, in this paper we only consider network flows of traffic among internal hosts, which are feasible to collect and analyse for online contexts [28]. These logs are processed through three main steps: *analytics core, attack prioritisation and autonomous triage*. The final output is a list of internal hosts ranked by a risk score representing the likelihood that each host is involved in one or more attacks.

FIGURE1. FRAMEWORK OVERVIEW



Starting from raw network data, the *analytics core* builds different *layers*, which are graph models whose nodes represent internal hosts and edges represent a metric of interest. Each layer portrays a different perspective of the events occurring in the monitored network. For example, if we consider three layers, then edges may represent the number of packets, the number of bytes, and the average duration of the transmissions between two hosts, respectively. Then, the analytics core applies anomaly detection algorithms to the activities of each internal host within each layer. This fine-grained analysis is motivated by the observation that an attack related to a single host within a large internal network causes very small alterations that are not visible in an aggregated model comprising all layers and all hosts. Similar 'global' approaches work well only to identify massive attacks or network-wide anomalies [29, 4].

As a further advantage, since anomaly detection in different layers and hosts can be performed in parallel, the analytics core scales linearly with respect to the number of monitored hosts and layers. In this way, we can extend and improve an instance of the framework by adding more layers and nodes without having to change the information flow and the overall architecture. The algorithms adopted by the analytics core for layers modelling and anomaly detection are presented in Section 4.

The *attack prioritisation* module takes as its input the anomalies identified by the analytics core, and correlates them with the goal of detecting different attack scenarios, each one corresponding to activities that an attacker may perform from a compromised internal host. It is also possible to include novel attack detection algorithms with limited computational effort, because they can leverage the common fine-grained analyses already performed by the analytics core. The details of the attack prioritisation algorithms are discussed in Section 5.

The output of the attack prioritisation module is a risk score assigned to each internal host for each considered attack. Attack-specific risk scores for all hosts represent the input of the *autonomous triage* module which aids security operators by visualising the few hosts with higher ranks and the attacks in which they are likely involved.

4. ANALYTICS CORE

This section describes the algorithms used by the analytics core for *layers modelling* and for *anomaly detection* within each layer. The objective is to identify statistical anomalies for each host in all the layers, which will be correlated and ranked by the attack prioritisation module. The analytics core is designed for *online* processing and *scalability*.

A. Layers Modelling

Raw data is collected from the probes as soon as it is produced, and temporarily stored for a time defined by the *current time window* of size Δ. If t denotes the current time, then the layers modelling module maintains all raw data generated between $t-\Delta$ and t. Since previous research shows that most network activities are characterised by a daily periodicity [29, 23], it is convenient to set Δ equal to one day. At every *sampling interval* τ, all raw data in the current time window is used to compute the *current representation* of all layers. Since anomalies can be detected only after their appearance in the current representation of a layer, 'early' prioritisation is influenced by the choice of the parameter τ that is conveniently chosen in the order of a few minutes. Lower values cause useless oversampling of data (as an example, Netflow records related to long-lived connections are refreshed every 2 minutes [30]), while higher values introduce detection delays. We use the notation $L_i(t)$ to identify the current representation of the layer i, built using raw data in the current time window.

As shown in Figure 2, each $L_i(t)$ is modelled as a graph whose nodes represent hosts of the internal network, and edges denote some specific features of network activities occurring between the two hosts. As an example, a layer representing the number of bytes exchanged between internal hosts can be defined as a directed and weighted graph, in which edge direction denotes the direction of data transfer (from source to destination) and the weight represents the amount of transferred bytes.

FIGURE 2. ACTIVITIES OF THE LAYER MODELLING MODULE

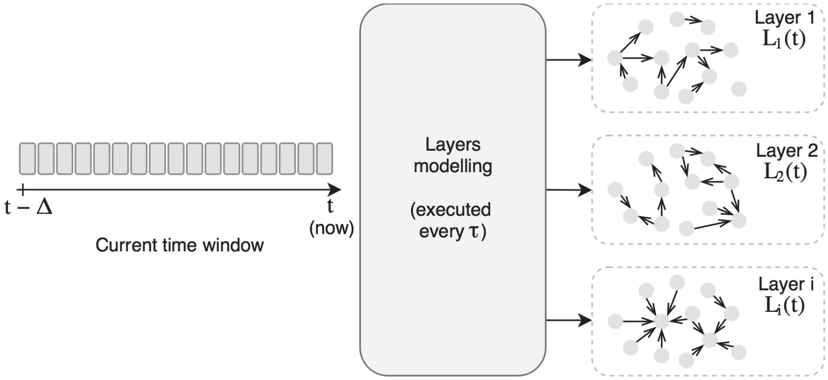


Table 1 reports the list of considered layers and their descriptions. These characteristics are commonly adopted to identify anomalies in traffic [31]. For example, time series of flows, packets, bytes and ports are used to identify reconnaissance activities [6] and data exfiltration [23]; graphs of internal communications are adopted for identification of worm propagation [7]; ARP messages can be useful for detecting eavesdropping activities [32].

TABLE 1. CONSIDERED LAYERS

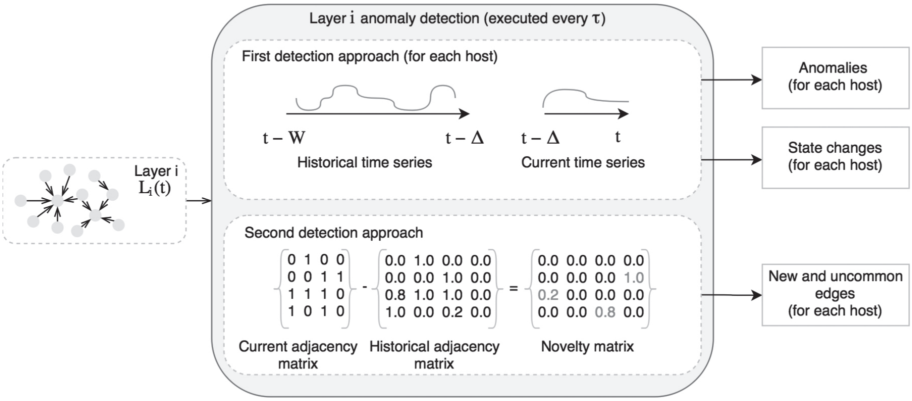
	Layer	Description
L_1	*Packets*	Directed weighted graph. Nodes are internal hosts and edges connect 2 nodes that exchange packets using any protocol. Direction is from source to target, and the weight of the edge is the total number of packets transmitted.
L_2	*Bytes*	Directed weighted graph. Nodes are internal hosts and edges connect 2 nodes that exchange packets using any protocol. Direction is from source to target, and the weight of the edge is the total number of bytes transmitted.
L_3	*Flows*	Directed weighted graph. Nodes are internal hosts and edges connect 2 nodes that exchange packets using any protocol. Direction is from source to target, and the weight of the edge is the total number of network flows.
L_4	*Ports*	Directed weighted graph. Nodes are internal hosts and edges connect 2 nodes communicating through TCP or UDP protocols. Direction is from source to target, and the weight of the edge is the number of different destination port numbers.
L_5	*Durations*	Directed weighted graph. Nodes are internal hosts and edges connect 2 nodes that exchange packets using any protocol. Direction is from source to target, and the weight of the edge is the average duration of network flows.
L_6	*Conns*	Directed unweighted graph. Nodes are internal hosts and edges connect 2 nodes that exchange IP datagrams. Direction is from source to target.
L_7	*Paths*	Bipartite directed graph. Both sets of nodes represent internal hosts. Edges connect each host from the first set to the hosts of the second set reachable by it through a ‘path’ composed of at least 3 hosts.
L_8	*DNS*	Bipartite directed graph. One set of nodes represents hostnames of internal hosts, the other set of nodes represents IP addresses. Edges connect a hostname to the associated IP address in DNS resolutions.
L_9	*ARP*	Bipartite directed graph. One set of nodes represents IP addresses of internal hosts, the other set of nodes represents MAC addresses. Edges connect the IP address and the MAC address that are bound as part of an ARP transaction.

The representations of all layers are passed as input to the processing modules that perform anomaly detection.

B. Layer Anomaly Detection

The goal is to identify hosts that exhibit anomalous behaviours in any of the layers. This step does not depend on identifiable attacks nor on the feature represented by each layer, hence all layers are subject to the same anomaly detection algorithms, which can be executed in parallel and independently. For each layer, we adopt two complementary detection approaches, as shown in Figure 3: the former identifies quantitative anomalies and state changes; the latter detects novel or uncommon events.

FIGURE 3. STRUCTURE OF A LAYER ANOMALY DETECTION MODULE



The former approach processes all current layer representations $L_i(t)$. The goal is to extract scalar values from graphs and to build time series. For each host, the framework computes two scalar values: the weighted in-degree and the weighted out-degree [33] representing the number of incoming and outgoing connections of each host in the current layer, respectively. Since a new $L_i(t)$ is received by the layer anomaly detection module (one at every sampling interval τ), scalar values for consecutive $L_i(t)$ are used to build two current time series representing recent values of in-degree and out-degree for each host. If t denotes the current instant of time, the current time series includes values between $t-\Delta$ and t. Moreover, to perform anomaly detection, it is necessary to build two historical time series including older scalar values between $t-W$ and $t-\Delta$ (excluded), where W represents the size of the historical window. W should be large enough (in the order of few weeks) to have a reliable baseline for the past behaviour of each host [28]. We observe that traffic among internal hosts exhibits more stability with respect to traffic among internal and external hosts, characterised by higher variability and consequent difficulties to achieve stable baseline models [29]. Anomaly detection is performed on each current time series through the online and adaptive detection algorithm proposed in [28] trained over the period W. This algorithm identifies both point anomalies and state changes [1] that reflect different kinds of relevant deviations between the current and past behaviours of an internal host.

The latter approach (cf. Figure 3) identifies new edges that never appeared in the historical window. For example, these edges may represent novel persistent connections of an attacker trying to perform lateral movement activities. For each $L_i(t)$, the detection algorithm computes its *current adjacency matrix* [33], which is a mathematical representation of the edges in $L_i(t)$, whose rows and columns represent the internal hosts: the matrix element (j, k) is set to 1 if $L_i(t)$ has an edge from host j to k; to 0 otherwise.

Older versions of the current adjacency matrix, built on previous $L_i(t)$ belonging to the historical time window W, are used to compute the *historical adjacency matrix*. Its values are rational numbers between 0 and 1. In particular, (j, k) denotes the frequency of occurrence of the edge from j to k in the older instances of $L_i(t)$. For example, the value (j, k) is set to 1 if all older instances of $L_i(t)$ layer contain an edge from j to k; if one-fifth of older $L_i(t)$ layers include an edge from j to k, then the value (j, k) is set to 0.2. The historical time window is updated every Δ.

At every sampling interval τ, the detection algorithm subtracts the historical adjacency matrix to the current adjacency matrix. The result, defined as *novelty matrix*, allows an immediate identification of new or uncommon edges that are present in $L_i(t)$, but never or seldom appeared in the historical time window. Uncommon edges having a low value in the historical adjacency matrix will result in values that are close to 1 in the novelty matrix; common edges with high values in the historical adjacency matrix will result in values close to 0 in the novelty matrix. The layer anomaly detection algorithm can sum the values included in each row of the novelty matrix to evaluate the ‘novelty’ of all the edges starting from the corresponding host. Similarly, the ‘novelty’ of all edges that end in any internal host is computed by summing the values on the corresponding column of the novelty matrix.

Anomalies, state-changes and novel edges detected by the analytics core are then used by the algorithms which evidence malicious activities and prioritise them.

5. ATTACK PRIORITISATION AND RANKING

The main goal is to prioritise signals of malicious activities that may be part of an advanced attack. To this purpose, we correlate the anomalies, state-changes and novel edges detected by the analytics core.

