IEEE
Advancing Technology for Humanity

2014
6th International Conference on Cyber Conflict

PROCEEDINGS

P.Brangetto, M.Maybaum, J.Stinissen (Eds.)

CCDCOE
NATO Cooperative Cyber Defence Centre of Excellence
Tallinn, Estonia

CyCon
International Conference on Cyber Conflict
TALLINN, ESTONIA

2014
6th International Conference on Cyber Conflict

PROCEEDINGS

P.Brangetto, M.Maybaum, J.Stinissen (Eds.)

3-6 JUNE, 2014 TALLINN, ESTONIA

2014 6TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT (CYCON 2014)

IEEE Catalog Number: CFP1426N-PRT
ISBN (print): 978-9949-9544-0-7
ISBN (pdf): 978-9949-9544-1-4

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
Layout: Jaakko Matsalu

CYCON 2014 SPONSORS

TECHNICAL SPONSOR

DIAMOND SPONSOR

GOLD SPONSORS

Co-Sponsored By

SPONSORS

ABOUT THE NATO CCD COE

The NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE) is an international military organisation accredited in 2008 by NATO´s North Atlantic Council as a “Centre of Excellence”. The Centre is not part of NATO’s command or force structure, nor is it funded by NATO. However, it is part of a wider framework supporting NATO Command Arrangements.

The NATO CCD COE’s mission is to enhance capability, cooperation and information-sharing between NATO, NATO member States and NATO’s partner countries in the area of cyber defence by virtue of research, education and consultation. The Centre has taken a NATO-orientated, interdisciplinary approach to its key activities, including academic research on selected topics relevant to the cyber domain from legal, policy, strategic, doctrinal and/ or technical perspectives, providing education and training, organising conferences, workshops and cyber defence exercises and offering consultations upon request.

For more information on the NATO CCD COE, please visit the Centre’s website at http://www.ccdcoe.org.

FOREWORD

This is the second time that the CyCon conference devotes its attention to advanced methods of cyber conflict and their strategic and policy implications. In 2013 we discussed the role and methods of automated cyber defense. We looked at automation not only as an enabling technological device or method that allowed us to increase the effectiveness and sophistication of cyber defensive and offensive actions, but also as a social factor which borders with the political, legal, moral and ethical framework of modern societies. All these factors remain in place as we move to the territory of Active Cyber Defense (ACD), the focus of CyCon 2014.

Historically, under the umbrella of Information Technology (IT)-centric cyber security many defensive cyber security paradigms were proposed, including perimeter bound cyber defense, cyber defense based on protection of mission-critical assets, and network-centric cyber defense. The perimeter bound cyber defense is based on a collection of various outward-facing security measures including firewalls, intrusion detection systems, and anti-virus software. The essence of mission criticality in cyber security is in the idea of protection of some, not all assets, and protecting them, not always, but within some time window. The network-centric cyber security paradigm that was promoted by the U.S. Department of Defense was motivated by the acceleration of the speed and mobility of the modern battlespace, and aimed building a secure information space for connecting people and systems independent of time and location.

Regardless of the research and developments of numerous IT-centric cyber security standards and software/hardware solutions, it has become increasingly evident, from the large number of cyber security incidents collected by government, academic, and industrial cyber security organizations during the last 10 years, that cyber defense, as an institution and industry, is not adequately protecting the national interests of states. The practice of everyday usage of IT-centric cyber defense has revealed that it is technically and financially unconceivable to protect each and every IT component, especially while dealing with very large IT infrastructures, or where IT assets are used in dynamic and unpredictable operational environments. Second, although known for several years that main cyber threats are not coming from casual hackers and petty criminals, but from well organized, well-funded, and well-informed criminal groups and state-sponsored actors, only recently this fact has been institutionalized and elevated to the national security level. The term Advanced Persistent Threat (APT) is often used to describe those criminal and state-sponsored cyber attacks that penetrate specific industrial or governmental organizations and evade detection within those organizations for weeks, months or even years, sometimes even with the help of insiders. The traditional cyber security methods, mostly passive, perimeter-bound and reactive are not conceptually and technologically well-equipped to challenge APT.

The commonly used term for ACD is the one given in the 2011 US Department of Defense, Strategy for Operations in Cyberspace. It states that “Active cyber defense is DoD’s synchronized, real-time capability to discover, detect, analyze, and mitigate threats and vulnerabilities. […] It operates at network speed by using sensors, software, and intelligence to detect and stop malicious activity before it can affect DoD networks and systems.” One can

see two major activities defined in the above-given definition, (1) Cyber Situation Awareness, and (2) Cyber Defensive Actions Planning and Execution. The activities of Cyber Situation Awareness encompass a variety of real-time tasks of cyber infrastructure sensing, real-time data collection, information fusion, analysis, and attack detection. Cyber Defensive Actions include the tasks real-time and pro-active measures like attack neutralization, attacker deception, target masking, cyber forensics, cyber infrastructure and mission adaptation and self-organization in order to assure mission continuation, undertaking offensive measures against the threat agents, prediction of potential future threats and reconfiguration of the cyber infrastructure accordingly, and others.

The current state of ACD brings us to technically, strategically, politically and legally active territory, which is in the process of developing its own widely-accepted models, architectures and solutions, but still debates on some fundamentals on the role, concepts, and scope of ACD. All aspects of ACD mentioned above are reflected in the papers presented at CyCon 2014.

The mission and vision of this conference is to look at the issues, models, and solutions of ACD from a synergistic multi-disciplinary perspective. The conference intention was to underscore the fact that significant progress has been made recently in defenses, industrial and academic communities in developing a common and actionable understanding of the objective, models, as well as boundaries of ACD. In this context, the annual Cyber Conflict (CyCon) conferences conducted in Tallinn by the NATO Cooperative Cyber Defence Centre of Excellence are continuing to provide a unique perspective. This distinctiveness is marked by an innovative synergistic approach to the conceptual framework, architectures, processes and systems of cyber defense. It holistically examines computer science and IT, through the lenses of technology, law, strategy and policy, military doctrine, social and economic concerns and human behavioral modeling with respect to the security of cyber space.

The proceedings of this 6th International Conference on Cyber Conflict 2014 (CyCon 2014) are collected in this volume. The 20 papers were selected by the conference program committee following a rigorous peer review process. The papers are spread across the legal, policy, strategic, and technical spectra of cyber defenses, specifically focusing on the issues of Active Cyber Defence. They include in-depth analyses the concept of ACD as well as its legal and socio-political aspects, models of ACD cyber situational awareness and detection of malicious activities, and cyber operational activities.

This volume is arranged into five chapters. The first chapter, Active Cyber Defence – Concepts, Policy and Legal Implications, discusses the conceptual framework, and legal and (socio-) political aspects of ACD to modern societies. The second chapter, Models of Active Cyber Defense, discusses three important models of ACD, weaponization of code, attacker deception, and deployment of actor agnostic threat models, and the benefits and risk factors of exploitations of those models. The third chapter, Cyber Situation Awareness, collects four papers that examine automatic procedures advancing situation awareness in a tactical operational space, particularly high speed situation awareness algorithms, augmenting sensor situation awareness with intelligence data, collection of situational awareness data from critical IT infrastructure

components, and cyber situational awareness from a military tactical level. The fourth chapter, Deception and Detection, is devoted to attack detection based on signature and anomaly base attack pattern matching, however adapted to the requirements of ACD. The fifth and last chapter, Cyber Operational Activities analyses the concepts and methods used in (military) cyber operations.

We would like to thank the members of both the CyCon 2014 technical program committee and the distinguished peer reviewers for their tireless work in identifying papers for presentation at the conference and publication in this book. Most importantly, though, we are delighted to congratulate this volume's editors – Pascal Brangetto, Markus Maybaum and Jan Stinissen. Without their technical expertise, professional attitude, and personal dedication, this work would not have been possible.

Dr. Gabriel Jakobson
Chief Scientist, Altusys Corp

Dr. Rain Ottis
Associate Professor
Tallinn University of Technology

Brookline, Tallinn, April 2014

TABLE OF CONTENTS

INTRODUCTION

For the sixth year in a row the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE) invited experts from government, academia and industry to Tallinn to discuss recent trends in cyber defence. The *6th International Conference on Cyber Conflict* (CyCon 2014) brought together an international group of computer technology experts, national security thinkers, strategists, political scientists, policy makers and lawyers, all of whom shared a common interest in cyber defence, and served as a hub for knowledge and networking.

CyCon 2014 focused on 'active cyber defence'. Reflecting the interdisciplinary approach of NATO CCD COE, the topic was explored from the technical, conceptual, strategic, political, legal and ethical perspectives on two parallel tracks. The *Strategy and Law Track* was co-chaired by *Jan Stinissen* (NATO CCD COE) and *Dr Rain Ottis* (Tallinn University of Technology), and the *Technology Track* by *Markus Maybaum* (NATO CCD COE) and *Dr Gabriel Jakobson* (Altusys Corp.). Three pre-conference workshops were organised: one on responsive cyber defence, one on cyber norms development, and one on cyber exercise development and cyber ranges, organised in cooperation with the European Defence Agency.

The *Strategy and Law Track* started with a general introduction to active cyber defence, offering presentations on concepts, definitions and policy and strategy considerations, including legal aspects and ethics. The policy issues were addressed both from the perspectives of the State and of private industry. This general introduction was followed by examples of possible active measures and an overview of the present and possible future roles of artificial intelligence in conducting active cyber defence.

On the second day of the conference the legal framework was outlined, addressing the most relevant concepts of international law in the context of active cyber defence: self-defence, countermeasures and the plea of necessity. The second day offered a number of presentations on military cyber operations, discussing topics including situational awareness, key terrain in cyber space, cyber employed as fighting power, targeting and cyber fratricide. Subsequently operations and law were combined in a session that evaluated different operational scenarios from a legal point of view.

The *Strategy and Law Track* offered three panel discussions. The first was on the policy and strategy aspects of active cyber defence, and reflected on the issues addressed by the different speakers, giving the participants the opportunity to ask questions and engage in discussions. The second panel reflected on active cyber defence operations and legal aspects, offering a more in-depth discussion about the applicability of the different legal concepts. A separate panel session dealt with cyber and international relations.

The *Technology Track* focussed on three fields of expertise: cyber intelligence, network technologies and malware, all within the scope of active cyber defence. On the first day of the conference, situational awareness aspects were discussed, including presentations on a more dynamic response to cyber incidents, the fusion of intelligence information and dealing with situational awareness, especially regarding critical infrastructures.

During day two, new aspects within network technologies and their relation to active cyber defence were discussed, focussing on malicious activity detection. New ideas on well-known attack techniques, such as man-in-the-middle attacks, advanced intrusion detection or anomalies in routing, were introduced. A hot wash up of the day's agenda was done in a second panel session, where the challenges of future secure architectures was addressed.

On the third day the technical track focussed on detection of the use of malware in active defence scenarios. Despite having a closer look at the future risks of weaponising code, interesting approaches, such as hiding botnets in TOR networks and new automated botnet detection methodologies, were presented.

The *Joint Sessions*, which brought together both Strategy and Technology Track audiences, covered the field from the highest political level down to presentations on the operational and technical levels, giving insight from the perspective of government, the military, the law and industry.

The editors would like to thank the Co-Chairs and distinguished members of the Programme Committee for their efforts in reviewing, discussing and selecting the submitted papers, guaranteeing their academic quality.

Programme Committee Co-Chairs (in alphabetical order):

- Dr Gabriel Jakobson, Altusys Corp
- Markus Maybaum, NATO CCD COE
- Dr Rain Ottis, Tallinn University of Technology
- Jan Stinissen, NATO CCD COE

Members of the Programme Committee (in alphabetical order):

- Louise Arimatsu, Chatham House
- Dr Iosif I. Androulidakis, University of Ioannina
- Pascal Brangetto, NATO CCD COE
- Emin Caliskan, NATO CCD COE
- Prof. Thomas Chen, College of Engineering, Swansea University
- Steve Chan, Massachusetts Institute of Technology
- Dr Christian Czosseck, CERT BW
- Prof. Dorothy E. Denning, Department of Defense Analysis, Graduate School of Operational and Information Sciences
- Prof. Dr Gabi Dreo Rodosek, Uni Bw Munich
- Colonel Dr Paul Ducheine, Netherlands Defence Academy
- Dr Kenneth Geers, Fireeye
- Prof. Dr Michael R. Grimaila, Associate Professor of Systems Engineering and a member of the Cyberspace Center for Research at the Air Force Institute of Technology
- Dr Jonas Hallberg, Swedish Defence Research Agency
- Prof. David Hutchison, Lancester University
- Kadri Kaska, NATO CCD COE

- Dr Marieke Klaver, TNO
- Prof. Igor Kotenko, St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS)
- Dr Scott Lathrop
- Dr Sean Lawson
- Corrado Leita, Symantec Research Labs
- Samuel Liles, Purdue University
- Lauri Lindström, NATO CCD COE
- Eric Luiijf, TNO Defence, Security and Safety
- Prof. Dr Michael Meier, Uni Bonn, Informatik IV
- Dr Jose Nazario, Invincea Inc.
- Lars Nicander, Center for Asymmetric Threat Studies at the Swedish National Defence College
- D.Sc. Julie J.C.H. Ryan, George Washington University
- Prof. Alexander Smirnov, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS)
- Dr Pontus Svenson, Swedish Defence Research Agency
- Enn Tõugu, IEEE Estonian Section
- Dr Jens Tölle, Fraunhofer FKIE
- Dr Risto Vaarandi, NATO CCD COE
- Liis Vihul, NATO CCD COE
- Colonel Dr Joop Voetelink, Netherlands Defence Academy
- Dr Jozef Vyskoc, VaF Rovinka and Comenius University Bratislava
- Bruce Watson, Stellenbosch University
- Dr Sean Watts, Creighton University
- Prof. Stefano Zanero, Politecnico di Milano, Dip. Elettronica e Informazione

Special thanks are due to the Institute of Electrical and Electronics Engineers (IEEE), the world's largest professional association dedicated to advancing technological innovation and excellence for the benefit of humanity. The IEEE Estonia Section served as technical co-sponsor of CyCon 2014 and of the Conference Proceedings. Numerous IEEE members have supported the Programme Committee, ensuring the academic quality of the papers and supporting their electronic publication and distribution.

Last but not least we would like to thank the authors of the papers collated in this publication for their superb submissions and friendly cooperation during the course of the publication process.

Pascal Brangetto, Markus Maybaum, Jan Stinissen
NATO Cooperative Cyber Defence Centre of Excellence

Tallinn, Estonia
June 2014

Chapter 1
Active Cyber Defence: Concepts, Policy, and Legal Implications

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

The Triptych of Cyber Securityfl A Classi cation of Active Cyber Defence

Robert S. Dewar
Department of Politics
University of Glasgow
Glasgow, United Kingdom
r.dewar.1@research.gla.ac.uk

Abstract: In the field of cyber security, ill-defined concepts and inconsistently applied terminology are further complicating an already complex issue[1]. This causes difficulties for policy-makers, strategists and academics. Using national cyber security strategies to support current literature, this paper undertakes three tasks with the goal of classifying and defining terms to begin the development of a lexicon of cyber security terminology. The first task is to offer for consideration a definition of "active cyber defence" (ACD). This definition is based upon a number of characteristics identified in current academic and policy literature. ACD is defined here as the proactive detection, analysis and mitigation of network security breaches in real-time combined with the use of aggressive countermeasures deployed outside the victim network. Once defined, ACD is contextualised alongside two further approaches to cyber defence and security. These are fortified and resilient cyber defence, predicated upon defensive perimeters and ensuring continuity of services respectively. This contextualisation is postulated in order to provide more clarity to non-active cyber defence measures than is offered by the commonly used term "passive cyber defence". Finally, it is shown that these three approaches to cyber defence and security are neither mutually exclusive nor applied independently of one another. Rather they operate in a complementary triptych of policy approaches to achieving cyber security.

Keywords: *active cyber defence; resilience; cyber security; definition; classification; triptych; lexicon*

1. INTRODUCTION DEFINITION OF THE PROBLEM *IS* THE PROBLEM[2]

A fundamental difficulty facing the development of cyber defence measures, and the wider study of cyber security, is that of accurately defining the issues under scrutiny. Inconsistently applied terminology and concepts are further complicating an already complex issue. Raising

1 Dan Kruger, "Radically Simplifying Cybersecurity," 2012, 1, http://www.absio.com/sites/default/files/assets/Radically_Simplifying_Cybersecurity_V1.4_1.pdf.

2 Ibid.

this may appear pedantic, but the use of ill-defined and inconsistent terms creates difficulties for policy makers in developing strategies to address the risks inherent in an increasingly wired society[3]. In order to begin the process of developing a comprehensive, cohesive lexicon of cyber security terminology a definition of one key feature – active cyber defence – is proposed here. The definition offered is predicated upon proactive measures not only to detect and analyse security breaches in real time and mitigate any damage caused, but also upon aggressive countermeasures undertaken outside the victim network[4].

There are, however, a number of serious concerns with the implementation of active cyber defence (ACD) which will also be examined. There are questions regarding the legality of the use of aggressive countermeasures outside the defender's network, particularly by state actors. Such action can constitute armed attacks under international law which can be responded to with conventional military force. This in turn raises the issues of accurate attribution of incidents given the anonymising capacities of cyberspace, and the militarisation of cyberspace due to the involvement of state military and security apparatus in ACD measures.

To fully classify ACD, it is necessary to contextualise it with other approaches to cyber defence and security. In so doing, a more comprehensive and representative classification of active cyber defence will be made possible. However, this raises issues regarding the erroneous classification of non-ACD actions. Current analyses group together measures such as firewalls, good "cyber hygiene" and network resilience under the umbrella term "passive cyber defence"[5] – a mirror-image of active approaches. This term is not entirely accurate. A more nuanced classification of the actions collated under the term passive cyber defence will be proposed, categorising non-ACD measures as fortified cyber defence and resilient cyber defence.

Finally, it will be argued that the three approaches to cyber defence offered here do not operate in isolation from one another, as is implied by the use of dualistic terms such as "active" and "passive". An examination of the cyber security strategies of national actors will demonstrate that active, fortified and resilient cyber defence are employed in a collaborative triptych of approaches to cyber security: three independent but related concepts coming together to achieve the single goal of operating in cyberspace free from the risk of physical or digital harm.

3 A. Klimburg and H. Tiirmaa-Klaar, *Cybersecurity and Cyberpower Concepts, Conditions and Capabilities for Cooperation for Action within the EU* (European Parliament, April 2011), 11, http://www.europarl.europa.eu/committees/en/sede/studiesdownload.html?languageDocument=EN&file=41648; Sean Lawson, "Beyond Cyber-Doom: Cyberattack Scenarios and the Evidence of History," *Mercatus Center at George Mason University*, 2011, 25, http://www.voafanti.com/gate/big5/mercatus.org/sites/default/files/publication/beyond-cyber-doom-cyber-attack-scenarios-evidence-history_1.pdf.

4 The definition includes measures associated with offensive action in cyberspace, also known as Computer Network Operations (CNO) or Computer Network Attack (CNA). See Sandro Gaycken, *Cyberwar Das Internet als Kriegsschauplatz* (Munich, Germany: Open Source Press, 2011), 142; Heather Harrison Dinniss, *Cyber Warfare and the Laws of War*, 1st ed. (CUP, 2012), 37. Gaycken also discusses deterrence, stating that "a good offense is often the best defence" (Gaycken, *Cyberwar*, 149.). Deterrence is not specifically addressed here as many of the deterring measures employed are active in nature, and based around maintaining a credible second strike in the event of an incident. See Amit Sharma, "Cyber Wars: A Paradigm Shift from Means to Ends," *Strategic Analysis* 34, no. 1 (2010): 69, doi:10.1080/09700160903354450; K. A. Taipale, "Cyber-Deterrence," *LAW, POLICY AND TECHNOLOGY CYBERTERRORISM, INFORMATION, WARFARE, DIGITAL AND INTERNET IMMOBILIZATION*, January 1, 2009, 4, http://papers.ssrn.com/abstract=1336045.

5 James P. Farwell and Rafal Rohozinski, "The New Reality of Cyber War," Survival 54, no. 4 (2012): 109; Leyi Shi et al., "Port and Address Hopping for Active Cyber-Defense," in *Intelligence and Security Informatics* (Springer, 2007), 295.

2. ACTIVE CYBER DEFENCE

Although the term "active defence" is common in the military as the idea of offensive action and counterattacks to deny advantage or position to the enemy[6], the concept remains elusive when applied to the cyber domain[7] and suffers a lack of clarity in related law and national policy[8]. A recent policy brief from the Center for North American Security argued that there is currently no commonly accepted definition of the term "active cyber defence"[9], missing an opportunity to provide one. Nevertheless, attempts have been made to define the concept. Rosenzweig offers a provisional definition as:

> "…the synchronized, real_time capability to discover, detect, analyze, and mitigate threats. [Active cyber defence] operates at network speed using sensors, software and intelligence to detect and stop malicious activity ideally before it can affect networks and systems."[10]

This definition identifies a number of features of ACD, the most important of which is the real-time detection and mitigation of key threats before damage occurs. Specific measures include the deployment of "white worms"[11], benign software similar to viruses but which seek out and destroy malicious software, identify intrusions[12] or engage in recovery procedures[13]. A second active defence tactic is to repeatedly change the target device's identity during data transmission, a process known as address hopping[14]. This has the dual role of masking the target's identifying characteristics as well as confusing the attacker[15]. Address hopping can serve as a useful action to counter espionage by masking the identities of devices where particular data is stored. Active cyber defence therefore places emphasis on proactive measures to counteract the immediate effects of a cyber-incident, either by identifying and neutralising malicious software or by deliberately seeking to mask the online presence of target devices to deter and counter espionage.

There are, however, a number of more aggressive measures which can be taken to defend systems and networks. While white worms can be used to seek out and combat malicious software, Curry and Heckman describe how they can also be used to turn the tools of hackers and would-be intruders against them and identify not just the attacking software, but the servers

6 Shane McGee, Randy V. Sabett, and Anand Shah, "Adequate Attribution: A Framework for Developing a National Policy for Private Sector Use of Active Defense," *Journal of Business & Technology Law* 8, no. 1 (2013): 206.
7 Farwell and Rohozinski, "The New Reality," 110.
8 McGee, Sabett, and Shah, "Adequate Attribution," 2.
9 Irving Lachow, *Active Cyber Defense A Framework for Policymakers*, Policy Brief (Washington, DC: Center for North American Security, February 22, 2013), 3.
10 Paul Rosenzweig, "International Law and Private Actor Active Cyber Defensive Measures," *Stanford Journal of International Law* 47 (2013): 2.
11 Wenlian Lu, Shouhuai Xu, and Xinlei Yi, "Optimizing Active Cyber Defense," in *Decision and Game Theory for Security* (Springer, 2013), 207.
12 Dinniss, *Cyber Warfare*, 108.
13 Lu, Xu, and Yi, "Optimizing Active Cyber Defence," 210.
14 Shi et al., "Address Hopping," 295.
15 Keith A. Repik, *Defeating Adversary Network Intelligence Efforts with Active Cyber Defense Techniques* (DTIC Document, 2008), 22, http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA488411.

and other hardware devices hosting and distributing the attacking code[16]. This is a process known as "hack-back"[17]. Once the source devices of an intrusion or attack have been identified steps can be taken to render those devices inoperative or otherwise prevent them from carrying out their goals. What makes these measures significant is that they are aggressive, offensive techniques which operate beyond the boundaries of the defender's network[18]. They are taking the fight to the attackers.

ACD is therefore a security paradigm employing two methods: one, the real-time identification and mitigation of threats in defenders' networks; two, the capacity to take aggressive, external offensive countermeasures. For the purposes of establishing, or at least beginning the process of developing, a lexicon of cyber security terminology, ACD can therefore be described as:

an approach to achieving cyber security predicated upon the deployment of measures to detect, analyse, identify and mitigate threats to and from communications systems and networks in real-time, combined with the capability and resources to take proactive or offensive action against threats and threat entities including action in those entities' home networks.

Beyond the immediate purpose of establishing a definition of the term "active cyber defence" however, the concept of ACD as a combination of real-time detection and forceful external action raises four important concerns.

First, there are legal implications in the use of offensive external measures. Rosenzweig states that, within the United States (US), private companies are discouraged from using hack-backs as any unauthorised access to a computer or network violates the US Computer Fraud and Abuse Act[19]. This means that defenders who employ software to trace an attacking server and engage in retaliatory action in the attacker's network open themselves up to legal sanction as much as the initial attacker. Given that cyberspace is a series of global networks, this dubious legality is exacerbated when measures undertaken outside the victim network occur extra-territorially, i.e. across international borders[20]. Although such action, when carried out by private corporations, lacks legal cohesion and consensus[21] the concept is particularly problematic when the actors involved include nation-states rather than private companies[22].

The potential for the involvement of nation-states in aggressive cyber techniques is a serious problem because, according to Dinstein[23] and Schmitt[24], that involvement can constitute an armed attack if any action causes damage or disruption of "a scale…comparable to non-cyber

16 John Curry, "Active Defence," *ITNOW* 54, no. 4 (December 1, 2012): 26–27, doi:10.1093/itnow/bws103; Kristin E. Heckman et al., "Active Cyber Defense With Denial and Deception: A Cyber-Wargame Experiment," Computers & Security, 2013, 73, http://www.sciencedirect.com/science/article/pii/S016740481300076X.