A. Experimental Testbed

There are several possible indicators associated with malicious activities. In this paper, we consider: reconnaissance (R); data transfer to a dropzone (DTD); man in the middle (MITM); watering hole through DNS spoofing (WH); and lateral movement through pivoting (LM). Table 2 indicates which *layer models* are included in the analysis of each attack scenario. The presence of multiple layers increases confidence that a suspicious activity is actually occurring. The attack prioritisation module evaluates a *risk score* for each internal host by combining the anomalies, state-changes and novel edges detected by the analytics core. We refer to the following notations:

- $\boldsymbol{A_{L_i}^{in}}$ (resp. $\boldsymbol{A_{L_i}^{out}}$) denotes the intensity of the biggest point anomaly in the incoming (resp. outgoing) time series related to layer L_i of an internal host during the observed window [3]. For example, a burst in the outgoing bytes.
- $\boldsymbol{C_{L_i}^{in}}$ (resp. $\boldsymbol{C_{L_i}^{out}}$) denotes the intensity of possible state-changes in the incoming (resp. outgoing) time series related to layer L_i of an internal host. For example, a state-change is detected if the average number of packets in the current window doubles for an extended period (hence, it is not only a point anomaly [4]).
- $\boldsymbol{N_{L_i}^{in}}$ NLiin (resp. $\boldsymbol{N_{L_i}^{out}}$) denotes the number of new incoming (resp. outgoing) edges of an internal host in the graph of layer L_i. For example, it can be used to detect the number of newly contacted hosts in the current time window.

All formulas and scores in this section are computed for each host.

TABLE 2. LAYERS USED TO PRIORITISE DIFFERENT TYPES OF ATTACKER ACTIVITIES

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	L_9
Attack	**Packets**	**Bytes**	**Flows**	**Ports**	**Durations**	**Conns**	**Paths**	**DNS**	**ARP**
Reconnaissance			X	X	X	X			
Data transfer	X	X			X	X			
MITM	X	X	X			X			X
Watering hole					X	X		X	
Lateral movement					X		X		

In the experiments, we consider an internal network consisting of more than 1,000 hosts composed of about 800 clients and 200 servers. The client machines have heterogeneous operating systems including several versions of Mac OS, Linux and Windows. The server machines host mainly websites and DBMS, but also high performance computations, code versioning and NAS storage. We place monitoring probes in the main 1Gbit switches of the network. Our algorithms are executed on a cluster of eight blades, each having an Intel Xeon 2.6GHz CPU and 16GB of RAM. Network flows are sampled every five minutes.

To evaluate scalability, we consider three scenarios consisting of 96, 287 and 1,012 hosts, respectively. One cluster node is sufficient for computations related to 96 and 287 hosts, while four nodes are necessary for the scenario with 1,012 hosts. This scalability is achieved because all computations of the analytics core are performed independently for each host and for each layer. Operations of the attack prioritisation module do not scale linearly, but their computational cost is negligible with respect to the anomaly detection algorithms of the analytics core. We present the details of the prioritisation of the five attack scenarios, and how risk scores are shown to the security operators.

B. Prioritisation of Suspicious Activities

For each scenario, we inject multiple attacks in some hosts of the network, we apply our analytics and evaluate a *risk score* for each host.

1) Reconnaissance in Internal Network

An attacker having control of an internal host likely scans neighbour hosts looking for (known or zero-day) vulnerabilities [6, 26]. We define the risk score R for reconnaissance as follows:

$$R = \frac{A_{Flows}^{out} + A_{Ports}^{out} + N_{Conns}^{out}}{1 + A_{Durations}^{out}}$$

A higher value of R denotes a higher likelihood that an internal host is performing a scan. Intuitively, when an internal host performs a reconnaissance activity, the average duration of its connections decreases (due to many volatile communications) while the numbers of flows, ports and contacted hosts increase. To evaluate the risk score for this attack, we carry out reconnaissance activities from 10 hosts by varying the scan intensity in terms of number of scanned hosts and ports, as described in Table 3.

TABLE 3. RECONNAISSANCE ATTACKS INJECTED IN THE INTERNAL NETWORK FROM 10 HOSTS

Attack	#ports scanned	#hosts scanned
horizontal scan	1 single port	from 50 to 1,000 distinct ports
vertical scan	from 50 to 1,000 distinct hosts	1 single host
block scan	from 50 to 1,000 distinct hosts	from 50 to 1,000 distinct ports

Since our approach focuses on prioritisation, we evaluate how many times an internal host performing the attack is ranked within the top-K hosts. Table 4 reports the results of multiple experiments executed over several weeks, where each row represents the percentage of times an internal host performing the attack has been ranked within the top-K. Each column corresponds to horizontal, vertical, or block scan experiments as described in Table 3.

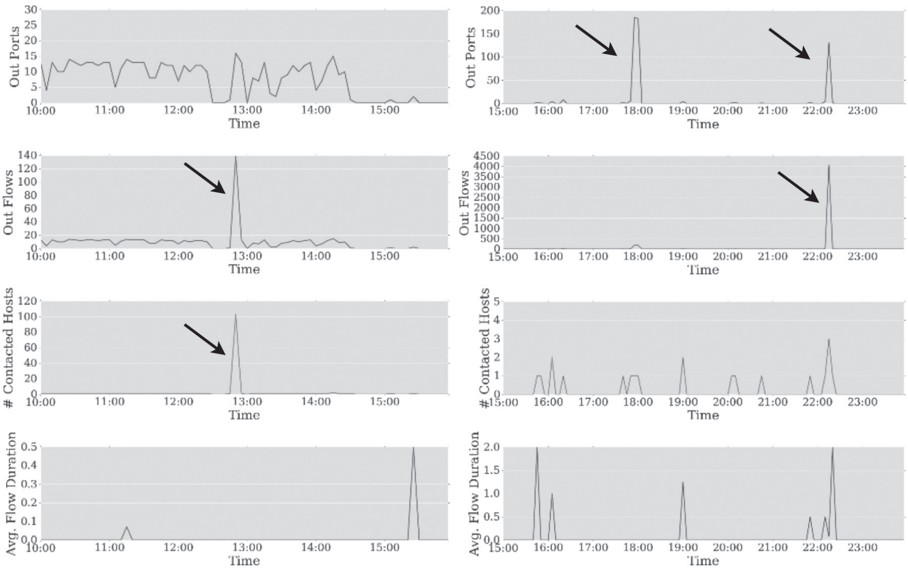
TABLE 4. PERCENTAGE OF TIMES A HOST PERFORMING A RECONNAISSANCE IS RANKED WITHIN THE TOP-K

In top-K	Horizontal scan	Vertical scan	Block scan
in top-5	94.3%	92.4%	99.2%
in top-10	99.5%	99.1%	99.7%
in top-25	100%	99.7%	100%
in top-50	100%	100%	100%

Table 4 shows that in more than 99% of the cases a host performing a reconnaissance activity is ranked within the top-10. Horizontal scans are easier to detect because they span over multiple hosts, whereas vertical scans are harder because it is more common for clients to contact servers on multiple ports if they offer more than one service. As expected, block scans have higher rankings, because they span both over multiple ports and multiple hosts.

Figure 4 reports an example of the traffic time series used to compute the risk score R, extracted from the layers *Ports, Flows, Conns, Durations*. The X-axis represents time, and the Y-axis reports the value of different metrics. Small arrows highlight significant anomalies: a horizontal scan around 12:48 and two vertical scans around 17:55 and 22:20.

FIGURE 4. TIME SERIES OF AN INTERNAL HOST PERFORMING HORIZONTAL AND VERTICAL SCANS



2) Data Transfer to Dropzone Before Exfiltration

Attackers often move data to be exfiltrated towards an internal *dropzone* [23, 12], used as intermediate point from which the exfiltration is easier to perform. These activities can be detected through the risk score DTD defined as follows:

$$DTD = \frac{A^{out}_{Durations} + A^{out}_{Bytes} + A^{out}_{Packets}}{1 + N^{out}_{Conns}}$$

A higher value of DTD suggests that an internal host is likely transferring data to an internal dropzone. In the numerator, we consider point anomalies instead of state changes because the higher bandwidth of internal networks – typically in the order of *Gbps* – allows for short transfer times. In the denominator, we consider N^{out}_{Conns} to rule out legitimate intensified network activity, such as p2p protocols.

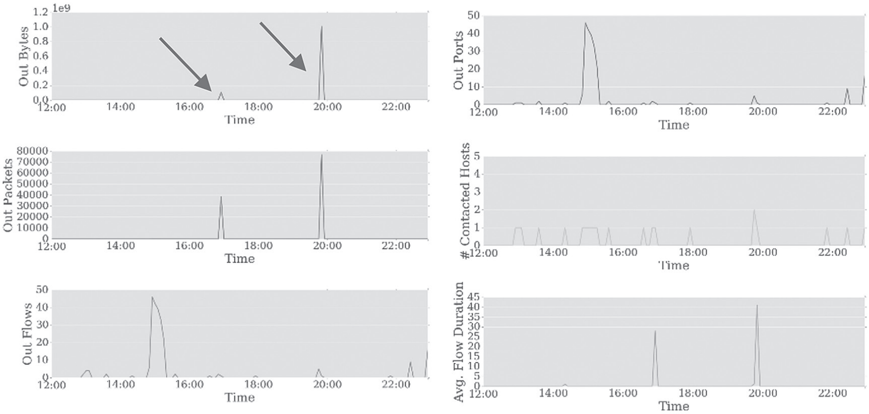
We perform several experiments in which we simulate DTD attacks of increasing transfer sizes from 10MB-50MB to 100MB-1GB. We use five controlled hosts as possible attackers and we transfer data to a Web server of the organisation as an emulated dropzone. Table 5 reports the percentage of times a host performing a DTD is ranked within the top-K: for small amounts of data (10-50MB), in about 92% of the cases the hosts are ranked within the top-10. This is the most challenging scenario for detection, because clients may use multiple hosts or devices for backup, hence it is difficult to identify anomalous transfers unless we integrate anomaly detection with white and black lists of internal hosts (where storage is or is not allowed), but similar integrations are out of the scope of this paper. The most compelling result is that in 99% of the cases, the 50-100MB internal transfers are ranked within the top-10.

TABLE 5. PERCENTAGE OF TIMES A HOST PERFORMING A DTD IS RANKED WITHIN THE TOP-K

In top-K	10-50MB	50-100MB	100MB-1GB
in top-5	87.5%	95.4%	99.7%
in top-10	91.7%	99.1%	100%
in top-25	95.6%	99.8%	100%
in top-50	99.5%	100%	100%

As an example, Figure 5 reports time series that show the evolution of several layers referring to an internal host used for injecting data transfers to the dropzone. The X-axis represents time, and the Y-axis reports the different metrics. Two arrows highlight peaks of about 100MB and 1GB, respectively – corresponding to the injected data exfiltration. We also observe an increment of the average flow duration in correspondence of the two data transfers, whereas other statistics (e.g., number of contacted hosts) remain stable.

FIGURE 5. TIME SERIES OF AN INTERNAL HOST IN WHICH TWO DTDS OF 100MB AND 1GB ARE INJECTED



3) MITM: Man in the Middle Attack

Man in the middle (MITM) attacks are used to perform advanced reconnaissance or to steal credentials, because an attacker can eavesdrop on communications of hosts within the same subnet. Here, we consider one of the most subtle forms of MITM, performed through ARP spoofing [34]. In this scenario, an attacker sends fake correspondence between IP and MAC addresses with the goal of acting as ‘hidden’ proxy between a victim and the gateway of its subnet. Netflows record no explicit communication between the eavesdropper and the victim, but our experiments show that once a host becomes a victim of MITM, then all packets sent and received by the victim pass twice through the switch. This attack can be captured by the state-change detection algorithm of the analytics core. In order to prioritise possible victims of MITM we define the following risk score:

$$MITM=\left(\frac{C_{Bytes}^{in}+C_{Bytes}^{out}+C_{Pkts}^{in}+C_{Pkts}^{out}}{1+C_{Flows}^{in}+C_{Flows}^{out}+N_{Conns}^{in}+N_{Conns}^{out}}\right)*N_{ARP}^{out}$$

In the numerator, we consider state-changes instead of point anomalies because MITM is usually an activity that lasts for some time to get useful information. The parameter N_{ARP}^{out} is a multiplicative factor because if $N_{ARP}^{out}=0$ there is no new correspondence in the ARP layer (see Section 4). In the denominator, we include state-changes and novel edges in *Flows* and *Conns* layers, because they must remain approximately stable with respect to a past window even if MITM is occurring.

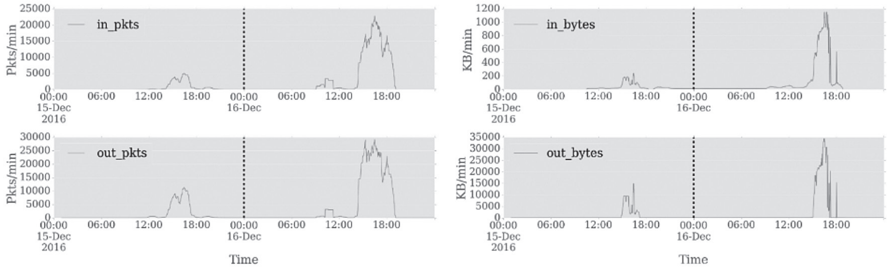
Experimental results are achieved through controlled MITM attacks of varying durations where we use one host as the eavesdropper and other 10 hosts as victims. Table 6 shows that in more than 95% of the cases even MITM lasting for just 15-30 minutes are prioritised in the top-10; if an attack lasts for at least 1 hour, the victim hosts are ranked within the top-5 in more than 98% of the cases.