17 McGee, Sabett, and Shah, "Adequate Attribution," 2; Rosenzweig, "International Law," 1.

18 Rosenzweig, "International Law," 3.

19 Ibid., 12.

20 Ronald J. Deibert, "The Geopolitics of Internet Control: Censorship, Sovereignty, and Cyberspace," in *Routledge Handbook of Internet Politics*, ed. A. Chadwick and P. N. Howard (London: Routledge, 2009), 334.

21 Rosenzweig, "International Law," 13.

22 It should be noted that in certain circumstances, states are responsible for the actions of private companies, such as state-sponsored private actors or contractors.

23 Yoram Dinstein, "The Principle of Distinction and Cyber War in International Armed Conflicts," *Journal of Conflict and Security Law* 17, no. 2 (July 1, 2012): 261.

24 Michael N. Schmitt, "Classification of Cyber Conflict," *Journal of Conflict and Security Law* 17, no. 2 (July 1, 2012): 250.

operations"[25], has a trans-border element and the attributable involvement of another state and its armed forces[26]. Consequently, a hack-back can be construed as an armed attack if its purpose is to render inoperative the source of the attack and if its effects are comparable to the use of conventional force. This is significant because, under international law, such attacks can be responded to with a range of action including "forcible responses"[27]. This raises the spectre of incidents escalating beyond the cyber-domain into the physical domain. A policy precedent has already been set by the US in this regard. In 2011 policy was issued stating that the US reserved the right to respond to a cyber-attack with military force as the option of last resort[28]. Nation-states have the right to defend themselves against any forms of attack and this right extends beyond kinetic incidents to those perpetrated entirely through cyber operations[29]. However, utilising ACD as a policy or strategic choice must be considered carefully, given its inherent characteristic of action beyond the defender's immediate network[30].

Such risks raise a second problem when employing aggressive, extra-territorial measures: the accurate attribution of the initial incident given the anonymising capacity of cyberspace and its effects on accurately identifying perpetrators. Although the problem of attribution has been extensively examined[31] it is pertinent to raise it here to highlight a major pitfall with the application of ACD as a security strategy, especially given the possibility of kinetic responses to cyber incidents. The basic premise of the attribution problem is that one cannot know with 100% certainty that the identified origin location of a security breach is the true origin of that breach[32]. While attribution is not impossible the anonymising effect of the digital domain makes it very difficult and resource-intensive[33], a feature exploited by malicious online actors as a protection against identification. To respond to an intrusion with a damaging hack-back therefore requires a high degree of certainty. The defender must be confident that the identified source of an intrusion is the genuine source given the legal ramifications examined above. This need for certainty is increased exponentially if nation-states are allegedly involved and reserve the right to deploy conventional weapons as a response to a cyber-incident.

The involvement of state actors and their security and military apparatus leads to a third concern with the use of active cyber defence. Malicious activity in cyberspace runs a gamut from viruses that steal or delete personal data and engage in espionage to acts of sedition and

25 Michael N. Schmitt, ed., *Tallinn Manual on the International Law Applicable to Cyber Warfare* (CUP, 2013), 45.

26 Schmitt, "Classification," 251; Schmitt, *Tallinn Manual*, 54. There is, however, currently an ongoing debate as to whether the actions described as "attacks" are in fact armed attacks or should more accurately be described as sabotage, subversion or espionage. In addition, very few incidents have occurred which qualify as attacks. See Thomas Rid, *Cyber War Will Not Take Place* (London: Hurst, 2013) and Brandon Valeriano and Ryan Maness, "The Dynamics of Cyber Conflict between Rival Antagonists, 2001–11 (in Press)," *Journal of Peace Research*, 2014.

27 Dinniss, *Cyber Warfare*, 108.

28 USA, *International Strategy for Cyberspace Prosperity, Security and Openness in a Networked World*, National Strategy (The White House, May 2011), 14, http://www.whitehouse.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

29 Schmitt, *Tallinn Manual*, 54.

30 Klimburg and Tiirmaa-Klaar, *Cybersecurity*, 13.

31 Dinniss, *Cyber Warfare*, 3,99; Gaycken, *Cyberwar*, 80–86; Schmitt, *Tallinn Manual*, 29–31; Nicholas Tsagourias, "Cyber Attacks, Self-Defence and the Problem of Attribution," *Journal of Conflict and Security Law* 17, no. 2 (2012): 229–44.

32 Dinniss, *Cyber Warfare*, 71.

33 Tsagourias, "Cyber Attacks, Self-Defence and the Problem of Attribution," 233.

the publishing of extremist propaganda[34]. Certain online content is banned in certain states, and so the authorities in those states filter that content. However, Deibert and Rohozinski[35] argue that there is the potential for a "mission creep" to set in when a state deploys the tools necessary to detect malicious activity before it causes any adverse effects. They cite the example of a crackdown on internet pornography by the Thai government leading to the complete blocking of access to YouTube.com[36] as a warning that, once the tools such as filters, address blocking and content analysis are in place, there is a great temptation to employ these tools for an ever expanding range of purposes. The allegations of mass surveillance of digital communications by Western security services published in the UK's Guardian newspaper[37] in 2013 demonstrate the risks of such a mission creep. What began as measures to combat terrorism have allegedly become programmes of mass data collection. The point here is that the use of ACD measures must be carried out with great care to avoid expanding a filtering remit beyond legitimate security concerns – such as preventing the spread of extremist propaganda – to overzealous measures such as unauthorised access to private correspondence.

The problem with such active filtering and surveillance is that, given the opportunities for the deployment of state apparatus[38], these actions are often carried out by national security or military institutions, leading to a potential militarisation of cyberspace[39]. The cyber security strategies of the actors adopting an ACD approach demonstrate the level to which military institutions are already being deployed as part of the security solution. In two specific cases – namely United Kingdom (UK) and the US – military institutions play a strong role in providing and ensuring cyber security though active cyber defence measures.

The UK Cyber Security Strategy identifies the proactive measures taken to disrupt threats to and from networked communications systems[40]. The Ministry of Defence (MoD) is tasked with improving the UK's ability to detect threats in cyberspace and to "anticipate, prepare for and disrupt" such threats[41]. To do this, resources have been provided to the MoD itself and the Government Communications Headquarters (GCHQ) to develop a range of techniques – including proactive measures – to disrupt those threats. This strategic approach falls neatly into Rosenzweig's definition of ACD – efforts to detect and hinder malicious activity – but implies the extension of action beyond the confines of national or UK government networks through proactive measures described by Curry and Heckman, as well as Lu et al[42]. The fact that the MoD has been assigned these tasks, despite UK cyber security strategy being led by the Cabinet

34 Maura Conway, "Cybercortical Warfare: Hizbollah's Internet Strategy," in *The Internet and Politics; Citizens, Voters and Activists*, ed. S. Oates, D. Owen, and R. Gibson (Routledge, 2005); Jialun Qin et al., "Analyzing Terror Campaigns on the Internet: Technical Sophistication, Content Richness, and Web Interactivity," *International Journal of Human-Computer Studies* 65, no. 1 (January 2007): 71–84.

35 Deibert, "The Geopolitics of Internet Control: Censorship, Sovereignty, and Cyberspace," 327.

36 Ibid.

37 The Guardian, "The NSA Files," Report Series, *The NSA Files | World News | The Guardian*, June 8, 2013, http://www.guardian.co.uk/world/the-nsa-files.

38 Curry, "Active Defence."

39 Ronald J. Deibert, "Militarizing Cyberspace," *Technology Review* 12 (August 2010), http://www.technologyreview.com/notebook/419458/militarizing-cyberspace/; Myriam Dunn Cavelty, "The Militarisation of Cyberspace: Why Less May Be Better," in *4th International Conference on Cyber Conflict*, ed. C. Czosseck, R. Ottis, and K. Ziolkowski (NATO CCD COE Publications, 2012), 141–53.

40 UK, *The UK Cyber Security Strategy Protecting and Promoting the UK in a Digital World*, National Strategy (UK Cabinet Office, 2011), 27.

41 Ibid., 39.

42 Curry, "Active Defence"; Heckman et al., "Active Cyber Defense With Denial and Deception"; Lu, Xu, and Yi, "Optimizing Active Cyber Defence."

Office – a civilian organ of central government – demonstrates a willingness to deploy military resources to provide cyber defence and security.

Such willingness is also present in the US's approach to cyber security. There are two documents which together expound American policy in this field: the White House's International Strategy for Cyberspace[43] and the Department of Defense (DoD)'s Strategy for Operating in Cyberspace[44]. The second document specifically cites the use of active cyber defence capabilities to prevent intrusions[45], clearly placing it within an active framework. Furthermore, as examined above, the prominence of military institutions in US cyber security policy and strategy is demonstrated by the explicit willingness of the American government to use military force (when all other avenues have been exhausted) in response to hostile acts in cyberspace[46]. If a key principle of ACD is the extension of measures beyond the immediate confines of victim systems and networks, then the use of kinetic military force in response to a cyber-attack is the ultimate example of such an extension and the example most prone to the issues of legality, attribution, mission creep and militarisation. Clearly therefore, the adoption of such active defence policies is concerning as it means military resources are being deployed to ensure security[47], necessarily increasing the level to which national military and security services are involved in cyber security policy decisions. Cyberspace has already been classified as a fifth military domain by the US and Japan[48] leading these states to seek military capacities and capabilities in that domain. The mission creep Deibert and Rohozinski warned against is manifesting itself in an increased military presence in cyberspace particularly if it takes on the task not only of restricting access to particular data, but also engages in measures outside the home networks of defended states.

The concept of combatting threats outside the network or systems under attack therefore raises a number of significant concerns, not least the capacity for defending actors to respond with kinetic military force and the ramifications of doing so. However, the extra-territoriality inherent to ACD is vital to our understanding of the concept as a methodological approach to cyber security due to the fact that it is this aggressive external action which differentiates ACD from other approaches. These other approaches have to date been described as "passive cyber defence"[49]. Such a description raises a fourth issue around ACD and current efforts to define the concept: the assumption that all other, non-active forms of cyber defence are "passive" or reactive in nature.

Farwell and Rohozinski describe passive cyber defence as an approach which includes:

> "firewalls, cyber 'hygiene' that trains an educated workforce to guard against errors or transgressions that can lead to cyber intrusion, detection technology, 'honey pots' or

43 USA, *International Strategy*.

44 USA, *Department of Defense Strategy for Operating in Cyberspace*, National Strategy (Department of Defense, 2011), http://www.defense.gov/home/features/2011/0411_cyberstrategy/docs/DoD_Strategy_for_Operating_in_Cyberspace_July_2011.pdf.

45 Ibid., 6.

46 USA, *International Strategy*, 14.

47 Dunn Cavelty, "Militarisation of Cyberspace," 141; Ronald J. Deibert, "Black Code: Censorship, Surveillance, and the Militarisation of Cyberspace," *Millennium-Journal of International Studies* 32, no. 3 (2003): 501–30.

48 Japan, "Cyber Security Strategy of Japan," June 2013, 41, http://www nisc.go.jp/eng/pdf/CyberSecurityStrategy.pdf; USA, *Strategy for Operating in Cyberspace*, 5.

49 Farwell and Rohozinski, "The New Reality," 109; Shi et al., "Address Hopping," 295.

> decoys that serve as diversions, and managing cyberspace risk through collective defence, smart partnerships, information training, greater situation awareness, and establishing secure, resilient network environments"[50]

Such actions, as well as installing intrusion detection and prevention measures[51] are not considered active defences. Rather they create a preventive environment[52] predicated on information-sharing and resilience. Lachow goes further, arguing that passive cyber defences which rely on perimeter sensors cannot adequately protect against sophisticated cyber-attacks[53] as these can adapt quickly and become more advanced than the defences of their targets. The term "passive" therefore implies a purely reactive approach: dealing with an incident once it has occurred rather than actively trying to prevent that occurrence in the first place. However, just as ACD is not as simple as taking proactive action of any kind, the construction of decoys, collective defence paradigms, information-sharing and the development of resilient networks which can cope with accidental or intentional damage are not simple reactions, and certainly not passive policies. They involve taking action to prevent and minimise the damage of a cyber-incident without resorting to the aggressive measures inherent to ACD.

In the interests of developing a consistent, coherent lexicon of terminology, active cyber defence is not the only term that suffers from a lack of definition. The same is true of that group of measures taken to mitigate the damage of cyber-incidents or return systems and networks to full functionality in the event of an incident. Instead of labelling these measures "passive cyber defence" – a simple mirror-image of "active cyber defence" – a clearer and more accurate categorisation of these measures would be to label them "fortified cyber defence" and "resilient cyber defence".

3. FORTIFIED CYBER DEFENCE

As discussed above, measures such as the establishment of firewalls, anti-virus software and detection technologies have been labelled by some commentators as passive, reactive forms of defence. However, if the ultimate aim of these actions is examined, the collection of measures involved cannot be accurately labelled as passive. The goal of firewalls and filters, and any other measures intended to prevent malicious access to key assets is just that – the prevention of access[54]. Steps are taken to reduce the chances of any intrusion or attack succeeding in its aims. An analogy to this is the construction of physical fortifications such as castles and fortresses. These were built with the intention of protecting those inside from outside attackers. Methods such as installing firewalls or placing filters and scanners on trunk cables are all intended to prevent malicious code, information or actors accessing network systems and exploiting assets[55]. These are not "passive" measures, taken in reaction to an incident; rather they are actions designed to build virtual fortifications.

In addition to the installation of firewalls and anti-virus software, fortified cyber defence (FCD)

50 Farwell and Rohozinski, "The New Reality," 109.
51 Shi et al., "Address Hopping," 295.
52 Lu, Xu, and Yi, "Optimizing Active Cyber Defence," 209.
53 Lachow, *Active Cyber Defense*, 1.
54 Ronald J. Deibert and Rafal Rohozinski, "Risking Security: Policies and Paradoxes of Cyberspace Security," *International Political Sociology* 4, no. 1 (2010): 25, doi:10.1111/j.1749-5687.2009.00088.x.
55 Deibert, "The Geopolitics of Internet Control: Censorship, Sovereignty, and Cyberspace," 325.

can be achieved by building security into the infrastructure supporting cyberspace: the software, computers, routers and other elements needed to enable the online domain to function[56]. The unpredictable and fragile nature of vast international computer networks creates a systemic ontological insecurity in cyberspace[57], making its infrastructure vulnerable to natural, accidental or malicious incidents. Data packets can be corrupted while in transit due to faulty cables, individual computers can themselves malfunction over time and software can fail. Building security measures into all the elements required for the international communications networks to function would mitigate against such systemic and exploitable vulnerabilities. In addition to providing a definition of active cyber defence, a definition of FCD is also offered here:

constructing systemically secure communications and information networks in order to establish defensive perimeters around key assets and minimise intentional or unintentional incidents or damage.

While the defining characteristic of ACD is aggressive action taken outside the defender's home network, the defining characteristic of FCD is that approach's preventive, introspective focus. FCD measures seek to establish defensive perimeters through systems of firewalls and antivirus software in order to minimise the chances of access to target systems and networks.

As discussed above, the US and UK cyber security strategies provide examples of national policies adopting ACD. Germany, on the other hand, provides an example of a national policy promoting FCD[58]. The focus for the German Cyber Security Strategy is ensuring that malicious intrusions are unsuccessful within a preventive security framework[59]. This is achieved through certain key objectives, including training and international co-operation as well as tackling cyber-crime. The ultimate aim of the German Strategy is to ensure that critical infrastructures and public and private IT systems are secure from threats which affect the confidentiality, integrity and availability of electronic data, and the availability of information and communications technology (ICT)[60]. The German approach to cyber security is therefore not a passive, reactive approach despite employing techniques Farwell and Rohozinski associate with passive cyber defence[61]. It is proactive in that it takes the issues seriously and aims to put in place particular measures to create a preventive environment where the possibility of breach success is minimised while not employing aggressive extra-territorial countermeasures designed for operation in an attacker's home network.

4. RESILIENT CYBER DEFENCE

A third approach to cyber defence is based not upon aggressively seeking perpetrators of security breaches or establishing fortifications around key assets. Instead it focusses on ensuring critical

56 Gary McGraw, "Cyber War Is Inevitable (Unless We Build Security In)," *Journal of Strategic Studies* 36, no. 1 (February 2013): 113.

57 Lene Hansen and Helen Nissenbaum, "Digital Disaster, Cyber Security, and the Copenhagen School," *International Studies Quarterly* 53, no. 4 (2009): 1160.

58 Germany, *Cyber Security Strategy for Germany (official Translation)*, National Strategy (Bonn: Federal Office for Information Security, 2011), http://www.bmi.bund.de/SharedDocs/Downloads/DE/Themen/OED_Verwaltung/Informationsgesellschaft/cyber.html?nn=109632.

59 Ibid., 5.

60 Ibid., 4.

61 Farwell and Rohozinski, "The New Reality," 109.

infrastructures and services which rely on networked communications continue to function and to provide the services for which they were designed. Rather than aggressive or fortified cyber defence, a potentially more pragmatic approach to cyber security in general is "resilient cyber defence" (RCD).

Resilience itself is predicated upon accepting that incidents will occur and focussing on the ability to recover from those incidents[62], either returning to the original state or adapting to generate a new, adjusted state[63]. In terms of precise technical measures, resilience in the cyber domain shares a number of traits with FCD: it requires practitioners and policy makers to focus their security efforts internally, making sure systems and networks are adaptable or can withstand incidents. Building security measures into those systems[64] is a key feature in such preparedness. RCD can therefore be defined as:

ensuring the continuity of system functionality and service provision by constructing communications and information networks with the systemic, inbuilt ability to withstand or adapt to intentional or unintentional incidents.

While ACD and FCD seek to identify threats and intrusions as soon as possible and deal with them, RCD advocates sharing vital information regarding security breaches among all interested parties and potential future victims[65].

Resilience is a common trait in current cyber security policy documents. The strategies of the European Union (EU) and Japan favour this approach. They concentrate on sharing information between public and private bodies, harmonising public infrastructure security measures and developing uniform standards of security[66] to ensure preparedness in the event of a natural or malicious incident. Other features of resilient cyber defence include ensuring that the private sector is actively involved in solution development, and promoting the recognition of shared responsibility amongst government agencies, private companies and individual users. That way, as many actors as possible know of a particular virus or intrusion mechanism and can take steps to ensure that system functionality continues should they be targeted.

The defining characteristic of RCD is this idea of functional continuity. Active paradigms concentrate on identifying threats and their origins and taking remedial and punitive external action. Fortified models focus on ensuring that network defences are in place to prevent, or at least minimise the success of, a security breach. Resilient models prioritise the continued functioning and service provision of the systems that rely on network communications so that

62 Christopher W. Zobel and Lara Khansa, "Quantifying Cyberinfrastructure Resilience against Multi-Event Attacks," *Decision Sciences* 43, no. 4 (2012): 688.

63 Myriam Dunn Cavelty, "Cyber-Security," in *Contemporary Security Studies*, ed. Alan Collins, 3rd ed. (OUP, 2012), 19.

64 Hansen and Nissenbaum, "Digital Disaster"; McGraw, "Cyber War."

65 European Commission, *JOIN (2013) 1 Final JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Cybersecurity Strategy of the European Union An Open, Safe and Secure Cyberspace*, Communication (European Commission, February 7, 2013), 6, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=JOIN:2013:0001:FIN:EN:pdf.

66 Japan, "Cyber Security Strategy," 30; European Commission, *Cybersecurity Strategy*, 5.

there is no break in that service[67]. To provide a simple example: if a power station suffers a cyber security breach, the first priority for an RCD approach would be to ensure that electricity production continues unaffected.

On examination therefore, fortified and resilience-based cyber defence solutions cannot be described as "passive cyber defence"[68]. Rather, they advocate a state of readiness, a capability to withstand malicious or natural incidents. Processes and procedures must be put in place to involve all interested actors in information-sharing, whether these are government agencies, public bodies or private sector companies. The EU is currently considering legislation which would make it a legal requirement for all relevant public and private actors to share security breach information[69]. Network fortification and resilience recommends that security and adaptability be built into the infrastructure supporting the online environment[70]. Given that cyber-incidents are varied and increasing[71], a state of readiness is a far more pragmatic option than aggressive techniques fraught with issues around accurate attribution, questionable legal standpoints and overzealous deployment of security and military resources and the consequences those actions risk.

The result of this classification is the identification of not two modes of cyber defence (active or passive), but three – active, fortified and resilient cyber defence. However the three paradigms are not mutually exclusive. While very different given their varying techniques, each approach operates in conjunction with the other to achieve a wider single goal, cyber security. By concentrating not on the implementation of the measures themselves but their ultimate goals these three paradigms together form a "Triptych of Cyber Security": three parallel approaches to achieving security when interacting with and utilising cyberspace.

5. CONCLUSION – THE 'TRIPTYCH' OF CYBER SECURITY

Active cyber defence (ACD) is an approach to cyber security predicated upon proactive measures to identify malicious codes and other threats, as well as aggressive external techniques designed to neutralise threat agents. ACD is defined by the capacity and willingness to take action outside the victim network[72]. Despite this, ACD is not mirrored by "passive cyber defence". The measures collated under this term should more accurately be classified as fortified and resilient cyber defence. These terms clarify the nature of the action taken by focussing on the end goals of the measures they describe.

The three types of cyber defence described here are not mutually exclusive. Instead they operate

67 European Commission, *Cybersecurity Strategy*, 6; Switzerland, *National Strategy for Switzerland's Protection against Cyber Risks*, National Strategy, 2012, 38, http://www melani.admin.ch/dokumentation/00123/01525/index.html?lang=en.

68 Farwell and Rohozinski, "The New Reality," 109; Lachow, *Active Cyber Defense*, 1; Shi et al., "Address Hopping," 295.

69 European Commission, "COM (2013) 48 Final Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL Concerning Measures to Ensure a High Common Level of Network and Information Security across the Union" (EUR-Lex, February 7, 2013), http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2013:0048:FIN:EN:PDF.

70 McGraw, "Cyber War."

71 European Commission, *Cybersecurity Strategy*, 3.

72 Lu, Xu, and Yi, "Optimizing Active Cyber Defence"; Rosenzweig, "International Law," 3.

in conjunction with one another in a triptych of measures further highlighting the inaccuracy of a simple divide between active and passive approaches. The goal of cyber security is to enable operations in cyberspace free from the risk of physical or digital harm. To that end, the three paradigms of defence postulated here work together to complement each other through a range of measures designed to address specific issues around online security. Active cyber defence focusses on identifying and neutralising threats and threat agents both inside and outside the defender's network, while fortified defence builds a protective environment. In its turn resilience focusses on ensuring system continuity. The national strategies developed over the last ten years demonstrate the complementarity of these three approaches. The US and UK categorically adopt an active paradigm, whereby all available resources are deployed to protect national interests, including proactively seeking out enemy actors and rendering them ineffective. The US further retains the right to deploy the ultimate sanction of kinetic military force in the event of a cyber-attack as a measure of last resort. However, neither the UK nor the US are ignorant of the benefits of fortifying assets, or of making critical national infrastructures resilient to the failures of the communications systems on which they rely[73]. For Germany the policy of choice is FCD but network resilience is recognised in a commitment to protecting and securing critical digital infrastructures due to their importance to physical social and economic services[74]. The EU and Japan adopt a resilience-based framework, yet both are seeking to develop active defence capabilities[75].

What this demonstrates is a conscious acknowledgement that one single approach to cyber security is not enough. Active cyber defence, including all the measures that that concept entails, is insufficient when seeking to achieve cyber security. Steps must be taken to fortify assets in order to minimise the likelihood and effectiveness of cyber-incidents, as well as ensure system and infrastructure continuity should an incident occur. Equally, FCD and RCD do not serve as effective deterrents to would-be attackers. The willingness to identify and pursue threat agents into their own home networks must be demonstrated alongside asset fortification and system resilience. In short, the paradigms of cyber defence are not stand-alone approaches. Even for those actors which place their strategies within an active framework, military or security agency resources are not the only ones utilised. The consequence of this is the deployment of elements of each approach simultaneously in a triptych of approaches intended to achieve a single goal.

By contextualising ACD as an approach which is used collaboratively with its fortified and resilient cousins in a triptych of cyber security, and highlighting the crucial difference of aggressive action beyond the victim network, it is possible to distil a definition of the term "active cyber defence". This is in spite of ACD being fraught with unresolved legal and diplomatic difficulties. For the purposes of classification, a definition of active cyber defence is proposed here:

a method of achieving cyber security predicated upon the deployment of measures to detect, analyse, identify and mitigate threats to and from cyberspace in real-time, combined with the capability and resources to take proactive or aggressive action against threat agents in those agents' home networks.

73 UK, *Cyber Security Strategy*, 39; USA, *Strategy for Operating in Cyberspace*, 6; USA, *International Strategy*, 18.
74 Germany, *Cyber Security Strategy*, 6.
75 European Commission, *Cybersecurity Strategy*, 11; Japan, "Cyber Security Strategy," 41.

The question of definition and classification in the cyber security debate will not be resolved overnight. While active cyber defence is one feature of that debate, the definition and classification offered here will go some way towards establishing a cohesive lexicon of terminology, an exercise which will assist the development of legal and political solutions to the complex issue of cyber security.

REFERENCES:

Conway, Maura. "Cybercortical Warfare: Hizbollah's Internet Strategy." In *The Internet and Politics; Citizens, Voters and Activists*, edited by S. Oates, D. Owen, and R. Gibson. Routledge, 2005.