TABLE 6. PERCENTAGE OF TIMES A HOST VICTIM OF A MITM IS RANKED WITHIN THE TOP-K

In top-K	15-30min	1-2hr	12-24hr	24-72hr
in top-5	89.8%	98.2%	99.4%	99.8%
in top-10	95.4%	99.1%	99.8%	100%
in top-25	99.0%	99.8%	100%	100%
in top-50	99.7%	100%	100%	100%

As an illustration, in Figure 6 we report the time series of a host related to *Packets* and *Bytes* layers, where the Y-axis denotes the different metrics, and the X-axis reports time. The plots report two days separated by a vertical dashed line. When the MITM occurs on the second day, it is possible to observe that the number of packets and bytes increases significantly.

FIGURE 6. COMPARISON SHOWING CHANGES IN PACKETS AND BYTES LAYERS WHEN MITM OCCURS



4) Watering Hole through DNS Spoofing

'Watering hole' is a technique used by attackers to increase their coverage and persistence by infecting multiple hosts of an organisation simultaneously. We consider a particular type of watering hole attack performed through DNS spoofing [35], where the attacker spoofs DNS responses to redirect victims to a compromised sever. To prioritise internal hosts that may correspond to watering holes, we define the risk score WH as follows:

$$WH=\left(N_{Conns}^{in}+C_{Conns}^{in}+C_{Durations}^{in}\right)*N_{DNS}^{out}$$

A high value of WH represents a higher likelihood that a host is performing a watering hole through DNS spoofing. Intuitively, this can be prioritised when a host has many new incoming connections, its IP corresponds to a new DNS resolution, and it has a state-change in the number and duration of incoming connections. We observe that N_{DNS}^{out} is a multiplicative factor, because N_{DNS}^{out} =0 implies that no DNS spoofing occurred.

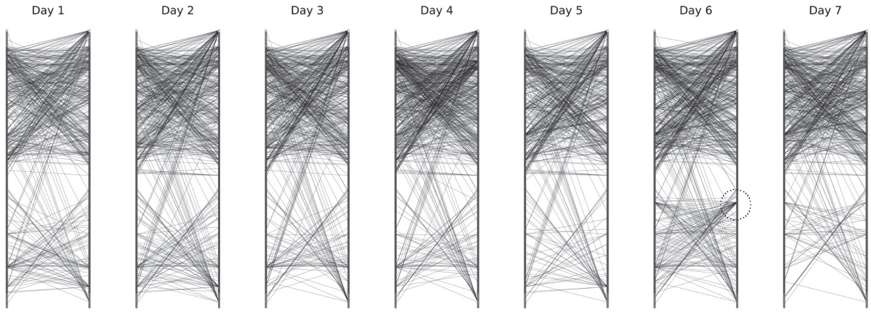
To evaluate the risk score WH, we use five internal clients as 'spoofers' of three internal Web servers offering different services: *Server 1 (small), Server 2 (medium)* and *Server 3 (large)*, which are used by an average number of clients per hour of about 10, 50 and 250, respectively. We perform a DNS spoofing at different times of the day over several weeks. Table 7 reports the percentage of times a watering hole host (that was redirecting traffic to itself through DNS spoofing) has been ranked within the top-K hosts in the different scenarios. This table shows that for Server 1 (having small activity) the spoofer is prioritised in the top-10 in more than 96% of the cases, while this percentage is even higher for servers with high number of clients where the intensity of the redirect is more evident.

TABLE 7. PERCENTAGE OF TIMES A HOST PERFORMING 'WATERING HOLE' IS RANKED WITHIN THE TOP-K

In top-K	Server 1 (small)	Server 2 (medium)	Server 3 (large)
in top-5	94.7%	98.9%	99.9%
in top-10	96.2%	99.8%	100%
in top-25	99.5%	100%	100%
in top-50	99.8%	100%	100%

As an example, Figure 7 reports a bipartite graph representation of the *Conns* layer over different days: on day 6, the dotted circle highlights a host that started performing a watering hole attack through DNS spoofing. We can observe an increase in the number of the incoming communications.

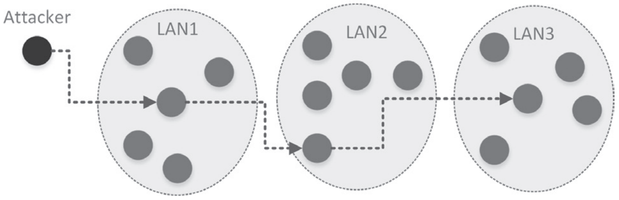
FIGURE 7. BIPARTITE COMMUNICATIONS GRAPH DERIVED FROM CONNS LAYER OVER 7 DIFFERENT DAYS



5) Lateral Movement through Pivoting

To get closer to his target, an attacker tends to compromise several internal hosts with higher privileges or to access the most internal parts of the corporate network. This activity is called *lateral movement* [12]. Figure 8 reports an example through a common technique named *pivoting* [36], where an attacker creates a tunnel of communications among multiple hosts *(pivoters)* to access a LAN that cannot be reached directly from the outside.

FIGURE 8. EXAMPLE OF LATERAL MOVEMENT THROUGH PIVOTING



To prioritise lateral movements, we define the risk score LM as follows:

$$LM = \left(N_{Paths}^{in} + N_{Paths}^{out} + C_{Durations}^{in} + C_{Durations}^{out} \right)$$

where N_{Paths}^{in} and N_{Paths}^{out} take into account new paths in the communications graph (see Section 4), while $C_{Durations}^{in}$ and $C_{Durations}^{out}$ check whether an increment in the average duration of the flows has occurred. (We recall that a *pivoting* tunnel has to last for some time [36]). We perform several experiments involving up to 10 controlled clients as intermediate pivoter hosts in the lateral movement (see Figure 8). Table 8 reports the percentages of times a pivoter host is ranked within the top-K risky nodes. The different columns correspond to different *lengths* of the pivoting tunnel. For example, Figure 8 presents a tunnel of length 2 with two pivoter hosts.

TABLE 8. PERCENTAGE OF TIMES A HOST PERFORMING LM IS RANKED WITHIN THE TOP-K

In top-K	1 pivoter	3-5 pivoters	8-10 pivoters
in top-5	96.2%	99.7%	99.9%
in top-10	97.9%	99.9%	100%
in top-25	99.1%	100%	100%
in top-50	99.8%	100%	100%

C. Autonomous Triage to Support Security Analysts

We present how results can be combined to produce an overall ranking, useful for security analysts to focus on few suspicious hosts.

FIGURE 9. ONLINE AUTONOMOUS TRIAGE OF INTERNAL HOSTS FOR DIFFERENT ATTACK SCENARIOS

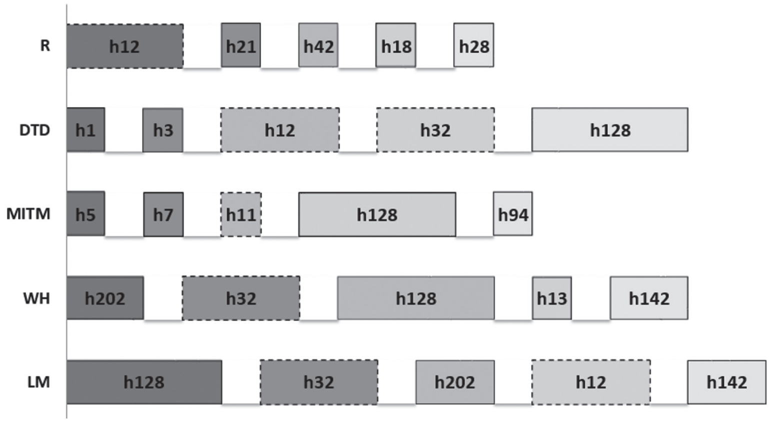


Figure 9 reports the overall rankings, with each line corresponding to a different attack: R for reconnaissance, DTD for data transfer to dropzone, MITM for man in the middle, WH for watering hole through DNS spoofing, and LM for lateral movement through pivoting. On the leftmost or rightmost side is the host with higher or lower risk score, respectively, on that line. All hosts are represented as rectangles whose size is proportional to the number of top-K rankings in which a host appears. For example, host *h202* appears in two top-five rankings (first for WH, and third for LM), hence its rectangle has a double size. As ranking operations are evaluated online, rectangles with a dashed outline denote hosts that recently entered a top-five ranking. A similar visualisation supports security analysts in monitoring several cyber threats occurring in the core of large networks. The number of top-K hosts to show can be adapted, depending on the size of the organisation and the amount of human resources. As a final remark, it is important to observe that our proposal makes it hard for an attacker to evade ranking, because we monitor changes in activity of each *individual host* with respect to its history, and we produce an overall ranking considering all hosts of the internal network. Hence, an attacker would require a *complete* view of all hosts' behaviours and history to evade prioritisation successfully.

6. CONCLUSIONS

In this paper, we propose a novel approach based on ranking and prioritisation instead of 'guaranteed' detection. We consider an innovative perspective in which we start by analysing *individual host* behaviours, and then post-correlate outputs to compute various indicators corresponding to different attacker activities. A prioritised list of likely compromised hosts is passed to human analysts, who can focus their attention only on the most suspicious hosts and activities. Experimental evaluations and use-case examples in real-world internal networks of more than 1,000 hosts demonstrate the feasibility and scalability of the proposed approach for online autonomous triage of different attack scenarios. Future works include the integration of attack indicators from external traffic, such as text analysis and statistical characterisation of DNS queries to identify possible C&Cs.

REFERENCES

[1] V. Chandola, A. Banerjee and V. Kumar, 'Anomaly detection: A survey,' *ACM Computing Surveys (CSUR)*, 2009.

[2] E.S. Pilli, R.C. Joshi and R. Niyogi, 'Network forensic frameworks: Survey and research challenges,' *Elsevier Digital Investigation*, 2010.

[3] T.-F. Yen, A. Oprea, K. Onarlioglu, T. Leetham, W. Robertson, A. Juels and E. Kirda, 'Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks,' in *Proceedings of the 29th ACM Annual Computer Security Applications Conference*, 2013.

[4] M. Andreolini, M. Colajanni and M. Marchetti, 'A collaborative framework for intrusion detection in mobile networks,' *Elsevier Information Sciences*, 2015.

[5] W. Wang and T. E. Daniels, 'A graph based approach toward network forensics analysis,' *ACM Transactions on Information and System Security (TISSEC)*, 2008.

[6] S. J. Stolfo, 'Worm and attack early warning: piercing stealthy reconnaissance,' *IEEE Security & Privacy*, 2004.

[7] M. P. Collins and M. K. Reiter, 'Hit-list worm detection and bot identification in large networks using protocol graphs,' in *International Workshop on Recent Advances in Intrusion Detection*, 2007.

[8] M. Bailey, E. Cooke, F. Jahanian, Y. Xu and M. Karir, 'A survey of botnet and botnet detection,' in *Conference for Homeland Security, CATCH '09, Cybersecurity Applications & Technology*, 2009.

[9] 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 ACM 28th Annual Computer Security Applications Conference*, 2012.

[10] G. Gu, P.A. Porras, V. Yegneswaran, M.W. Fong and W. Lee, 'BotHunter: Detecting Malware Infection Through IDS-Driven Dialog Correlation,' in *Usenix Security*, 2007.

[11] G. Gu, R. Perdisci, J. Zhang and W. Lee, 'BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection,' in *Usenix Security Symposium*, 2008.

[12] R. Brewer, 'Advanced persistent threats: minimising the damage,' *Network Security*, pp. 5-9, 2014.

[13] M. Colajanni, D. Gozzi and M. Marchetti, 'Enhancing interoperability and stateful analysis of cooperative network intrusion detection systems,' in *Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems*, 2007.

[14] M. Marchetti, M. Colajanni and F. Manganiello, 'Framework and Models for Multistep Attack Detection,' *International Journal on Security and Its Applications*, 2011.

[15] K. Ruan, J. Carthy and T. Kechadi, 'Survey on cloud forensics and critical criteria for cloud forensic capability: A preliminary analysis,' in *Proceedings of the Conference on Digital Forensics, Security and Law*, 2011.