Curry, John. "Active Defence." *ITNOW* 54, no. 4 (December 1, 2012): 26–27. doi:10.1093/itnow/bws103.

Deibert, Ronald J. "Black Code: Censorship, Surveillance, and the Militarisation of Cyberspace." *Millennium-Journal of International Studies* 32, no. 3 (2003): 501–30.
"Militarizing Cyberspace." *Technology Review* 12 (August 2010). http://www.technologyreview.com/notebook/419458/militarizing-cyberspace/.
"The Geopolitics of Internet Control: Censorship, Sovereignty, and Cyberspace." In *Routledge Handbook of Internet Politics*, edited by A. Chadwick and P. N. Howard, 323–36. London: Routledge, 2009.

Deibert, Ronald J , and Rafal Rohozinski. "Risking Security: Policies and Paradoxes of Cyberspace Security." *International Political Sociology* 4, no. 1 (2010): 15–32. doi:10.1111/j.1749-5687.2009.00088 x.

Dinniss, Heather Harrison. *Cyber Warfare and the Laws of War*. 1st ed. CUP, 2012.

Dinstein, Yoram. "The Principle of Distinction and Cyber War in International Armed Conflicts." *Journal of Conflict and Security Law* 17, no. 2 (July 1, 2012): 261–77. doi:10.1093/jcsl/krs015.

Dunn Cavelty, Myriam. "Cyber-Security." In *Contemporary Security Studies*, edited by Alan Collins. 3rd ed. OUP, 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2055122.
"The Militarisation of Cyberspace: Why Less May Be Better." In *4th International Conference on Cyber Conflict*, edited by C. Czosseck, R. Ottis, and K. Ziolkowski, 141–53. NATO CCD COE Publications, 2012.

European Commission. "COM (2013) 48 Final Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL Concerning Measures to Ensure a High Common Level of Network and Information Security across the Union." EUR-Lex, February 7, 2013. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2013:0048:FIN:EN:PDF.
JOIN (2013) 1 Final JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Cybersecurity Strategy of the European Union An Open, Safe and Secure Cyberspace. Communication. European Commission, February 7, 2013. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=JOIN:2013:0001:FIN:EN:pdf.

Farwell, James P., and Rafal Rohozinski. "The New Reality of Cyber War." *Survival* 54, no. 4 (2012): 107–20.

Gaycken, Sandro. *Cyberwar Das Internet als Kriegsschauplatz*. Munich, Germany: Open Source Press, 2011.

Germany. *Cyber Security Strategy for Germany (official Translation)*. National Strategy. Bonn: Federal Office for Information Security, 2011. http://www.bmi.bund.de/SharedDocs/Downloads/DE/Themen/OED_Verwaltung/Informationsgesellschaft/cyber.html?nn=109632.

Hansen, Lene, and Helen Nissenbaum. "Digital Disaster, Cyber Security, and the Copenhagen School." *International Studies Quarterly* 53, no. 4 (2009): 1155–75. doi:10.1111/j 1468-2478.2009.00572.x.

Heckman, Kristin E., Michael J. Walsh, Frank J. Stech, Todd A. O'Boyle, Stephen R. DiCato, and Audra F. Herber. "Active Cyber Defense With Denial and Deception: A Cyber-Wargame Experiment." *Computers & Security*, 2013. http://www sciencedirect.com/science/article/pii/S016740481300076X.

Japan. "Cyber Security Strategy of Japan," June 2013. http://www.nisc.go.jp/eng/pdf/CyberSecurityStrategy.pdf.

Klimburg, A., and H. Tiirmaa-Klaar. *Cybersecurity and Cyberpower Concepts, Conditions and Capabilities for Cooperation for Action within the EU*. European Parliament, April 2011. http://www.europarl.europa.eu/committees/en/sede/studiesdownload html?languageDocument=EN&file=41648.

Kruger, Dan. "Radically Simplifying Cybersecurity," 2012. http://www.absio.com/sites/default/files/assets/Radically_Simplifying_Cybersecurity_V1.4_1.pdf.

Lachow, Irving. *Active Cyber Defense A Framework for Policymakers*. Policy Brief. Washington, DC: Center for North American Security, February 22, 2013. http://www.cnas.org/publications/policy-briefs/active-cyber-defense-a-framework-for-policymakers.

Lawson, Sean. "Beyond Cyber-Doom: Cyberattack Scenarios and the Evidence of History." *Mercatus Center at George Mason University*, 2011. http://www.voafanti com/gate/big5/mercatus.org/sites/default/files/publication/beyond-cyber-doom-cyber-attack-scenarios-evidence-history_1.pdf.

Lu, Wenlian, Shouhuai Xu, and Xinlei Yi. "Optimizing Active Cyber Defense." In *Decision and Game Theory for Security*, 206–25. Springer, 2013. http://link.springer.com/chapter/10.1007/978-3-319-02786-9_13.

McGee, Shane, Randy V. Sabett, and Anand Shah. "Adequate Attribution: A Framework for Developing a National Policy for Private Sector Use of Active Defense." *Journal of Business & Technology Law* 8, no. 1 (2013): 1.

McGraw, Gary. "Cyber War Is Inevitable (Unless We Build Security In)." *Journal of Strategic Studies* 36, no. 1 (February 2013): 109–19. doi:10.1080/01402390.2012.742013.

Qin, Jialun, Yilu Zhou, Edna Reid, Guanpi Lai, and Hsinchun Chen. "Analyzing Terror Campaigns on the Internet: Technical Sophistication, Content Richness, and Web Interactivity." *International Journal of Human-Computer Studies* 65, no. 1 (January 2007): 71–84. doi:10.1016/j.ijhcs.2006.08.012.

Repik, Keith A. *Defeating Adversary Network Intelligence Efforts with Active Cyber Defense Techniques*. DTIC Document, 2008. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA488411.

Rid, Thomas. *Cyber War Will Not Take Place*. London: Hurst, 2013.

Rosenzweig, Paul. "International Law and Private Actor Active Cyber Defensive Measures." *Stanford Journal of International Law* 47 (2013). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2270673.

Schmitt, Michael N. "Classification of Cyber Conflict." *Journal of Conflict and Security Law* 17, no. 2 (July 1, 2012): 245–60. doi:10.1093/jcsl/krs018.
, ed. *Tallinn Manual on the International Law Applicable to Cyber Warfare*. CUP, 2013.

Sharma, Amit. "Cyber Wars: A Paradigm Shift from Means to Ends." *Strategic Analysis* 34, no. 1 (2010): 62–73. doi:10.1080/09700160903354450.

Shi, Leyi, Chunfu Jia, Shuwang Lü, and Zhenhua Liu. "Port and Address Hopping for Active Cyber-Defense." *In Intelligence and Security Informatics*, 295–300. Springer, 2007. http://link.springer.com/chapter/10 1007/978-3-540-71549-8_31.

Switzerland. *National Strategy for Switzerland's Protection against Cyber Risks*. National Strategy, 2012. http://www melani.admin ch/dokumentation/00123/01525/index.html?lang=en.

Taipale, K. A. “Cyber-Deterrence.” *LAW, POLICY AND TECHNOLOGY CYBERTERRORISM, INFORMATION, WARFARE, DIGITAL AND INTERNET IMMOBILIZATION*, January 1, 2009. http://papers.ssrn.com/abstract=1336045.

The Guardian. “The NSA Files.” Report Series. *The NSA Files | World News | The Guardian*, June 8, 2013. http://www.guardian.co.uk/world/the-nsa-files.

Tsagourias, Nicholas. “Cyber Attacks, Self-Defence and the Problem of Attribution.” *Journal of Conflict and Security Law* 17, no. 2 (2012): 229–44.

UK. *The UK Cyber Security Strategy Protecting and Promoting the UK in a Digital World*. National Strategy. UK Cabinet Office, 2011. http://www.cabinetoffice.gov.uk/resource-library/cyber-security-strategy.

USA. *Department of Defense Strategy for Operating in Cyberspace*. National Strategy. Department of Defense, 2011. http://www.defense.gov/home/features/2011/0411_cyberstrategy/docs/DoD_Strategy_for_Operating_in_Cyberspace_July_2011.pdf.
International Strategy for Cyberspace Prosperity, Security and Openness in a Networked World. National Strategy. The White House, May 2011. http://www.whitehouse.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

Valeriano, Brandon, and Ryan Maness. “The Dynamics of Cyber Conflict between Rival Antagonists, 2001–11 (in Press).” *Journal of Peace Research*, 2014.

Zobel, Christopher W., and Lara Khansa. “Quantifying Cyberinfrastructure Resilience against Multi-Event Attacks.” *Decision Sciences* 43, no. 4 (2012): 687–710.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Socio-Political E□ ects of Active Cyber Defence Measures

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

Kim Hartmann
Otto-von-Guericke-Universität
Magdeburg, Germany
kim hartmann@ovgu.de

Abstract: This paper compares public and political attitudes across a range of countries to systems for monitoring and surveillance of internet usage. U.S. and Russian data collection and mining systems are taken as case studies. There are wide variations in societal acceptability of these systems based on the perceived acceptable balance between personal privacy and national security. Disclosures of covert internet monitoring by U.S. and other government agencies since mid-2013 have not led to a widespread public rejection of this capability in the U.S. or Europe, while in Russia, internet users show acceptance of limitations on privacy as normal and necessary. An incipient trend in EU states toward legitimisation of real-time internet monitoring is described.

Keywords: *active cyber defence, Russia, UK, monitoring, surveillance, US, PRISM, SORM*

1. INTRODUCTION

Like many other concepts relating to cyberspace, the term "Active Cyber Defence" at present lacks a universally accepted definition. But any such definition must encompass proactive measures in cyberspace for the purpose of incident prevention, and these measures must not necessarily be limited to technical means.[1] In this paper, we examine social and political, rather than technical, aspects of a national proactive cyber defence posture, by examining two sets of preventive measures related to monitoring and surveillance of an online population.

In China, as well as to some extent in Russia, misuse of social media is perceived as a significant national security issue. The perceived threat is from "the rapid growth of social networking and instant communication tools, like Weike and WeChat, which disseminate information rapidly,

1 According to one authoritative US official, cyber defence is the "ability to draw on the strengths of our partners and bring to bear the best technical skills against any existing or evolving threat. Effective cyber defenses ideally prevent an incident from taking place. Any other approach is simply reactive". See testimony to U.S. Committee on Homeland Security and Governmental Affairs by Sallie McDonald, Assistant Commissioner for the U.S. Office of Information Assurance and Critical Infrastructure Protection, published 31 July 2012, available at http://hsgac-amend.senate.gov/old_site/100401mcdonald.htm

have a large influence and broad coverage, and have a strong ability to mobilize society."[2] Close control of social media, and warning and punishing abusers in order to prevent uncontrolled distribution of information which is hostile to the ruling powers is a prime example of proactive online defence to protect national security.[3]

In this paper, one U.S. and one Russian online data collection and mining system intended to exploit the internet to defend against threats to national security will be reviewed. These two programmes, known to the public as PRISM and SORM respectively, are instructive not only because they demonstrate two different approaches to a similar problem set, but also because they were initiated and continue to be operated in two very different legal and social contexts. Thus conclusions can be drawn for the legal status, and social acceptability, of other possible active cyber defence measures relating to surveillance of online activity.

The paper will review considerations regarding the broad effects of PRISM and SORM on national and international security and privacy issues, as well as whether and where these programmes are operated entirely in accordance with national law. The range of public and official reaction to both these systems in various countries will also be considered, allowing conclusions to be drawn about the extent to which proactive measures would be palatable to public opinion in the future.

2. THE INTERNATIONAL DEBATE

Disclosures of alleged U.S. surveillance activities to the public by former National Security Agency (NSA) contractor Edward Snowden in June 2013 sparked heated international debate on telecommunications monitoring as an act of prevention (i.e. as a form of proactive defence). Public discussion in the U.S., Europe, Russia and beyond revealed widely varying societal attitudes to the issues involved.