[16] J. Grover, 'Android forensics: Automated data collection and reporting from a mobile device,' *Elsevier Digital Investigation*, 2013.

[17] P. Bhatt, E. Toshiro Yano and P. M. Gustavsson, 'Towards a Framework to Detect Multi-stage Advanced Persistent Threats Attacks,' in *IEEE International Symposium on Service Oriented System Engineering (SOSE)*, 2014.

[18] E. M. Hutchins, M. J. Cloppert and R. M. Amin, 'Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains,' in *Proceedings of the 6th International Conference on i-Warfare and Security*, 2011.

[19] R. Brewer, 'Advanced persistent threats: minimising the damage,' *Network Security*, pp. 5-9, 2014.

[20] I. Jeun, Y. Lee and D. Won, 'A practical study on advanced persistent threats,' *Computer Applications for Security, Control and System Engineering*, pp. 144-152, 2012.

[21] N. Virvilis and D. Gritzalis, 'The big four – what we did wrong in advanced persistent threat detection?' in *IEEE International Conference on Availability, Reliability and Security (ARES)*, 2013.

[22] M. Marchetti, F. Pierazzi, A. Guido and M. Colajanni, 'Countering Advanced Persistent Threats through security intelligence and big data analytics,' in *Cyber Power, 2016 8th International Conference on Cyber Conflict*, Tallinn, Estonia, 2016.

[23] M. Marchetti, F. Pierazzi, M. Colajanni and A. Guido, 'Analysis of high volumes of network traffic for Advanced Persistent Threat detection,' *Elsevier Computer Networks*, 2016.

[24] J. McPherson, K.-L. Ma, P. Krystosk, T. Bartoletti and M. Christensen, 'Portvis: a tool for port-based detection of security events,' in *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, 2004.

[25] V. Sekar, N. G. Duffield, O. Spatscheck, J. E. van der Merwe and H. Zhang, 'LADS: Large-scale Automated DDoS Detection System,' in *Usenix Annual Technical Conference, General Track*, 2006.

[26] M. H. Bhuyan, D. Bhattacharyya and J. K. Kalita, 'Surveying port scans and their detection methodologies,' *The Computer Journal*, 2011.

[27] S. Staniford-Chen, S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip and D. Zerkle, 'GrIDS-a graph based intrusion detection system for large networks,' in *Proceedings of the 19th National Information Systems Security Conference*, 1996.

[28] S. Casolari, S. Tosi and F. Lo Presti, 'An adaptive model for online detection of relevant state changes in Internet-based systems,' *Performance Evaluation*, pp. 206-226, 2012.

[29] F. Pierazzi, S. Casolari, M. Colajanni and M. Marchetti, 'Exploratory security analytics for anomaly detection,' *Computers & Security*, pp. 28-49, 2016.

[30] 'nProbe,' [Online]. Available at: http://www.ntop.org/products/netflow/nprobe/.

[31] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras and B. Stiller, 'An overview of IP flow-based intrusion detection,' *IEEE communications surveys & tutorials*, 2010.

[32] P. Goyal, S. Batra and A. Singh, 'A literature review of security attack in mobile ad-hoc networks,' *International Journal of Computer Applications*, 2010.

[33] M. Newman, *Networks: An introduction*, Oxford University Press, 2010.

[34] V. Ramachandran and S. Nandi, 'Detecting ARP spoofing: An active technique,' in *International Conference on Information Systems Security*, 2005.

[35] U. Steinhoff, A. Wiesmaier and R. Araujo, 'The State of the Art in DNS Spoofing,' in *Proc. 4th Intl. Conf. Applied Cryptography and Network Security (ACNS)*, 2006.

[36] Offensive-Security.com, 'Pivoting,' 2016. [Online]. Available: https://offensive-security.com/metasploit-unleashed/pivoting/.

2017 9th International Conference on Cyber Conflict
Defending the Core
H. Rõigas, R. Jakschis, L. Lindström, T. Minárik (Eds.)
2017 © NATO CCD COE Publications, Tallinn

On the Security of the Blockchain BIX Protocol and Certificates

Riccardo Longo
Department of Mathematics
University of Trento
Trento, Italy
riccardolongomath@gmail.com

Federico Pintore
Department of Mathematics
University of Trento
Trento, Italy
federico.pintore@unitn.it

Giancarlo Rinaldo
Department of Mathematics
University of Trento
Trento, Italy
giancarlo.rinaldo@unitn.it

Massimiliano Sala
Department of Mathematics
University of Trento
Trento, Italy
maxsalacodes@gmail.com

Abstract: In recent years certification authorities (CAs) have been the target of multiple attacks due to their sensitive role in internet security. In fact, with access to malicious certificates it is possible to mount effective large-scale man-in-the-middle attacks that may become very vicious, especially if the incident is not properly handled. Many attacks, such as the 2011 ones against DigiNotar and Comodo, also show strong hints of state sponsorship; thus, CAs have to be considered primary targets in a scenario of (possibly state-sponsored) large-scale cyber attacks. Therefore, there is a need for a PKI protocol which is more resilient and without single points of failure, such as the CAs. The BIX protocol is a blockchain-based protocol that allows distribution of certificates linking a subject with their public key, hence providing a service similar to that of a PKI but without the need for a CA. In this paper, we analyse the security of the BIX protocol in a formal way. First, we identify formal security assumptions which are well-suited to this protocol. Second, we present some attack scenarios against the BIX protocol. Third, we provide formal security proofs that these attacks are not feasible under our previously established assumptions.

Keywords: *PKI, security proof, blockchain*

1. INTRODUCTION

Blockchain is an emerging technology that is becoming widely adopted to solve a myriad of problems where the classic centralised approach can be substituted by decentralisation. Indeed, centralised computations, albeit efficient, are possible only if there is a trusted third party (TTP) that everybody trusts. Nowadays, this is sometimes felt as a limitation and a possible vulnerability.

The general idea behind blockchain technology is that blocks containing information are created by nodes in the network, and these blocks are both public and cryptographically linked, so that an attacker should be unable to modify them without the users noticing the tampering. Also, the information contained in any block comes from the users and any user signs their own information cryptographically. Some examples of blockchain applications can be found in [1], [2] and [3].

A very sensitive security aspect which is usually kept centralised is the issuing of digital certificates, which form the core of a public key infrastructure (PKI). A certificate contains at least a cryptographic public key and it is digitally signed by a TTP. An example is X.509 certificates [4], mostly containing RSA public keys, which are widely used in the Internet for establishing secure transactions (e.g., an e-payment with an e-commerce site like Amazon). Since every user of a PKI must trust the certification authority (CA), which acts as a TTP, the identity of a web site is checked by verifying the CA's signature via the CA's public key. In a scenario of (possibly state-sponsored) large-scale cyber attacks, CAs may become primary targets because of their strategic role in guaranteeing the authentication and security of most web resources. Unfortunately, their role becomes a liability if they are compromised in the attack, since it becomes impossible for the attacked infrastructure to distinguish fake servers from real ones. In recent years, there have been multiple attacks against CAs, one of the most notable being the one that brought DigiNotar to bankruptcy in 2011. In that case, an intrusion led to the issuing of multiple malicious certificates and poor handling of the crisis left users exposed (with evidence of exploitation of these certificates in Iran) for months, and almost crippled the Dutch PKI. Other attacks saw prominent CAs among the targets, such as Comodo, StartSSL and Verizon. For details about the listed attacks we refer to [5].

Therefore, there is the need for a PKI protocol which is more resilient to wide cyber attacks and which does not introduce *single points of failure*, such as the CAs. This is exactly the idea behind the so-called BIX certificates. The BIX protocol aims to distribute the role of the CAs while preserving the security features. Indeed, the BIX protocol is designed with a blockchain-like structure that provides integrity to data, showcasing the possibility of a distributed PKI. A certificate is a block in a blockchain and a valid user interacting properly with the protocol will be able to attach their certificate to the blockchain. The protocol works with very few assumptions on the underlying network, but the original paper by Sead Muftic [6] focuses on the innovative ideas and the technology behind them, leaving formal proofs of security as a stimulating open research problem. Fascinated by his approach, in our present paper we prove the security of the BIX protocol, while providing suitable formal models for threat scenarios.

We achieve this by giving a mathematical reduction of the attacks to the solution of some (well-known) hard cryptographic problems. First, we suppose that an attacker tries to attach their certificate to a pre-existing certificate chain *without* interacting properly with the protocol. This is equivalent to having a malicious user trying to forge a valid certificate for themself (or for an innocent user). The second attack scenario considers that an adversary tries to modify an existing chain of certificates, distributing it as a proper chain.

In Section 2 of this paper we first define the security assumptions on which the security relies, giving formal definitions of the cryptographic primitives (i.e., hash functions and digital signatures) that act as the building blocks of the protocol, highlighting their security features that will eventually guarantee the security of the whole construction. In other words, the first step is the statement of supposedly intractable problems related to these primitives, which will become the goal of the formal reduction.

In Section 3 we provide a sketch of Muftic's scheme, highlighting the characteristics that are instrumental in its security.

In Section 4 and Section 5, we first proceed to formalise the threat scenarios and their actors, stating their capabilities and goals, in order to build realistic models of the attacks, suitable for formal analysis. This translation of protocol and malicious interaction into a formal language allows the reduction of an effective attack against the protocol to the disruption of the security of well-studied and established cryptographic primitives, such as hash functions and digital signatures.

Finally, we draw our conclusions regarding BIX protocol's resiliency against large-scale cyber attacks.

2. PRELIMINARIES

A. Formal Proofs of Security

In cryptography, the security of a scheme usually relies on the difficulty of a particular mathematical problem. So, in a formal proof of security the goal is to model the possible attacks on the scheme and prove that a successful breach implies the solution of a hard, well-known mathematical problem. Some security parameters may be chosen in such a way that the problem guaranteeing the security becomes almost impossible to solve in a reasonable time, and thus the scheme becomes impenetrable. More formally, the scheme is supposed secure if an *Assumption* holds on the related mathematical problem. Generally, an *Assumption* is that there is no *polynomial-time* algorithm that solves a problem $\mathcal{P}$ with *non-negligible* probability. For example, see the problem in the following subsection on hash functions.

In a formal proof of security of a cryptographic scheme there are two parties involved: a *Challenger* $\mathcal{C}$ that runs the algorithms of the scheme and an Adversary $\mathcal{A}$ that tries to break the scheme making queries to $\mathcal{C}$. In a *query* to $\mathcal{C}$, depending on the security model, $\mathcal{A}$ may request private keys, the encryption of specific plaintexts, and so on. The goal of $\mathcal{A}$ also depends on the security model, for example it may be to recover a key, or to forge a digital signature.

Hence, the security proofs follow a general path. Suppose there is an *Adversary* $\mathcal{A}$ that breaks the scheme with non-negligible probability p_1. A Simulator $\mathcal{S}$ is built such that if $\mathcal{A}$ breaks the scheme then $\mathcal{S}$ solves $\mathcal{P}$. So, given an instance of $\mathcal{P}$, $\mathcal{S}$ runs a challenger $\mathcal{C}$ that interacts with $\mathcal{A}$, simulating the scheme correctly with non-negligible probability p_2. Thus $\mathcal{S}$ solves $\mathcal{P}$ with non-negligible probability, which is usually p_1p_2, contradicting the Assumption.

To summarise, a formal proof of security is a reduction from the problem attack the scheme to the problem solve $\mathcal{P}$. Typically, $\mathcal{P}$ is a well-studied problem, so the assumption on its insolvability is accepted by the academic community.

B. Hash Functions

Commonly, the messages to be signed are compressed in fixed-length binary strings via a cryptographic hash function. A hash function H can be idealised as a function whose set of inputs is the set of all possible binary strings, denoted by $(\mathbb{F}_2)^*$, while its set of possible outputs is the set of all binary strings of given length (called *digest*). Real-life hash functions have a *finite* input set, but it is so large that can be thought of as infinite.

Cryptographic hash functions can need several security assumptions, however for the goals of this paper the following definitions are sufficient.

Definition 1: Collision Problem for a Class of Inputs. Let $r \geq 1$l, h:$(\mathbb{F}_1)^* \rightarrow (\mathbb{F}_2)^{\mathrm{r}}$ be a *hash function*, and $L \subseteq (\mathbb{F}_2)^l$ be a class of inputs. The *collision problem* for h and L consists in finding two inputs m_1,$m_2 \in L$, with $m_1 \neq m_2$, such that $h(m_1) = h(m_2)$.

Definition 2: Collision Resistance of Hash Functions (Assumption 1). Let h be a *hash function*. We say that h is *collision resistant* for a class of inputs L if there is no polynomial-time algorithm $\mathcal{B}(h, L) \rightarrow \{m_1,m_2\}$ that solves the *Collision Problem* for h and L with non-negligible probability. The complexity parameter is r.

C. Digital Signatures and ECDSA

With the name *Digital Signature Scheme*, we refer to any asymmetric cryptographic scheme for producing and verifying digital signatures, consisting of three algorithms:

- *Key Generation* - $\mathsf{KeyGen}(\kappa) \rightarrow (\mathsf{SK}, \mathsf{PK})$: given a security parameter κ generates a public key PK, that is published, and a secret key SK.
- *Signing* - $\mathsf{Sign}(m, \mathsf{SK}) \rightarrow s$: given a message m and the secret key SK, computes a digital signature s of m.
- *Verifying* - $\mathsf{Ver}(m, s, \mathsf{PK}) \rightarrow r$: given a message m, a signature s and the public key PK, it outputs the result $r \in \{\mathsf{True}, \mathsf{False}\}$ that says whether or not s is a valid signature of m computed by the secret key corresponding to PK.

We measure the security of a Digital Signature Scheme by the difficulty of forging a signature in the following scheme (which results in an existential forgery):

Definition 3: Digital Signature Security Game. Let $\mathcal{DSS}$ be a Digital Signature Scheme. Its security game, for an adversary $\mathcal{A}$, proceeds as follows:

1) *Setup*
The challenger $\mathcal{C}$ runs the KeyGen algorithm, and gives to the adversary the public key PK.
2) *Query*
The adversary issues signature queries for some messages m_i the challenger answers giving $s_i = \mathsf{Sign}(m_i, \mathsf{SK})$.
3) *Challenge*
The adversary is able to identify a message $\neq m_i \, \forall i$, and tries to compute s such that $\mathsf{Ver}(m, s, \mathsf{PK}) = \mathsf{True}$. If $\mathcal{A}$ manages to do so, they win.

Definition 4: Security of a Digital Signature Scheme (Assumption 2). A Digital Signature Scheme $\mathcal{DSS}$ is said to be secure if there is no polynomial-time algorithm $\mathcal{A}$ (w.r.t.κ) that wins the *Digital Signature Security Game* with non-negligible probability.

Ideally, a Digital Signature Scheme is designed in such a way that forging a signature in the scheme is equivalent to solving a hard mathematical problem. Although this equivalence is usually assumed but not proved, we say that the Digital Signature Scheme is based on that mathematical problem. Several Digital Signatures Schemes (e.g. [7]), are based on the discrete logarithm problem (DLOG), although other approaches exist, see e.g. [8], [9]. Among them, the Elliptic Curve Digital Signature Algorithm (ECDSA), which uses elliptic curves, is widespread. We refer to [10] for the details about the design of ECDSA.

If an attacker is able to solve the DLOG on an elliptic curve $\mathbb{E}$, then they can break the corresponding ECDSA. The converse is much less obvious. In [12], the authors provide convincing evidence that the unforgeability of several discrete logarithm-based signatures cannot be equivalent to the DLOG problem in the standard model. Their impossibility proofs apply to many discrete logarithm-based signatures like DSA, ECDSA and KCDSA, as well as standard generalisations of these. However, their work does not explicitly lead to actual attacks. Assuming that breaking the DLOG is the most efficient attack on ECDSA, then nowadays recommended key lengths start from 160 bits.

3. A DESCRIPTION OF BIX CERTIFICATES

In this section, we describe the BIX certificates and the structure containing them, called the BIX Certification Ledger (BCL). BIX certificates share many similarities with X.509 certificates, but the identities are anonymous. For a detailed comparison, we refer to [6, Section 2.1]; here we only highlight their characteristics that are instrumental for our security proofs.

The BCL collects all the BIX certificates filling a double-linked list, in which every certificate is linked to the previous and the next. To simplify our notation, we define the BCL as a 'chain of certificates', CC, of n certificates, that we may consider as a sequence:

$$\mathsf{CC}: c_0, \ldots, c_{n-1}.$$

We denote with λ a function returning the length of a chain, that is $\lambda(\mathsf{CC}) = n$. Also, $||$ denotes string concatenation.

TABLE 1. STRUCTURE OF A BIX CERTIFICATE

	Header (H_i)	
	Sequence number Version, Date.	
Issuer (S_{i-1})	Subject (S_i)	Next Subject (S_{i+1})
Bix ID of S_{i-1} Public key (PK$_{i-1}$)	Bix ID of S_i Public key (PK$_i$)	Bix ID of S_{i+1} Public key (PK$_{i+1}$)
Issuer Signature	Subject Signature	Next Subject Signature
Backward cross-signature		
Signature of $(H_i \| \| h(S_{i-1}) \| \| h(S_i))$ by SK$_{i-1}$ Signature of $(H_i \| \| h(S_{i-1}) \| \| h(S_i))$ by SK$_i$		
	Forward cross-signature	
	Signature of $(H_i \| \| h(S_i) \| \| h(S_{i+1}))$ by SK$_i$ Signature of $(H_i \| \| h(S_i) \| \| h(S_{i+1}))$ by SK$_{i+1}$	

Remark 1. The owner of the certificate c_i has a double role: as a **user**, with c_i that certificates their identity; as an **issuer**, providing the certificate c_{i+1} to the next user. In this way, there is no need of a CA.

To each certificate corresponds a user having a pair private key/public key, which we denote with SK_i and PK_i. Certificate c_0 is called the root certificate and certificate c_{n-1} is called the tail certificate.

In this paper, a certificate c_i for $i = 1, \ldots, n-2$ is defined by the following fields (and subfields) (necessary for our proofs of security), while the complete list can be found in [6]. Root and tail certificates are described later on.

1) Header ($\boldsymbol{H}_i$ *)*

In this field, there is general information such as timestamps and protocol version, but the relevant information for our analysis is the **Sequence number** i, that is the identification number of the certificate and also the position with respect to certificates of other BIX members.

2) Subject ($\boldsymbol{S}_i$)

The Subject contains the personal information that identifies the i-th user (S_i), in particular:

- **Subject BIX ID.** The unique global identifier of the user who owns the certificate. All BIX IDs contained in the Subject fields of a valid chain are distinct.
- **Public key.** The cryptographic public key of the owner of the certificate PK_i.

3) Subject signature

This contains the signature over the Subject attributes via the private key SK_i associated to PK_i.

4) Issuer ($\boldsymbol{S}_{i-1}$)

The Issuer field enjoys the same attribute structure as the Subject field, but it identifies the BIX member who certified S_i, i.e., it contains the Subject attributes of c_{i-2}, which identifies S_{i-2} (the previous member in the BCI).

5) Issuer signature

This field contains the signature over the Issuer attributes created by the Issuer, that is, performed via the private key SK_{i-1} associated to PK_{i-1}.

6) Backward cross-signature

The Backward_Cross_Signature contains two signatures, one created by the Issuer S_{i-1} and the other created by the Subject S_i, over the same message: $(H_i||h(S_{i-1})||h(S_i))$. Note that this field guarantees validity of the Header and binding between the Subject and the Issuer.

7) Next Subject ($\boldsymbol{S}_{i+1}$)

The Next_Subject field enjoys the same attribute structure of the Subject field, but it identifies the BIX member who is certified by S_i, i.e., it contains the Subject attributes of c_{i+1}, which identifies S_{i+1} (the next member in the BCI).

8) Next Subject signature

This is the same field as Subject signature, except it is created by the Next Subject over its own data, that is, performed via the private key SK_{i+1} associated to PK_{i+1}.

9) Forward cross-signature

The Forward_Cross_Signature contains two signatures, one created by the Subject S_i and the other created by the Next Subject S_{i+1}, over the same message: $(H_i||h(S_i)||h(S_{i+1}))$.

Note that this field guarantees binding between the current user acting as an issuer and the next user (to whom the next certificate c_{i+1} is issued).

We now describe the special certificates:

- The certificate c_0, called the *root certificate*, has the same structure of a standard certificate, but the Issuer field and the Subject field contain the same data. Indeed, the root user S_0 is not a normal user but rather an entity that initiates the specific BCL.

- The certificate c_{n-1} has the same structure of a standard certificate, but some fields are not populated because the next user is still unknown: Next_Subject, the Next_Subject signature, the Forward_Cross_Signature. However, we underline that it is regularly published in the chain and considered valid by other users.
  The last user that owns the last certificate, c_{n-1} will then become the issuer for the next certificate (see *Remark 1*).

In the BIX protocol a new user requests the issuing of a new certificate through a query to the BIX community, which is processed only by the user that owns the tail certificate of the chain. For further details about the BIX protocol we refer to [6, Section 3.3].

4. CHAIN LENGTHENING ATTACK SCENARIO

The first attack scenario that we consider supposes that an attacker tries to attach their certificate to a pre-existing certificate chain without interacting properly with the last user of the chain. More precisely, the attacker $\mathcal{A}$ should not interact with the subject of the last certificate in the chain according to the BIX protocol.

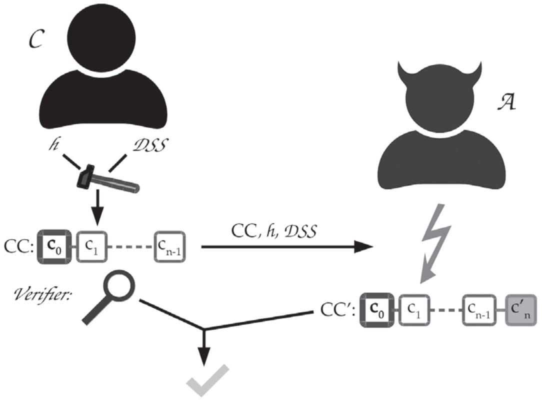
A. The Security Game

For this attack, we consider a game where an adversary $\mathcal{A}$ aims to add a certificate to the tail of a certificate chain CC. We will call it *Static Chain Lengthening (SCL) Game* and it proceeds as follows:

- The challenger $\mathcal{C}$ builds a certificate chain CC according to the BIX protocol with root certificate c_0, using a hash function h and a digital signature scheme $\mathcal{DSS}$.
- $\mathcal{C}$ passes to $\mathcal{A}$ the chain CC together with h and $\mathcal{DSS}$.
- $\mathcal{C}$ builds an honest verifier $\mathcal{V}$ that given a certificate c^* and a certificate chain CC* outputs True if the root certificate of CC* is c_0 and c^* is a valid certificate of CC*, False otherwise.
- $\mathcal{A}$ tries to build a forged certificate chain CC′, $\lambda(\mathrm{CC}') = n + 1$, such that:
  - o CC′ truncated before the last certificate c'_n is identical to CC if the Next_Subject and Forward_Cross_Signature fields of the second-to-last certificate of CC′ are not considered (i.e. we obtain CC′ by adding a certificate to CC and completing c_{n-1} accordingly);
  - o user S_{n-1} did not take part in the creation of c'_n and so in particular they did not perform the Forward_Cross_Signature of c_{n-1} and the Backward_Cross_Signature of c'_n;
  - o $\mathcal{V}(c', \mathrm{CC}') = \mathsf{True}$ where c'_n is the last certificate of CC′.

$\mathcal{A}$ wins the SCL Game if they build a CC′ that satisfies these last three points.

FIGURE 1. SCL GAME


$\mathcal{C}$
$\mathcal{A}$
$\mathcal{h}$
$\mathcal{DSS}$
CC: c_0 c_1 c_{n-1}
CC, $\mathcal{h}$, $\mathcal{DSS}$
Verifier:
CC′: c_0 c_1 c_{n-1} c'_n

Definition 5: Security against SCL. The BIX protocol is said to be **secure against static chain lengthening** if there is no adversary $\mathcal{A}$ that in polynomial time wins the *SCL Game* with non-negligible probability.

B. Security Proof

In the following we will prove that winning the SCL game implies the contradiction of our security assumptions introduced in Section 2.

Theorem 1. Let $\mathcal{A}$ be an adversary that wins the *SCL Game* with probability ϵ. Then a simulator $\mathcal{S}$ might be built that, with probability at least ϵ, either solves the *Collision Problem*, with L the set of all possible Subject fields, or wins the *Digital Signature Security Game*.