Although during the early stages of disclosure public dismay and strident political disapproval was primarily directed at the NSA and its British counterpart, GCHQ, as the Snowden disclosures progressed it became increasingly evident that many other states had been engaging in their own analogous monitoring and surveillance programmes, constrained only by the limitations of geography, political ambition and budget.[4] In the words of one authoritative commentator, this reflected the "big difference between the public outrage of politicians and the day-to-day reality of intelligence co-operation between Americans and Europeans".[5]

According to Finnish Foreign Minister Erkki Tuomioja, "All states spy on each other… All states are also being spied upon."[6] And Russian Foreign Minister Sergey Lavrov is reported to have

2 Paul Mozur, "China Wants to Control Internet Even More", *Wall Street Journal*, November 15, 2013, http://blogs.wsj.com/chinarealtime/2013/11/15/china-wants-greater-internet-control-public-opinion-guidance/

3 Josh Chin And Paul Mozur, China Intensifies Social-Media Crackdown, *Wall Street Journal*, September 19, 2013, http://online.wsj.com/news/articles/SB10001424127887324807704579082940411106988

4 Nigel Inkster, "Snowden – myths and misapprehensions", IISS, 15 November 2013, http://www iiss.org/en/politics%20and%20strategy/blogsections/2013-98d0/november-47b6/snowden-9dd1

5 Julian Lindley-French, "What U.S. Intelligence Really Says About Europe", Speaking Truth Unto Power, October 31, 2013, http://lindleyfrench.blogspot co.uk/2013/10/what-us-intelligence-really-says-about.html

6 "Foreign Minister: All states involved in spying", Yle news, November 3, 2013, http://yle.fi/uutiset/foreign_minister_all_states_involved_in_spying/6914489

commented on the monitoring of world leaders' phones: "It's a little boring to even comment. I mean, really, everybody already knew."[7] But elsewhere, especially in Western Europe, calm and reasoned reaction from responsible politicians was strikingly rare. Well-informed British expert Nigel Inkster notes that "countries that considered themselves to have friendly relations with the United States but which [had] been the subject of U.S. covert intelligence collection... reacted with varying degrees of outrage – some of it real, but much of it manufactured either for domestic political reasons or in the hope of leveraging some policy advantage from U.S. discomfiture."[8]

Meanwhile, sections of the English-language media appointed themselves to the role of gatekeepers and arbiters, deciding for themselves what classified information they would release to the public, according to their own definitions of national security.[9] But this approach failed to reflect the overall attitudes of internet users in the Anglosphere, and even less so those of internet users overall.

The recent growth of non-Anglophone online populations has led to a rapid movement away from Euro-Atlantic views of the nature of the internet and how it and its freedoms should be regulated. In 1996, the U.S. made up over 66% of the world's online population, whereas in 2012, it accounted for only 12%.[10] According to one assessment, India saw an increase in numbers of internet users of 32% just in the year to March 2012.[11] One effect of this shift is an adjustment in median attitudes of internet users to the ideal balance of privacy against security on the internet. Russia provides a clear example of this different approach and set of assumptions by the broad mass of users,[12] and it is for this reason that this paper uses a Russian system to compare and contrast with U.S. surveillance programmes.

3. INTERNET SURVEILLANCE TWO SYSTEMS COMPARED

In November 2013 a delegation of representatives of Russia's Federation Council (the parliament's upper house) and Foreign Ministry visited the U.S. with the intention of taking American service providers to task for not guaranteeing user privacy against government intrusion - a reversal of roles which six months earlier would have seemed laughable.[13] Yet the Snowden allegations conclusively dislodged the United States from the moral high ground of internet user freedom.

7 As reported by TIME's Moscow correspondent Simon Shuster on Twitter: https://twitter.com/shustry/status/395640131547189248

8 Nigel Inkster, "Snowden – myths and misapprehensions", IISS, 15 November 2013, http://www iiss.org/en/politics%20and%20strategy/blogsections/2013-98d0/november-47b6/snowden-9dd1

9 "Guardian worldview at root of national security row", *The Commentator*, October 10, 2013, http://www.thecommentator.com/article/4250/guardian_worldview_at_root_of_national_security_row

10 "State of the Internet in Q3 2012", comScore, December 5, 2012, http://www.comscore.com/Insights/Presentations_and_Whitepapers/2012/State_of_the_Internet_in_Q3_2012

11 "State of the Internet in Q1 2012", comScore, available at http://www.slideshare.net/alcancemg/state-of-theinternetq12012webinar-copy

12 Keir Giles, "After Snowden, Russia Steps Up Internet Surveillance", Chatham House, October 29, 2013, http://www chathamhouse.org/media/comment/view/195173

13 "U.S. ready to discuss cyber security with Russia - Ruslan Gattarov", Voice of Russia, November 15, 2013, http://voiceofrussia.com/2013_11_15/US-ready-to-discuss-cyber-security-with-Russia-Ruslan-Gattarov-6191/?print=1

A. PRISM

PRISM, an online mass electronic data collection tool operated by U.S. security agencies, was the first alleged classified monitoring and surveillance system to be made public by Snowden.[14] The word PRISM has since entered common usage as a shorthand for a whole range of different alleged U.S. surveillance and query mechanisms.[15] But for the purposes of this paper, reference will only be made to disclosures relating to this specific system. The description of this system below is drawn from media reporting, and it should be noted that no reported details have been confirmed, and furthermore much reporting on this topic substantially misunderstands and/or misrepresents the source documents. The details on PRISM repeated below are useful only to the extent that they reflect what has been presented to internet users worldwide, and they are the information on the basis of which public opinion has been formed.

It is important to note that according to the publicly available reports, PRISM is not an interception or intrusion but rather a data mining tool. This implies that PRISM is not used to break into personal computer systems, but analyses data. The data analysed is provided by companies providing internet or computing services. Hence, only data transferred to these companies is monitored by PRISM.

In June 2013, the Washington Post released a list of nine U.S. service providers known to have cooperated with the NSA. These companies were:

- **Microsoft.** In June 2013 Microsoft released a press statement claiming to have only forwarded data to the authorities if legitimised through a legally binding document.[16]

- **Google.** Google states that data is only being exchanged with the U.S. authorities when legally demanded.[17]

- **Facebook**, known originally as a social network only but expanding into other services and especially known for its massive data collection policies. Following Google, Facebook also stated that it only provides data to the U.S. authorities when legally obliged to do so.[18]

The remaining service providers were Apple, Youtube, Skype, AOL and Yahoo. Open source

14 Greenwald, Glenn and MacAskill, Ewen, "NSA Prism program taps in to user data of Apple, Google and others", *The Guardian*, June 6, 2013, http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data

15 Gellman, Barton, "U.S. surveillance architecture includes collection of revealing Internet, phone metadata", *The Washington Post*, June 16, 2013, http://www.washingtonpost com/investigations/us-surveillance-architecture-includes-collection-of-revealing-internet-phone-metadata/2013/06/15/e9bf004a-d511-11e2-b05f-3ea3f0e7bb5a_story.html

16 "Statement of Microsoft Corporation on Customer Privacy", Microsoft, June 6, 2013, http://www.microsoft com/en-us/news/press/2013/jun13/06-06statement.aspx

17 Page, Larry and Drummond, David, "Official Google Blog", June 7, 2013, http://googleblog.blogspot.de/2013/06/what.html

18 Gellmann, Barton and Poitras, Laura, "U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program", *The Washington Post*, June 6, 2013, http://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html?hpid=z1

reporting also suggested that Dropbox provided data for the PRISM programme, but Dropbox denies any knowledge of this.[19]

1) Technical Aspects

Being a data mining tool, PRISM relies heavily on multiple data sources. There are few technical details publicly available about the technical implementation of PRISM and its exact functions provided. However, as far as details are available, it seems that the collection process of PRISM is limited to providing an interface to request data from cooperating service providers. The requested data is then transferred from the service provider's database to local servers directly accessible by PRISM.

It is known that certain user actions (such as logging on or off) yield notifications in the system, initiating new data requests or suggesting the request to an operator. According to slides supposedly explaining PRISM published through the Washington Post, the data collection is initiated and operated through the FBI "Data Interception Technology Unit" (DITU). The DITU forwards the data to the NSA program "PRINTAURA" that seems to be used to control the traffic flow, passing it on to "SCISSORS" and "PROTOCOL EXPLOITATION S3132" used to distinguish between different data types (voice, video, call and internet records). The appropriate path (NUCLEON, PINWALE, MAINWAY or MARINA) is chosen accordingly for further processing/analysing of the obtained data. After having passed through these programs, the data is indexed according to a code containing information about the provider, type of data collected, source and date as well as a serial number.

The slides provided do not include information about when and how it is decided to add a user to the PRISM database, i.e. how it is decided to monitor a specific user continuously. However, this aspect is crucial to the public debate as it yields both privacy and ethical issues.

Once in the database, PRISM seems to automatically retrieve information about certain user actions, triggering a new data collection process. This implies that once a user is added to the database, legal actions such as logging in to the e-mail provider may trigger monitoring and data collection routines. The user is put under general suspicion. This practice is not uncommon in criminal investigations, but it seems that the legal barriers for the non-digital surveillance of individuals are higher than those for PRISM observations, yielding legal and ethical questions.[20]

2) Legal Aspects

PRISM was initiated by the Protect America Act under the administration of President George W. Bush. As PRISM collects data from companies under Section 702 of the Foreign Intelligence Surveillance Act (FISA) Amendments Act 2008, PRISM is operated under the supervision of the U.S. Foreign Intelligence Surveillance Court (FISC).[21] FISA regulates procedures to physically and digitally monitor and collect foreign intelligence information. The monitoring may be extended to any individual being suspected of espionage or terrorism world-wide, although the law is not applicable outside the U.S.

19 Lardinois, Frederic, "Google, Facebook, Dropbox, Yahoo, Microsoft, Paltalk, AOL And Apple Deny Participation In NSA PRISM Surveillance Program", Tech Crunch, June 6, 2013, http://techcrunch.com/2013/06/06/google-facebook-apple-deny-participation-in-nsa-prism-program/

20 "NSA slides explain the PRISM data-collection program", *The Washington Post*, June 6, 2013. http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/

21 "NSA slides explain the PRISM data-collection program", *The Washington Post*, June 6, 2013, http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/

It should be noted that knowledge of this capability and its application was already in the public domain long before disclosures by Snowden. Reporting by the New York Times in December 2005 described how the Bush administration secretly authorized the NSA to eavesdrop on both Americans and other individuals within the U.S. in order to counteract terrorism without court-approved warrants. This amendment provided the NSA with the ability to decide on the monitoring of individuals without any further court-approval necessary. Although this report led to discussions within the U.S., including both official concerns over disclosure and public concerns over privacy which foreshadowed the much more substantial debates triggered by Snowden, there appears to have been little visible impact at that time outside the U.S.[22] Now, in 2013, the extent to which FISA has been used in order to monitor both foreigners and Americans has led to controversial discussion among lawmakers, lawyers and researchers both within the U.S. and abroad, with sharply divided opinions on both the legality and constitutionality of operations.[23]

B. SORM

In marked contrast to information on PRISM, which took many internet users by surprise, large parts of the Russian internet surveillance and monitoring system have been public knowledge since their inception.

While disclosure of the capabilities of U.S. monitoring systems including PRISM provoked widespread reactions of shock in Europe (whether genuine or otherwise), reactions in Russia were tempered by the knowledge that Russia has been operating the SORM system openly, and governed by laws and regulations which are publicly accessible, for over a decade. In short, in Russia, an online public that is entirely accustomed to being monitored by the state approached the problem with a different set of presumptions.

SORM, an abbreviation for *Sistema operativno-rozysknykh meropriyatiy*, or System for Operational Investigative Activities, is a well-documented and long-established system for monitoring use of the internet through Russian internet service providers (ISPs) and enabling access to this monitoring for a range of Russian law enforcement bodies. One important contrast with PRISM is that SORM is primarily directed at collection of communications data from all communications users within Russia, whereas PRISM is a global programme mining data from selected highly specific targets worldwide. In other words, while both PRISM and SORM are capable of monitoring foreign users' data, PRISM is part of an active collection programme which "goes outside" to collect data, while SORM is instead passive and waits for the data to get "inside" the Russian national network. It is still the case, however, that some international users may be just as unaware of their data being automatically monitored through SORM as they were unaware of the potential of being monitored through U.S. systems.

Thus the legality and public acceptability, or otherwise, of covert interception of foreign nations' telecommunications raises different considerations in the Russian case from that of the U.S. At present, SORM is the only Russian programme named in the public domain with which

22 Risen, James und Lichtblau, Eric, "Bush lets U.S. spy on callers without court", *The New York Times*, December 16, 2005, http://www.nytimes.com/2005/12/16/politics/16program.html?pagewanted=print&_r=0

23 Donohue, Laura, "NSA surveillance may be legal - but it is unconstitutional", *The Washington Post*, http://www.washingtonpost com/opinions/nsa-surveillance-may-be-legal--but-its-unconstitutional/2013/06/21/b9ddec20-d44d-11e2-a73e-826d299ff459_story.html

these comparisons can be drawn; the likelihood of a Russian Snowden emerging to disclose the extent of other Russian measures directed abroad seems remote.

Sampling of opinion among Russian internet users suggests an acceptance of SORM and similar programmes based on greater relative weight given to security concerns over personal privacy, and an implicit understanding that use of the internet means a renunciation of privacy.[24] It should be noted that a significant proportion of media coverage implying criticism of Russian monitoring arrangements derives from a single source, the husband-and-wife team of Andrei Soldatov and Irina Borogan, who write and are quoted extensively on SORM and its derivatives in both Russian and foreign media.[25] Without their contributions and opinions, the open source picture on Russian internet surveillance would look substantially different.