Proof. Let $\mathcal{DSS}$ be the digital signature scheme and h the hash function used in the BIX protocol, and $L \subseteq (\mathbb{F}_2)^l$ be the class of all possible Subject fields. We will build a simulator $\mathcal{S}$ that simultaneously plays the *Digital Signature Security (DSS) Game* and tries to solve an instance of the *Collision Problem* for L. It does so by simulating an instance of the *SCL Game* and exploiting $\mathcal{A}$. We will prove that if wins the SCL Game then either $\mathcal{S}$ finds a solution for the Collision Problem or $\mathcal{S}$ wins the DSS Game.

$\mathcal{S}$ starts with taking as input an instance (h, L) of the Collision Problem and a public key PK^* given by the $\mathcal{DSS}$ challenger (i.e., the output of the first phase of the DSS Game for the scheme $\mathcal{DSS}$). $\mathcal{S}$ then proceeds to build a certificate chain CC^* following the BIX protocol. $\mathcal{S}$ builds all but the last certificate normally, running the KeyGen algorithm of the $\mathcal{DSS}$ to choose public keys for the Subject fields, so the corresponding secret keys are available to sign these certificates properly. Then let $n = \lambda(\mathsf{CC})^* \geq 2$ (i.e. the number of certificates contained in CC^*), c_0^* its root certificate and c_{n-1}^* the last one. $\mathcal{S}$ sets the Subject of c_{n-1}^*, that we will denote by

S^*_{n-1}, such that its public key is PK^*, then it queries the challenger of the DSS Game to obtain three valid signatures, respectively, on:

- the hash $h(S^*_{n-1})$ of this subject,
- $(H^*_{n-1}||h(S^*_{n-2})||h(S^*_{n-1}))$ for the Backward_Cross_Signature of c^*_{n-1},
- $(H^*_{n-2}||h(S^*_{n-2})||h(S^*_{n-1}))$ for the Forward_Cross_Signature of c^*_{n-2},

where H^*_{n-2} is the Header of c^*_{n-2}, H^*_{n-1} is the Header of c^*_{n-1}, and $h(S^*_{n-2})$ is the hash of the Issuer of c^*_{n-1}, that is the Subject of c^*_{n-2}. In this way $\mathcal{S}$ completes a certificate chain CC* of length n, that it passes to $\mathcal{A}$.

$\mathcal{A}$ responds with a counterfeit chain CC' of length $\lambda(\mathrm{CC}) = n + 1$. If CC' is not valid (the chains CC' and CC* do not correspond up to the n-th certificate, or an integrity check fails) then $\mathcal{S}$ discards this answer and gives up ($\mathcal{S}$ fails).

Otherwise, if the verifier outputs True, the chain CC' is valid. Denote by l' the string $(H_n'||h(S_{n-1}')||h(S_n'))$ signed in the Backward_Cross_Signature of c'_n (the last certificate of CC') by the private key corresponding to PK^*. We have two cases:

- l' is equal to a message for which $\mathcal{S}$ requested a signature.
  Because of its bit-length, l' may be equal to $l^*_0{:=}(H^*_{n-2}\ ||h(S^*_{n-2})||h(S^*_{n-1}))$ or $l^*_1{:=}(H^*_{n-1}\ ||h(S^*_{n-2})||h(S^*_{n-1}))$, but not to $h(S^*_{n-1})$. In either case, $l'{=}l^*_0$ or $l'{=}l^*_1$, the equality implies that $h(S_n'){=}h(S^*_{n-1})$, but the specification of the BIX protocols supposes that different certificates have a different BIX ID in the Subject (and we know that CC' is valid). So $S'_{n-1} = S^*_{n-1} \neq S'_n$, because of the BIX ID's, but they have the same hash so $\mathcal{S}$ may submit (S^*_{n-1}, S'_n) as a solution to the Collision Problem.
- l' is different from all messages for which $\mathcal{S}$ requested a signature.
  In the Backward_Cross_Signature of c'_n there is a signature s of l' such that $\mathsf{Ver}(l', s, \mathsf{PK}*) = \mathsf{True}$ (remember that PK^* is the public key of the Issuer of c'_n and that CC' is considered valid, so the signatures check out), so $\mathcal{S}$ may submit (l', s) as a winning answer of the challenge phase of the DSS Game.

So, if $\mathcal{S}$ does not fail, it correctly solves the Collision Problem or wins the DSS Game, and since $\mathcal{A}$ is a polynomial-time algorithm, $\mathcal{S}$ is a polynomial-time algorithm too, given that the other operations performed correspond to the building of a certificate chain and this must be efficient. $\mathcal{S}$ might fail only if the chain given by $\mathcal{A}$ is not valid (i.e. if $\mathcal{A}$ fails). Since the simulation of the SCL Game is always correct, $\mathcal{A}$'s failure happens with probability $1 - \epsilon$, then the probability that $\mathcal{S}$ wins is $1 - (1 - \epsilon) = \epsilon$.

Corollary 1: *SCL Security*. If the DSS is secure *(Assumption 1)* and the hash function is collision resistant for the class L *(Assumption 2)*, where L is the set of all possible Subject fields, then the BIX protocol is secure against the Static Chain Lengthening.

Proof. Thanks to *Theorem 1*, given a polynomial-time adversary that wins the SCL Game with non-negligible probability ϵ, a polynomial-time simulator might be built that with the same probability either solves the *Collision Problem* or wins the *DSS Game*. So, let C be the event 'solution of the Collision Problem' and D be the event 'victory at the DSS Game'. We have that

$$\epsilon = P(C \vee D) \leq P(C) + P(D)$$

The sum of two negligible quantities is itself negligible, so the fact that ϵ is non-negligible implies that at least one of $P(C)$ and $P(D)$ is non-negligible, and this means that *Assumption 1* or *Assumption 2* is broken.

Remark 2. The infeasibility of the above attack guarantees also the non-repudiation property of the last certificate in the chain. That is, if Alice (the user of the second-to-last certificate) tries to repudiate Bob (the user of the last certificate), with an eye to issuing another certificate, then Bob might claim his rightful place showing a version of the chain containing his certificate. This chain is then the proper one, since no one can attach its certificate to the tail of the certificate chain without being a proper user.

Remark 3. We have assumed (see *Assumption 2*) that $\mathcal{DSS}$ is secure against existential forgery, but in the proof of Theorem 1 the freedom of the attacker in the choice of the message to be signed is limited. In fact, it has to forge a signature of $l' := (H_n' || h(S_{n-1}') || h(S_n'))$, where $h(S_{n-1}')$ is given by $\mathcal{S}$, and even H_n' is not completely controlled by $\mathcal{A}$ (e.g. the sequence number is given). So, a large part of the string to be signed is beyond the control of the forger, hence the challenge is something in between an existential and a universal forgery, which is the weaker assumption on the Digital Signature Scheme ($\mathcal{DSS}$). However, the security of $\mathcal{DSS}$ against universal forgery is not sufficient for our purposes, so we settled on the stronger assumption.

5. CERTIFICATE TAMPERING

In the second attack scenario that we consider, a malicious attacker tries to corrupt a chain of certificates built on a trusted root certificate, resulting in another chain that may redistribute as a proper chain with the same root but with altered information. As we will see, the security against this attack would guarantee that no external attacker can modify any certificate in the chain, including deleting or inserting a certificate in any non-ending point, as long as the root certificate is safe (no unauthorised use), secure (cannot be broken) and public (anyone can check it). If the security proved in the previous section is also considered, then a certificate chain is also secure at the end point (no one can wrongfully insert themself at the end or disavow the last certificate) achieving full security from external attacks to the BIX protocol.

A. The Security Game

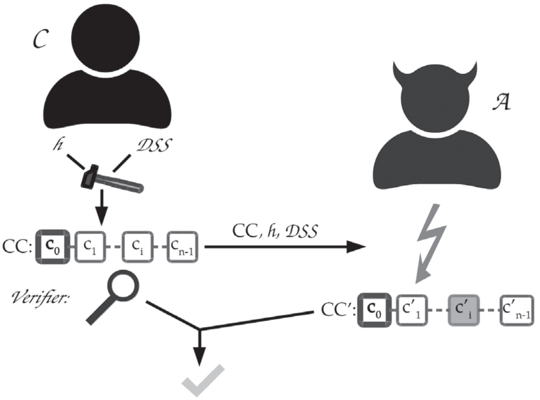
For this attack, we consider a game where an adversary $\mathcal{A}$ aims to modify information contained in the *Subject* field of a certificate c_i contained in a certificate chain $\mathtt{CC}$, with $1 \leq i \leq n-1$,

$n = \lambda(\mathrm{CC})$. We will call it the *Static Tampering with Subject (STS) Game* and it proceeds as follows:

- The challenger $\mathcal{C}$ builds a certificate chain CC with root certificate c_0, according to the BIX protocol and using a hash function h and a Digital Signature Scheme $\mathcal{DSS}$. Let $n = \lambda(\mathrm{CC})$.
- $\mathcal{C}$ passes to $\mathcal{A}$ the chain CC together with h and $\mathcal{DSS}$.
- $\mathcal{C}$ builds an honest verifier $\mathcal{V}$ that, given a certificate c^* and a certificate chain CC^* outputs True if the root certificate of CC^* is c_0 and c^* is a valid certificate of CC^*, False otherwise.
- $\mathcal{A}$ tries to build a forged certificate chain CC' such that:
  - exists $1 \leq i \leq n-1$ such that Subject fields of c_i and c'_i are different, that is, $S_i \neq S_i'$.
  - $\mathcal{V}(c_i', \mathrm{CC}') = \mathsf{True}$

$\mathcal{A}$ wins the *STS* Game if they successfully build such a CC' satisfying the last two items.

FIGURE 2. STS GAME



Definition 6: Security against STS. The BIX protocol is said secure against Static Tampering with Subject if there is no adversary $\mathcal{A}$ that in polynomial time wins the *STS Game* with non-negligible probability.

B. Security Proof

In the following, we will prove that winning the SCL game implies the contradiction of our security assumptions introduced in Section 2.

Theorem 2. Let $\mathcal{A}$ be an adversary that wins the *STS Game* with probability ϵ. Then a simulator $\mathcal{S}$ might be built that with probability at least $\frac{\epsilon}{n-1}$ either solves the *Collision Problem*, where L is the set of all possible Subject fields, or wins the *Digital Signature Security Game*, where n is the length of the certificate chain that $\mathcal{S}$ gives to $\mathcal{A}$.

Proof. Let $\mathcal{DSS}$ be the Digital Signature Scheme and h the hash function used in the BIX protocol, and $L \subseteq (\mathbb{F}_2)^l$ be the class of all possible Subject fields. We will build a simulator $\mathcal{S}$ that simultaneously plays the digital signature security (DSS) Game and tries to solve an instance of the Collision Problem for L. It does so by simulating an instance of the STS Game and exploiting $\mathcal{A}$. We will prove that when $\mathcal{A}$ wins the STS Game, at least one in $n-1$ times $\mathcal{S}$ is successful. To be more precise, if $\mathcal{S}$ does not find a solution for the collision problem then $\mathcal{S}$ wins the DSS Game.

$\mathcal{S}$ starts with taking as input an instance (h, L) of the Collision Problem and a public key PK^* given by the $\mathcal{DSS}$ challenger (i.e., the output of the first phase of the Digital Signature Security Game for the scheme $\mathcal{DSS}$).

$\mathcal{S}$ now proceeds to build a certificate chain CC following the BIX protocol, as follows. First, $\mathcal{S}$ chooses $n \geq 2$ (possibly depending on the $\mathcal{A}$'s requirements). Then $\mathcal{S}$ selects $1 \leq k \leq n-1$ at random to be the index of a certificate c_k in CC. $\mathcal{S}$ builds the first $k-1$ certificates normally, running the KeyGen algorithm of the $\mathcal{DSS}$ scheme to choose public keys for the Subject fields, so the corresponding secret keys are available (to $\mathcal{S}$) to sign these certificates properly. So $c_0,\ldots, c_{k-3}$ are complete certificate and c_{k-2} is a tail certificate. Then it sets the Subject of c_{k-1} such that its public key is PK^*, and a header H_{k-1}. It queries the challenger of the DSS Game to obtain three valid signatures, respectively, on:

- the hash $h(S_{k-1})$ of this subject,
- $(H_{k-1}||h(S_{k-2})||h(S_{k-1}))$ for the Backward_Cross_Signature of c_{k-1} (if $k > 1$),
- $(H_{k-2}||h(S_{k-2})||h(S_{k-1}))$ for the Forward_Cross_Signature of c_{k-2} (if $k > 1$),

where we recall that H_{k-2} is the Header of c_{k-2}, H_{k-1} is the Header of c_{k-1}, $h(S_{k-2})$ is the hash of the Issuer of c_{k-1}. Then $\mathcal{S}$ builds the $k+1$-th certificate, choosing a H_k and S_k, using again the KeyGen algorithm to sign S_k, querying the DSS challenger for two valid signatures, respectively, on:

- $(H_k||h(S_{k-1})||h(S_k))$ for the Backward_Cross_Signature of c_k,
- $(H_{k-1}||h(S_{k-1})||h(S_k))$ for the Forward_Cross_Signature of c_{k-1},

where we recall that H_k is the Header of c_k and $h(S_k)$ is the hash of the Subject of c_k. Finally, $\mathcal{S}$ completes the chain CC (following the protocol and choosing everything, including the SK_i's), so that it has n certificates, and passes it to $\mathcal{A}$.