At the same time, when legitimate concerns over online privacy are raised in Russia, official responses to them can on occasion spectacularly miss the point. For example, since mid-2013, Russia has moved to strengthen the role of the Federal Security Service (FSB) in ensuring domestic cyber security, both institutionally and technically.[26] Under a draft order sponsored by the Russian Ministry of Communications, as of July 1st 2014, Russian ISPs may be obliged to store records of all data and activities of users processed for a period of 12 hours, with provision for direct and immediate access to this information by the FSB.[27] But, it was reported, this new level of intrusion would not compromise the right to privacy because "personal information would only be available to specific organisations" rather than being made public.[28]

One under-reported potential consequence of the new requirement for 12-hour storage of user activity is a compromise of the security of the stored data. The new regulations will place a substantial financial burden on ISPs,[29] who will be under pressure to store very large quantities of data as cheaply as possible, with consequences for its secure handling. This has the potential to make Russian ISPs tempting targets for espionage and criminal activity.

Further proposed national security measures include close surveillance of visitors to the Sochi Winter Olympics 2014. According to experienced observer Mark Galeotti, intensive monitoring of electronic communications at Sochi is likely to be used as a test case for rolling out more intrusive and extensive systems than SORM, to include deep packet inspection (DPI) capability.[30] Yet media reporting of the proposed measures within Russia, including by

24 "Вы теперь интернетом как будете пользоваться?", Kommersant, October 21, 2013, http://kommersant.ru/doc/2324794

25 For example, Shaun Walker, "Russia to monitor 'all communications' at Winter Olympics in Sochi", *The Guardian*, October 6, 2013, http://www.theguardian.com/world/2013/oct/06/russia-monitor-communications-sochi-winter-olympics?CMP=twt_gu, and Andrei Soldatov, "Russia's Spying Craze", *The Moscow Times*, October 31, 2013 http://www.themoscowtimes.com/opinion/article/russias-spying-craze/488773.html

26 ГД одобряет передачу ФСБ полномочий по интернет-безопасности RIA-Novosti, November 15, 2013, http://ria.ru/defense_safety/20131115/977204644.html

27 Владислав Ъ-Новый, Елена Ъ-Черненко, Роман Ъ-Рожков, "Федеральный сервер безопасности", *Kommersant*, 21 October 2013, http://kommersant.ru/doc/2324684

28 "Хинштейн: доступ ФСБ к интернет-трафику не нарушит тайну личной жизни", RIA-Novosti, October 21, 2013, http://ria.ru/defense_safety/20131021/971490496.html

29 Keir Giles, "After Snowden, Russia Steps Up Internet Surveillance", Chatham House, October 29, 2013, http://www chathamhouse.org/media/comment/view/195173

30 Mark Galeotti, "On your marks, get set… intercept!", oDRussia, October 29, 2013, http://www.opendemocracy.net/od-russia/mark-galeotti/on-your-marks-get-set%E2%80%A6-intercept

independent media citing foreign sources, gave the impression of general indifference to plans for pervasive monitoring.

4. PERCEPTIONS OF INTERNET SURVEILLANCE

This section reviews and reflects on some of the remarkable international reactions to the debate on internet surveillance which was triggered within Europe by the Snowden defection. The selected examples demonstrate specific reactions by social groups and their leaders, which illustrate the implications of covert versus acknowledged internet monitoring and surveillance, depending on the socio-cultural background of the public. A clear distinction needs to be drawn between average societal attitudes overall, and the public reactions of leadership figures – with even sympathetic commentators noting "the EU's theatrical outraged reaction".[31]

A. Germany

Sudden and uncontrolled disclosure of monitoring and surveillance systems affecting Germany triggered interesting socio-political reactions, partly related to Germany's unique history in Europe as a nation previously divided into one state with a strong respect for individual rights, and another where state surveillance and control of the population were all-pervasive.

Although privacy and data protection are major concerns in modern Germany and treated as fundamental rights, the initial German reactions to disclosures of NSA internet monitoring activities were untroubled. In August 2013, Ronald Pofalla, Chief of Staff of the German Chancellery and Federal Minister for Special Affairs, stated that the NSA and GCHQ had acted in accordance with German law,[32] and that any scandal was now "over".[33]

Subsequently, however, it was reported in October 2013 that Chancellor Angela Merkel's personal mobile phone was under surveillance by U.S. agencies.[34] During investigation of what became known in Germany as the "Handygate affair", further monitoring of German citizens and leaders was revealed. Public disapprobation was fuelled by disconcerting allegations that the German Bundestag was being monitored from the nearby U.S. embassy. With the embassy under special protection by German police and military services, the suggestion that German taxes had been used to protect an installation spying on German leaders and citizens contributed to a strong public backlash against monitoring and surveillance activities.[35]

31 Bérénice Darnault, "Why the EU response to NSA leaks is contradictory", The World Outline, October 28, 2013, http://theworldoutline.com/2013/10/eus-response-nsa-leaks-spying-scandal-contradictory/

32 Carstens, Peter, "Pofalla: Amerikaner und Briten halten sich an deutsches Recht", Frankfurter Allgemeine Zeitung, August 1, 2013, http://www.faz.net/aktuell/politik/inland/spaehaffaere-pofalla-amerikaner-und-briten-halten-sich-an-deutsches-recht-12528037.html

33 "Pofalla erklärt NSA-Affäre für beendet", Die Zeit, August 12, 2013, http://www.zeit.de/politik/deutschland/2013-08/nsa-bnd-pofalla--bundestag-spaehaffaere-snowden-abkommen

34 "Zu Informationen, dass das Mobiltelefon der Bundeskanzlerin möglicherweise durch amerikanische Dienste überwacht wird", Bundesregierung Pressemitteilung, October 23, 2013, http://www.bundesregierung.de/Content/DE/Pressemitteilungen/BPA/2013/10/2013-10-23-merkel-handyueberwachung.html

35 Smale, Alison, "Anger Growing Among Allies on U.S. Spying", The New York Times. October 23, 2013, http://www nytimes.com/2013/10/24/world/europe/united-states-disputes-reports-of-wiretapping-in-Europe html?_r=0

Commentators compared early bland government assurances that all actions were legal, and a refusal to engage with public concerns, followed by sudden and shocking disclosures, to the erection of the Berlin Wall in 1961. With public concern directed primarily at the United States, and only occasional reminders that "the U.S. isn't the only country German intelligence believes may be spying on the country's leadership",[36], Germany was forced to remonstrate publicly with its U.S. allies, with further potential severe implications for future legitimate monitoring operations within Germany.[37]

B. Nordic States

Conversely, Nordic EU member states have challenged assumptions with their reactions in the aftermath of the Snowden defection. The debate in Nordic countries, which might ordinarily have been expected to be staunch advocates of privacy rights, has been tempered by a more specific threat perception and an acute awareness of the vulnerabilities of those states.[38] In Finland, news of a sophisticated attack and data breach at the Ministry for Foreign Affairs (MFA), which private sources blamed on Russia,[39] gave impetus to public discussion of possible new laws on legal intercept - with much of the debate focusing not on whether this should take place, but under which government agency it would best fit.[40] Swedish Foreign Minister Carl Bildt described cooperation with foreign intelligence services on communications intelligence gathering against Russia as "hardly sensational".[41] And authorities in Denmark felt sufficiently secure in the legitimacy of their work to pre-empt inaccurate reporting by journalists supplied with Snowden material by going on the record to describe previously classified collection programmes.[42]

C. United Kingdom

The British debate is coloured by the particular role of the UK in two key aspects of the 2013 disclosures on internet surveillance: the prominent role of GCHQ as a partner of the NSA in facilitating surveillance, and the prominent role of The Guardian newspaper in disseminating stolen classified information on alleged surveillance activities.

Public perception of internet surveillance by the authorities also differs in the UK. Polling suggests that "60% plus" say the intelligence services have the right amount of power to monitor activity on the internet or need more – even though there is a perceived need for more transparency and an "informed dialogue with the public".[43]

36 Anton Troianovski, "Germany to Boost Anti-Spy Efforts", *Wall Street Journal*, November 20, 2013, http://online.wsj.com/news/articles/SB10001424052702304791704579209740311164308

37 Troianovski, Anton, "Germany Warns of Repercussions from U.S. Spying", The Wall Street Journal, October 28, 2013, http://online.wsj.com/news/articles/SB10001424052702304200804579163760331107226

38 "Swedes 'not afraid' of internet surveillance", *The Local*, November 8, 2013, http://www.thelocal.se/20131108/swedes-not-worried-about-internet-surveillance-survey

39 Keir Giles, "Cyber Attack on Finland is a Warning for the EU", Chatham House, November 8, 2013, http://www.chathamhouse.org/media/comment/view/195392

40 "Verkkovalvonta keskittymässä yhdelle taholle", *Ilta-Sanomat*, 18 November 2013, http://m.iltasanomat.fi/kotimaa/art-1288622010437.html

41 "Bildt defends Sweden surveillance", *The Local*, November 3, 2013, http://www.thelocal.se/20131103/bildt-defends-sweden-surveillance

42 Claus Blok Thomsen, Jakob Sorgenfri Kjær, Jacob Svendsen, "Presset FE fortæller om dansk spionage", *Politiken*, November 20, 2013, http://politiken.dk/indland/ECE2138411/presset-fe-fortaeller-om-dansk-spionage/

43 UK Home Secretary Hazel Blears, speaking at Intelligence and Security Committee open evidence session, November 7, 2013, UK Parliament website, http://www.parliamentlive.tv/Main/Player.aspx?meetingId=14146

The appearance before Parliament's Intelligence and Security Committee of the chiefs of the three UK intelligence and security services[44] began a significant shift in public opinion.[45] Afterwards, there were indications that even the most liberal-minded of observers were beginning to realise the extent of the damage done by The Guardian's misguided crusade.[46] At the time of writing, unease at The Guardian's continued support for Snowden associate Glenn Greenwald was beginning to grow. This was aided by mistakes by both parties, including insistence on the palpably untrue assertion that limited damage had been done by releasing the files, since 850,000 individuals already had access to them,[47] and easily detected misinformation by Greenwald on the content of individual files, as in the case of allegations that millions of telephone calls in Norway had been intercepted by the NSA.[48] According to one expert assessment, Snowden "did not understand the significance of much of the material he did read and that the same was true for the newspapers that published it. The resulting confusion and misapprehensions that have taken hold within the media and shaped the public debate".[49]

Broadly, UK public opinion appears to be in line with the perception reflected in U.S. polls that releasing classified information on internet surveillance was harmful to national security[50] - to the palpable frustration of liberal journalists that the rest of the UK does not see it their way.[51] It has been argued that, in a curious parallel with Russia, this results from a higher British perception of the security interests that are at stake. As described in the Financial Times:

"The basic narrative of British history... is of a country that has had to ward off a succession of attempted foreign invasions. The role of the intelligence services in protecting the UK is both noted and celebrated… Most British citizens accept and, indeed, celebrate the role of the state in keeping the country free and independent – and the role of the intelligence services has historically been integral to that task. The threat from terrorism, as witnessed in the London bombings of 2005, has only increased the awareness of the need for good intelligence."[52]

44 Intelligence and Security Committee open evidence session, November 7, 2013, UK Parliament website, http://www.parliamentlive.tv/Main/Player.aspx?meetingId=14146

45 Catherine A. Traywick, "British Spies Aren't James Bonds, and 7 Other Things We Learned from Britain's Landmark Intelligence Hearing", Foreign Policy, November 7, 2013, http://blog.foreignpolicy com/posts/2013/11/07/british_spies_arent_james_bonds_and_7_other_things_we_learned_from_the_uks_landmar

46 Andrew Sparrow, "Guardian faces fresh criticism over Edward Snowden revelations", *The Guardian*, November 10, 2013, http://www.theguardian.com/media/2013/nov/10/guardian-nsa-revelations-edward-snowden

47 Nicholas Watt, "Threat from NSA leaks may have been overstated by UK, says Lord Falconer", *The Guardian*, November 17, 2013, http://www.theguardian.com/world/2013/nov/17/threat-nsa-leaks-snowden-files

48 Kjetil Magne Sørenes, "Dette dokumentet viser ikke overvåking av Norge, ifølge E-tjenesten", *Dagbladet*, 19 November 2013, http://www.dagbladet.no/2013/11/19/nyheter/snowden_i_norge/edward_snowden/innenriks/samfunn/30395928/

49 Nigel Inkster, "Snowden – myths and misapprehensions", IISS, 15 November 2013, http://www iiss.org/en/politics%20and%20strategy/blogsections/2013-98d0/november-47b6/snowden-9dd1

50 Scott Clement, "Poll: Most Americans say Snowden leaks harmed national security", *The Washington Post*, November 20, 2013, http://www.washingtonpost.com/politics/poll-most-americans-say-snowden-leaks-harmed-national-security/2013/11/20/13cc20b8-5229-11e3-9e2c-e1d01116fd98_story.html