$\mathcal{A}$ responds with a counterfeit chain CC$'$. $\mathcal{A}$ fails if and only if CC$'$ is not valid, which happens when there is no $1 \leq i \leq n-1$ such that $S'_i \neq S_i$ or when the integrity check of the verifier fails.

If we are in this situation, $\mathcal{S}$ discards CC' and gives up ($\mathcal{S}$ fails).

Otherwise, let $1 \leq i \leq n-1$ be the first index for which $S'_i \neq S_i$. Since k is chosen at random, we have that $k = i$ with probability $\frac{1}{n-1}$. In this case, there are two possibilities:

- $h(S_k) = h(S'_k)$, but $S_k \neq S_k'$ for hypothesis, then $\mathcal{S}$ outputs the pair (S_k, S'_k) as a solution to the collision problem.
- Otherwise, we have that $S_{k-1} = S_{k-1}'$ and PK^* is the public key of the issuer of c_k'. Then in the Backward_Cross_Signature of the certificate c_k' there is the digital signature s for which holds the relation $\mathrm{Ver}((H_k' || h(S_{k-1}') || h(S_k')), s, \mathrm{PK}^*) = \mathrm{True}$ (remember that CC' is considered valid, so the signatures check out). So $\mathcal{S}$ may submit $((H_k' || h(S_{k-1}') || h(S_k')), s)$
- as a winning answer of the challenge phase of the DSS Game, since it is different from the messages $\mathcal{S}$ queried for signatures, that are $[h(S_{k-1}), (H_{k-1} || h(S_{k-2}) || h(S_{k-1})), (H_{k-2} || h(S_{k-2}) || h(S_{k-1})),$ $(H_k || h(S_{k-1}) || h(S_k)), (H_{k-1} || h(S_{k-1}) || h(S_k))]$

So $\mathcal{S}$ correctly solves the collision problem or wins the DSS Game at least when $\mathcal{A}$ wins and $i{=}k$. The probability of this event is at least the probability of the two cases and so it is

$$\epsilon \cdot \frac{1}{n-1} = \frac{\epsilon}{n-1}$$

Note also that since $\mathcal{A}$ is a polynomial time algorithm, so is $\mathcal{S}$.

Corollary 2: STS Security. If the DSS is secure (see *Assumption 2*) and the hash function is collision resistant for the class L (*Assumption 1*) where L is the set of all possible Subject fields, then BIX protocol is secure against the Static Tampering with Subject.

Proof. For the BIX protocol to be functional the length of the chain must be polynomial. So, for the result of *Theorem 2*, given a polynomial time adversary that wins the STS Game with non-negligible probability ϵ, a polynomial time simulator might be built that with probability at least $\frac{\epsilon}{n-1}$ either solves the Collision Problem or wins the DSS Game, where n is the length of the chain. So, let C be the event 'solution of the Collision Problem' and D be the event 'victory at the DSS Game'. We have that

$$\frac{\epsilon}{n-1} \leq P(C \vee D) \leq P(C) + P(D)$$

The fact that $\frac{\epsilon}{n-1}$ is non-negligible implies that at least one of $P(C)$ and $P(D)$ is non-negligible, and this means that *Assumption 1* or *Assumption 2* is broken.

6. CONCLUSIONS

In this paper, the BIX certificates protocol proposed in [6] has been formally analysed from a security point of view. In particular, the security against static attacks that aim to corrupt a chain has been proven, reducing the security to the choice of adequate hash function and digital signature scheme. For this reason, the security of ECDSA, the main DSS nowadays, has also been discussed.

The current BIX protocol is still insufficiently complete for it to be considered a full PKI. Possibly, the main lack is the absence of a procedure to revoke or to renew certificates. This is an open problem and further research effort is needed. However, the security proofs given in this paper show how the BIX infrastructure is a reliable structure for storing public identities in a distributed and decentralised way. While a targeted attack on a CA can result in the issuing of malicious certificates or revocation of valid ones, shattering every certificate it issued, in the case of a cyber attack BIX certificates could still be trusted because no single entity could be targeted and exploited to take down the entire system. Indeed, we suppose that BIX certificates are issued and distributed in *peacetime*, so that when an emergency breaks out the infrastructure is ready to cope with possible attacks. Indeed, the properties proven in this work guarantee the integrity of the information contained in the certificate chain, so users can rely upon it even in the middle of a cyber attack. It is true that a targeted offensive against the owner of the last certificate would disrupt the protocol, preventing the issuing of new certificates. Nevertheless, if this user is taken down, the validity of existing certificates will still hold.

It may seem that the BIX protocol relies on a trusted third party, the messaging system. However, it is not a third party, as highlighted by Muftic [6], since it only passively broadcasts certificates and for purely addressing purposes it verifies the uniqueness of BIX identifiers (in Muftic's construction the IM protocol used is [13]).

Our conclusion is that a PKI system based on the BIX protocol is more resilient to a wide-scale cyber attack than the standard PKI protocols based on CAs.

Regarding related research, the idea of using a public ledger for digital identities has prominent applications in the distribution of Bitcoin wallet addresses (see for example [14]), but there are also applications that try to leverage the functionalities of cryptocurrencies to improve PKI. For example, Matsumoto and Reischuk [15] exploit smart contracts on Ethereum to deter misbehaviour of CAs and to incentive extended vigilance over CAs' behaviour. However, we fear that this is not sufficient in case of a large-scale cyber attack, because the financial losses that this solution enforces would affect the attacked CA and not the attackers themselves.

REFERENCES

[1] S. Wilkinson et al., *Storj: A Peer-to-Peer Cloud Storage Network*, 2014 [Online]. Available: http://storj.io/storj.pdf.

[2] M. Araoz, *Proof of Existence*, 2015 [Online]. Available: https://proofofexistence.com/about.

[3] Ethereum Foundation, *Ethereum Project*, 2015 [Online]. Available: https://www.ethereum.org/.

[4] D. Cooper et al., ‘Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile’, IETF RFC 5280, 2008.

[5] A. Niemann, and J. Brendel, *A survey on CA compromises*, 2013 [Online]. Available: https://www.cdc.informatik.tu-darmstadt.de/fileadmin/user upload/Group CDC/Documents/Lehre/SS13/Seminar/CPS/cps2014 submission (Vol. 8).

[6] S. Muftic, ‘Bix certificates: Cryptographic tokens for anonymous transactions based on certificates public ledger’, Ledger, vol. 1, pp. 19–37, 2016.

[7] T. Elgamal, ‘A public key cryptosystem and a signature scheme based on discrete logarithms’, *IEEE Trans. on Inf. Th.*, vol. 31, no. 4, pp. 469–472, 1985.

[8] M. O. Rabin, ‘Digital signatures and public-key functions as intractable as factorization’, MIT laboratory for computer science, MIT/LCS/TR-212, Jan. 1979.

[9] R. L. Rivest, A. Shamir, and L. M. Adleman, ‘A Method for Obtaining Digital Signatures and Public-Key Cryptosystems’, *Commun. ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[10] D. Johnson, A. Menezes, and S. Vanstone, ‘The Elliptic Curve Digital Signature Algorithm (ECDSA)’, *Certicom*, 1998.

[11] B. Preneel, ‘The State of Cryptographic Hash Functions’, *LNCS*, vol. 1561, pp. 158–182, 1999.

[12] P. Paillier and D. Vergnaud, ‘Discrete-Log-Based Signatures May Not Be Equivalent to Discrete Log’, LNCS, vol. 3788, pp. 11–20, 2005.

[13] XMPP Standards Foundation, *Extensible Messaging and Presence Protocol*, 2015 [Online]. Available: https://www.xmpp.org/.

[14] BitID, *BitID Open Protocol*, 2015 [Online]. Available: http://bitid.bitcoin.blue/

[15] S. Matsumoto and R. M. Reischuk, *IKP: Turning a PKI Around with Blockchains*, 2016 [Online]. Available: https://eprint.iacr.org/2016/1018.pdf.

BIOGRAPHIES

Editors and Co-Editors

Cpt **Raik Jakschis** is a member of the NATO CCD COE Technology Branch and currently focuses on cyber security research into ICS/SCADA systems. Prior to his post at NATO CCD COE, Raik worked at the Bundeswehr Communications and Information Systems Service Centre (BwCISSC) to enhance and secure the IT infrastructure of the German Armed Forces. He holds a Master's degree in Information Technology from Helmut-Schmidt-University, University of the Bundeswehr, Hamburg.

Lauri Lindström has been a Researcher at NATO CCD COE since May 2013. Prior to joining NATO CCD COE he worked at the Estonian Ministry of Foreign Affairs (2007-2012) as the Director General of Policy Planning and held various positions at the Ministry of Defence (1995-2007), dealing mainly with issues related to international cooperation, Estonia's accession to NATO, defence planning and security policy. Lauri Lindström holds a PhD from the Tallinn University, Estonia.

Tomáš Minárik is a Researcher at the NATO CCD COE's Law & Policy Branch. He holds a law degree from the Charles University in Prague. He worked as a legal adviser at the International Law Department of the Czech Ministry of Defence, and later at the National Cyber Security Centre of the Czech Republic. His current research focuses on the legal aspects of active cyber defence, the right to privacy, anonymity networks, and activities of international organisations regarding cyberspace.

Mauno Pihelgas is a Researcher at the Technology branch of the NATO CCD COE, where his area of expertise is monitoring and situational awareness. His prior experience includes 5 years as a monitoring administrator and developer for the largest telecoms operator in Estonia. In addition to being a GIAC GMON Continuous Monitoring Certified Professional, he is also a Red Hat Certified System Administrator and a Red Hat Certified Engineer. Mauno holds a Master of Science degree in Cyber Security, and is a 3rd year PhD candidate researching log analysis, data mining and machine learning.

LtCol **Nikolaos Pissanidis** (GWAPT) is a Greek Army officer with more than 20 years of professional experience in the field of IT and IT security. Before his current assignment as a researcher and web security expert at NATO CCD COE's Research and Development Technology Branch, he worked in several different national management and technical positions focusing on information technology, software development, cyber security and web penetration testing. Besides a diploma from the Hellenic Army Academy, Niko holds a Master's degree in New Technologies in Informatics and Telecommunications from the Department of Informatics and Telecommunications at the National and Kapodistrian University, Athens.

Henry Rõigas is a Researcher in the Law and Policy Branch at the NATO CCD COE. His research in the Centre focuses mainly on the political aspects of international cyber security.

Henry was the co-editor of the book *International Cyber Norms: Legal, Policy & Industry Perspectives* and project manager of the book *Cyber War in Perspective: Russian Aggression against Ukraine*. He is also responsible for managing the agenda for the 9th International Conference on Cyber Conflict (CyCon 2017) and for editing the proceedings book. Henry holds a Master's degree in International Relations from the University of Tartu.

Authors

Giovanni Apruzzese received a Master's degree in computer engineering from the University of Modena and Reggio Emilia, Italy, in 2016, where he is currently working towards a PhD at the International Doctorate School in Information and Communication Technologies (ICT). His research interests include security analytics and network science.

Brad Bigelow is the Principal Technical Advisor to the Deputy Chief of Staff, Communications and Information Systems and Cyber Defence (DCOS CCD), Supreme Headquarters Allied Powers Europe (SHAPE). He has extensive experience in network and information systems, intelligence, satellite operations, information warfare, cyber defence, 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 (NSTAC), where he led assessments of the information security practices of the power and financial services industries.

LtCol **Jeffrey Biller** is a Judge Advocate in the United States Air Force assigned as a Military Professor to the Stockton Center for the Study of International Law at the US Naval War College in Newport, Rhode Island. His previous Air Force positions included assignment as the Staff Judge Advocate for the Air Force's two operational cyberspace wings and the Deputy Staff Judge Advocate for the Air Force Intelligence, Surveillance and Reconnaissance Agency. Prior to service as a Judge Advocate, LtCol Biller was an Air Force intelligence officer. He received his JD from the University of Kansas and has a LLM in National Security Law from George Washington University.

Dr **Michele Colajanni** has been Professor of Computer Engineering at the University of Modena and Reggio Emilia since 2000. He received a Master's degree in Computer Science from the University of Pisa, and a PhD in Computer Engineering from the University of Rome in 1992. He manages the Interdepartmental Research Centre on Security and Safety (CRIS), and the Master's in Information Security, Technology and Law. His research interests include security of large-scale systems, performance and prediction models, and web and cloud systems.

Dr **Kenneth Geers** (PhD, CISSP) is Comodo Senior Research Scientist, NATO CCD COE Ambassador, Atlantic Council Senior Fellow, Digital Society Institute-Berlin Affiliate, and Visiting Professor at Taras Shevchenko National University of Kyiv in Ukraine. Dr Geers spent 20 years in the US Government, with time in the US Army, NSA, NCIS, and NATO, and was Senior Global Threat Analyst at FireEye. He is the author of *Strategic Cyber Security*, editor of *Cyber War in Perspective: Russian Aggression against Ukraine*, editor of *The Virtual*

Battlefield: Perspectives on Cyber Warfare, technical expert to the *Tallinn Manual*, and author of many articles and chapters on cyber security.

Alessandro Guido is a PhD candidate at the University of Modena and Reggio Emilia, Italy. He received a Master's degree in Computer Engineering from the same University in 2012. His research interests include network security and all aspects of information security.

Dr **Robert Koch** is an IT staff officer of the Federal Armed Forces and a research assistant in the Department of Computer Science at Universität der Bundeswehr, München, and member of the University's Research Centre for Cyber Defence (CODE). He received his PhD in 2011 and is now a senior research assistant and lecturer in Computer Science. His main areas of research are network and system security with focus on intrusion and extrusion detection in encrypted networks, security of COTS products, security visualisation and the application of artificial intelligence. He has several years of experience in the operation of high security networks and systems, and his research papers have been published in many international conferences and journals. He serves as programme chair and member of the technical program committee for numerous conferences.

Teo Kuhn is an officer in the German Federal Armed Forces, who achieved his MSc in Computer Sciences with the focus on cyber defence at the Universität der Bundeswehr, München. He gained practical experience during internships at renowned German companies, e.g. secunet Security Networks AG. Driven by his interests in IT security and IT forensics, he developed a current-based IDS for SCADA and ICS, which aroused global interest and received positive feedback. He will soon be going to work for the Cyber and Information Space Command of Germany's military.

Dr **Vincent Lenders** is at armasuisse, where he works as cyber research director for the Swiss Federal Department of Defence. He earned both his MSc (2001) and PhD (2006) in Electrical Engineering at ETH Zürich, Switzerland. He was a postdoctoral research faculty member at Princeton University, N.J., USA. Dr Lenders is the co-founder and on the board of the OpenSky Network and Electrosense associations. His current research interests are in the fields of cyber security, information management, big data and crowdsourcing. He is the author of more than 90 publications that have appeared in renowned international journals and peer-reviewed conferences.

Dr **Martin Libicki** (PhD., U.C. Berkeley 1978) holds the Keyser Chair of Cybersecurity Studies at the US Naval Academy. He is the author of a 2016 textbook on cyberwar, *Cyberspace in Peace and War*, as well as two other commercially published books: *Conquest in Cyberspace: National Security and Information Warfare*, and *Information Technology Standards: Quest for the Common Byte*. He is also the author of numerous RAND monographs, notably *Defender's Dilemma, Brandishing Cyberattack Capabilities, Crisis and Escalation in Cyberspace, Cyberdeterrence and Cyberwar, How Insurgencies End* (with Ben Connable), and *How Terrorist Groups End* (with Seth Jones). His prior employment includes 12 years at NDU.

Ricardo Longo obtained a Bachelor's degree in Mathematics in 2012 and a Master's degree in Mathematics in 2014, attending a curriculum in Coding Theory and Cryptography, both at the University of Trento, Italy. He is now a PhD candidate in Mathematics at the same university. For his thesis, he is working on formal proofs of security of cryptographic protocols and schemes. His research interests are focused on: pairing-based cryptography, especially attribute-based encryption, tokenisation algorithms, cryptocurrencies and blockchain technologies, zero-knowledge proofs of knowledge, and complexity theory.

Dr **Kubo Mačák** is a Senior Lecturer in Law at the University of Exeter (United Kingdom). He holds a DPhil in international law from the University of Oxford. Kubo has held research positions at the Universities of Bonn (Germany), Haifa (Israel), and Wuhan (China). He has worked at the United Nations ad hoc tribunals for the former Yugoslavia and Rwanda. He is currently serving as a core expert in the International Humanitarian Law Group of the Manual on International Law Applicable to Military Uses of Outer Space (MILAMOS) project. His research interests span the law of cyber security, international humanitarian law, and general international law.

Dr **Mirco Marchetti** is a researcher at the Department of Engineering 'Enzo Ferrari' of the University of Modena and Reggio Emilia (Italy). He received a PhD in Information and Communication Technologies in 2009. His research interests include all aspects of system and network security, security for cyber-physical systems and automotive, cryptography applied to cloud security, and outsourced data and services.

Dr **Ivan Martinovic** is an Associate Professor at the Department of Computer Science, University of Oxford. Before coming to Oxford he was a postdoctoral researcher at the Security Research Lab, UC Berkeley (2011) and at the Secure Computing and Networking Centre, UC Irvine (2009/2010). From 2009 until 2011 he enjoyed a Carl-Zeiss Foundation Fellowship and he was an associate lecturer at TU Kaiserslautern, Germany. He obtained his PhD from TU Kaiserslautern under supervision of Prof. Jens B. Schmitt and MSc from TU Darmstadt, Germany.

Dr **Fabio Pierazzi** is a Research Assistant working on big data security analytics at the University of Modena and Reggio Emilia, Italy. At the same institute, he completed a Master's degree in Computer Engineering in 2013 and a PhD in Computer Science in 2017. During 2016, he spent 10 months as a visiting research scholar at the University of Maryland, USA, working with Prof V. S. Subrahmanian on machine learning for malware classification and cyber deception. His research interests focus on security analytics applied to network security, intrusion detection, malware classification, and in general on all solutions that combine cybersecurity and analytics.

Dr **Federico Pintore** obtained a Master's degree in Mathematics in 2011 at the University of Cagliari. He received a PhD in Mathematics in 2015 at the University of Trento, with a thesis on number theory and elliptic curves. He is currently a research fellow in the Department of Mathematics at the University of Trento, supervising a project on cryptocurrencies and electronic payment cards. His research interests are focused on: number theory (class field theory, integral binary quadratic forms); elliptic curve cryptography (point counting problem, attacks on ECDLP, EC index-calculus, pairing-based cryptography); cryptocurrencies; and security proofs.

Dr **Giancarlo Rinaldo** is an Assistant Professor in the University of Trento. He has taught a dozen courses in the area of algebra and symbolic computation. His research interests cover different areas of mathematics: commutative algebra, homological algebra, combinatorics, symbolic computation and applications (e.g. vertex covering and cutpoints search on graphs, groebner bases computation, resolution of modules, coding theory and cryptography). He has published more than 30 papers on those subjects. He also has a good knowledge of many Information Technology aspects: from low level programming (c/c++) to web protocols.

Dr **Massimiliano Sala** received an MSc degree in 1995 from the University of Pisa, Italy, and a PhD in 2001 from the University of Milan, Italy. From 2002 to 2006 he was a Senior Research Fellow with the Boole Research Centre in Informatics, University College Cork, Ireland. He is currently a Full Professor at the Department of Mathematics, University of Trento, where he founded the Laboratory of Cryptography (CryptoLabTN), which he is still leading. His research interests are in computational algebra, algebraic coding theory, algebraic cryptography, Boolean functions, and their mutual interactions. His recent industrial collaborations focus on blockchain technology and e-payment security.

Matthias Schäfer is a PhD candidate in the Department of Computer Science at the University of Kaiserslautern, Germany, where he also received his MSc degree in Computer Science in 2013. Between 2011 and 2013, he worked for the Information Technology and Cyberspace group of armasuisse, Switzerland and was a visiting researcher at the Department of Computer Science of Oxford University. He is also a co-founder and board member of the OpenSky Network association and managing director of SeRo Systems GmbH.

Tassilo Singer was Research Associate and lecturer at the University of Passau (2012-2013; 2015-2017) and at the European University Viadrina (2013-2015). He has also been a guest lecturer at a number of universities across Europe. He has published several works in the field of international law and in particular concerning the *ius ad bellum* and the law of armed conflict. Since October 2013 he has been a PhD candidate under the auspices of Prof. Wolff Heintschel von Heinegg and writes his thesis with the working title: *Dehumanization of Warfare – Challenges for International Law*. His current research focus is on modern weapon technologies such as unmanned systems, autonomy and cyber warfare with view to international law.

Ido Sivan-Sevilla is a computer and social scientist. He is currently completing his PhD at the Hebrew University of Jerusalem, on the dynamics between security and privacy in cyberspace. Ido completed his BA in Computer Science at the Technion, Israel's Institute of Technology, and served as a Captain in the Israeli Air Force (IAF), leading its operational cyber-security team. He has experience from the private sector as a product manager, and also at government level, as he joined Israel's Prime Minister's Office to establish its cyber security unit. He completed his MA in Public Policy through the US State Department Fulbright programme, and worked as a legislative assistant in the US Congress. Beyond his PhD research, he also teaches on the Israeli Cadets Programme and serves as a Research Fellow at the Institute for National Security Studies (INSS), through which he is writing a book about cybersecurity regulation in the civic sector.

Max Smeets is a College Lecturer in politics at Keble College, University of Oxford, and Research Affiliate of the Oxford Cyber Studies Programme. He is pursuing a DPhil in International Relations at the University of Oxford, St. John's College. His expertise focuses on the causes underlying cyber proliferation and restraint. Max was a Carnegie Visiting Scholar at Columbia University SIPA and a Doctoral Visiting Scholar at Sciences Po CERI. He holds an undergraduate degree from University College Roosevelt, Utrecht University, and an MPhil in International Relations from the Brasenose College, University of Oxford.

Matthew Smith is a DPhil candidate at the Centre for Doctoral Training in Cyber Security in the Department of Computer Science, University of Oxford. His work focuses on the security and privacy challenges faced by avionic data link systems used for air traffic control and operations, and how these might be mitigated in the short-to-medium term. Before studying at Oxford, he received an MEng in Computer Science from the University of Warwick.

Peter Stockburger is a Senior Managing Associate with Dentons, the world's largest law firm. Dentons employs approximately 7,600 lawyers and professionals in more than 140 locations across 57 countries. Peter is a member of the firm's Global Cybersecurity and Employment Groups, and focuses his practice on cyber security and data privacy issues, employment law, and complex commercial litigation. Peter also specialises in public international law, and serves as an Adjunct Professor at the University of San Diego School of Law, where he teaches in the areas of international law and appellate advocacy.

Dr **Martin Strohmeier** is a post-doctoral researcher in the Department of Computer Science at the University of Oxford. His main research interests are currently in the area of network security, including wireless sensor networks and critical infrastructure protection. During his DPhil at Oxford, he extensively analysed the security and privacy of wireless aviation technologies of this generation and the next. His work predominantly focuses on developing cyber-physical approaches which can quickly and efficiently improve the security of air traffic control. He has received several best paper awards from both the aviation and computer security community and is a co-founder of the aviation research network OpenSky. Before coming to Oxford in 2012, he received his MSc degree from TU Kaiserslautern, Germany and joined Lancaster University's InfoLab21 and Deutsche Lufthansa AG as a visiting researcher.

Eliza Watt is a Visiting Lecturer and a Doctoral Researcher at the University of Westminster, London. She obtained an LLB, LLM and LLM from University of Westminster and King's College London. She is a non-practising barrister, called to the British Bar at the Honourable Society of the Inner Temple. Having worked as a legal consultant for an environmental remediation company, she returned to academia and will be submitting her PhD thesis in the Autumn of 2017, titled *Cyberspace, Surveillance, Law and Privacy*. She has presented her research at a number of conferences, including University College London, and has published in *The International Journal of Human Rights*.