51 John Naughton, "Edward Snowden: public indifference is the real enemy in the NSA affair", The Observer, October 20, 2013, http://www.theguardian.com/world/2013/oct/20/public-indifference-nsa-snowden-affair

52 Gideon Rachman, "Why the British like their spies", *Financial Times*, November 10, 2013.

5. CONSEQUENCES

The immediate consequence of Edward Snowden's distribution of classified information on alleged internet surveillance activities is a severe detriment to the national security of a number of states around the world. According to NSA Director General Keith Alexander the documents were "being put out in a way that does the maximum damage to NSA and our nation".[53] GCHQ Director Iain Lobban agrees, saying that the "cumulative effect of the global media coverage will make our job far, far harder for years to come".[54]

The defection of Snowden placed additional strain on an already challenging relationship between Russia and the U.S., with both sides expressing "disappointment" with each other, over Russia's acceptance of an application by Snowden for temporary asylum[55] and the subsequent decision by the U.S. to cancel a meeting between Presidents Obama and Putin scheduled for early September 2013.[56]

But the diplomatic effect extends beyond the U.S. and Europe. The Brazilian reaction to allegations of espionage by the USA and Canada was especially vehement.[57] Brazil will host a global conference on internet security in 2014 "to identify common objectives and ways of limiting espionage and monitoring operations".[58] Yet once again, there are indications that the outrage may be largely artificial. The suggestion that this came as a revelation to Brazil, giving rise to entirely new concerns, is belied by earlier plans for direct cable links with other countries "with the explicit aim of enhancing cyber security for the participating nations by bypassing the United States".[59]

In some cases, the diplomatic fallout has direct security consequences. For instance, diplomatic tensions between Australia and Indonesia peaked, reflected in an exchange of sexually lurid front-page cartoons in Australian and Indonesian newspapers, with the implication that surveillance of Indonesian targets "gave some kind of prurient pleasure to a brutish, hairy-legged Australia".[60] As a result, elements of intelligence cooperation between the two nations have been suspended, which is expected to result in an increased terrorism and criminal threat to Australia.[61]

53 Mark Hosenball, "NSA chief says Snowden leaked up to 200,000 secret documents", Reuters, November 14, 2013, http://www reuters.com/article/2013/11/14/us-usa-security-nsa-idUSBRE9AD19B20131114

54 Sir Iain Lobban, Director, GCHQ, speaking at Intelligence and Security Committee open evidence session, November 7, 2013, UK Parliament website, http://www.parliamentlive.tv/Main/Player.aspx?meetingId=14146

55 Luhn, Alec, Harding, Luke and Lewis, Paul, "Edward Snowden asylum: U.S. 'disappointed' by Russian decision", *The Guardian*, August 2, 2013, http://www.theguardian.com/world/2013/aug/01/edward-snowden-asylum-us-disappointed

56 "Russia 'disappointed' bilateral talks with U.S. cancelled", BBC, August 7, 2013, http://www.bbc co.uk/news/23608052

57 Tamara Santos, "Why is everyone spying on Brazil?", The World Outline, October 13, 2013, theworldoutline.com/2013/10/everyone-spying-brazil/

58 Tamara Santos, "Why is everyone spying on Brazil?", The World Outline, October 13, 2013, theworldoutline.com/2013/10/everyone-spying-brazil/

59 Keir Giles, "Russian Interests In Sub-Saharan Africa", U.S. Army War College Strategic Studies Institute, July 2013, p. 34.

60 Michael Bachelard, "Australia's reputation in Indonesia hits new low", *The Age*, November 23, 2013, http://m.theage.com.au/federal-politics/political-news/australias-reputation-in-indonesia-hits-new-low-20131123-2y2k2.html

61 John Schindler, "Snowden's Thunder Down Under", The XX Committee, November 21, 2013, http://20committee.com/2013/11/21/snowdens-thunder-down-under/

But in addition to the long-term national security implications, there have been direct and immediate consequences in both commercial and legal terms in a number of countries. "Fears about the NSA using American hardware to spy on the rest of the world"[62] have led to severe revenue implications for U.S. companies, with major players such as CISCO and IBM suffering badly.[63] As pointed out by Nigel Inkster, "the major U.S. technology companies and service providers which have to varying degrees collaborated with the NSA, either voluntarily or in response to judicial warrants, have experienced a decline in trust with uncertain but potentially significant implications for their future business prospects."[64] Businesses promoting cloud services in particular have reportedly experienced a significant drop in demand due to security fears, while firms in Switzerland are benefiting from that country's current perceived status as unaffected by surveillance concerns.[65]

Most significantly for the purposes of this paper, one trend that was beginning to be observed at the time of writing is the move towards public legitimisation of internet interception and surveillance activities.

A conference at London's Chatham House in late November 2013 heard how online activity worldwide was in effect being governed by U.S. law, while in the USA itself, the response to disclosures of NSA activities was calls across the political spectrum not for a reduction in the extent of surveillance, but for greater oversight of its implementation.[66] In its work with overseas intelligence-gathering organisations, the NSA had been restricted, or in some cases assisted, by very different legal environments in the partner country. An unattributed document released in December 2013 and purporting to review NSA cooperation agreements with a range of foreign partner organisations refers to "legal and policy impediments on the partner side".[67] In a possibly unrelated example, domestic legal considerations caused the Japanese government to decline NSA requests for cooperation in tapping cables carrying phone and Internet data across the Asia-Pacific region in 2011.[68] But after October 2013, a number of European countries have moved to establish or reinforce a firm legal framework for their own interception and surveillance activities.

There are numerous and varying assessments of the legality of interception of communications in Europe, even within the narrow focus of privacy as a human rights issue. According to a draft of the "EU Human Rights Guidelines on Freedom of Expression Online and Offline",

62 Christopher Mims, "Cisco's disastrous quarter shows how NSA spying could freeze U.S. companies out of a trillion-dollar opportunity", Quartz, November 14, 2013, http://qz.com/147313/ciscos-disastrous-quarter-shows-how-nsa-spying-could-freeze-us-companies-out-of-a-trillion-dollar-opportunity/

63 Cyrus Farivar, "Cisco attributes part of lowered earnings to China's anger toward NSA", Ars Technica, November 14, 2013, http://arstechnica.com/business/2013/11/cisco-attributes-part-of-lowered-earnings-to-chinas-anger-towards-nsa/

64 Nigel Inkster, "Snowden – myths and misapprehensions", IISS, 15 November 2013, http://www iiss.org/en/politics%20and%20strategy/blogsections/2013-98d0/november-47b6/snowden-9dd1

65 Varying estimates given by multiple industry speakers at "e-Crime & Information Security Mid Year Meeting", London, October 24, 2013

66 "Power and Commerce in the Internet Age", Chatham House, London, November 25-26 2013, agenda available at http://www.chathamhouse.org/Internet2013/agenda

67 Unattributed document provided by Swedish SVT television's "Uppdrag Granskning" investigative programme, available at https://www.documentcloud.org/documents/894386-legal-issues-uk-regarding-sweden-and-quantum.html

68 "NSA asked Japan to tap regionwide fiber-optic cables in 2011", *The Japan Times*, October 27, 2013, http://www.japantimes.co.jp/news/2013/10/27/world/nsa-asked-japan-to-tap-regionwide-fiber-optic-cables-in-2011/#.UnqQx3B7KsY

"lack of respect for the right of privacy and data protection constitutes a restriction of freedom of expression. Illegal surveillance of communications, their interception, as well as the illegal collection of personal data violates the right to privacy and freedom of expression."[69]

Yet in 2007, the European Court of Human Rights ruled as inadmissible (manifestly ill-founded) a complaint by an Italian internet user under Article 8 (right to respect for private and family life) of the European Convention on Human Rights. Although the complaint related to spam rather than surveillance, the Court declared that "once connected to the Internet, e-mail users no longer enjoyed effective protection of their privacy".[70]

As noted above, a cyber attack on the Finnish Ministry for Foreign Affairs (MFA) spurred attempts there to legitimise active defence, in the form of pre-emptively screening both data traffic within Finland and that which passes through Finnish cables, as opposed to the current state of legislation where data can only be intercepted once a crime is suspected and an investigation in progress. The aim, according to the Finnish Minister of Defence, would be to enable Finland "to prevent and intervene if another country's intelligence operations focus on Finland and Finnish officials."[71]

In an apparent direct reference to the MFA attack, which Finland learned of through a tipoff from Sweden's FRA signals intelligence agency, National Police Commissioner Mikko Paatero noted that "we cannot follow signals in Finland or travelling through Finnish cables... but others can do it for Finland. In my opinion it's a little bit embarrassing that we can hear from somewhere else about what is happening here."[72] Meanwhile in Sweden, although interception is already legal under the "FRA Law", the authorities are now seeking to enhance their powers in a similar manner to Russia.[73]

Most recently at the time of writing, a law was passed in France in December 2013 allowing surveillance of internet users in real time and without prior legal authorisation, by a much increased range of public officials including police, gendarmes, intelligence and anti-terrorist agencies as well as several government ministries.[74] The law gave rise to accusations of cynicism, being passed just weeks after France expressed outrage that the NSA had allegedly been engaged in similar activities, at which President François Hollande expressed his "extreme reprobation".[75]

69 Draft "EU Human Rights Guidelines on Freedom of Expression Online and Offline", unpublished, version as at November 20, 2013.

70 Muscio v. Italy, European Court of Human Rights, "Information Note on the Court's case-law No. 102", November 2007, http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=002-2419

71 "Defence Minister: Police and defence forces to get wider web powers", Yle News, November 2, 2013, http://yle.fi/uutiset/defence_minister_police_and_defence_forces_to_get_wider_web_powers/6914546?origin=rss

72 "Finnish Police want web snooping powers", Yle news, November 7, 2013, http://yle fi/uutiset/finnish_police_want_web_snooping_powers/6923309

73 "Intel agency seeks direct access to Swedes' data", *The Local*, November 19, 2013, http://www.thelocal.se/20131119/swedens-security-service-seeks-direct-data-access

74 "Adoption de la loi controversée de programmation militaire", *Le Monde*, December 10, 2013, http://www lemonde.fr/international/article/2013/12/10/adoption-definitive-de-la-controverse-loi-de-programmation-militaire_3528927_3210.html

75 Kim Willsher, "French officials can monitor internet users in real time under new law", *The Guardian*, December 11, 2013, http://www.theguardian.com/world/2013/dec/11/french-officials-internet-users-real-time-law

In this way, disclosure of alleged surveillance activities by the NSA and GCHQ is having the effect, probably unanticipated by the disclosers, of ensuring that more of the U.S. and UK's partner nations are ensuring they have the legal framework in place to be able to participate in this activity on an unarguably legitimate basis.

6. CONCLUSION

Comparison of Russian, U.S. and British attitudes to internet monitoring demonstrates clearly that the common perception of legitimacy of that monitoring varies widely between nations.

Varying reactions to prior knowledge of Russian, and sudden disclosure of U.S. monitoring systems demonstrate that public responses are heavily influenced not only by national attitudes towards public security, but also by the extent of awareness of monitoring. A balance needs to be sought between the positive benefits of public knowledge of the precise limitations of privacy online, and the negative national and international security implications of widespread awareness of monitoring capabilities.

Direct comparison of the public reactions to PRISM and SORM supports this conclusion. Criticism of the aims and methods of PRISM and related systems was fuelled by their necessary lack of transparency. Failure to initiate public discussion about the nature of the threats which PRISM is intended to counter, and the nature of the counter-measures required, left the field open for wide-ranging and misinformed speculation. In particular, media coverage downplayed the legal controls and safeguards in place to protect the domestic US population from abuses of these capabilities. This situation was exacerbated by restraints on the U.S. intelligence community, which has been prevented from joining or contributing to the public narrative to correct speculation by the need to preserve what secrecy remains by not confirming or denying the accuracy of media allegations. By contrast, SORM is a system publicly avowed in the context of a well-developed threat narrative, and consequently does not excite similar reactions or wildly misinformed reporting.

Although disclosure of the alleged capability and reach of U.S. and allied surveillance mechanisms prompted strident and outraged reportage in some sections of the English-language media, public opinion has not followed suit. Instead, a more balanced and sober assessment of national security needs is leading European states to pass legislation through due democratic process to ensure that internet monitoring of specific threats to security continues unhindered. It follows that active cyber defence in the sense of active measures online in order to prevent and pre-empt threats to national security will continue to be perceived as legitimate, and these measures should be expected to continue unrestrained by the new environment of enhanced public awareness.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

The Drawbacks and Dangers of Active Defense

Oona A. Hathaway
Yale Law School
New Haven, CT
oona.hathaway@yale.edu

Abstract: The growing prevalence of cyber-attacks on states, businesses, and individuals has raised new and urgent questions about the legal framework that governs states' capacity to respond such attacks. An issue that has proven particularly vexing is what actions a state may take in response to attacks that fall into the gap between the actions that constitute a prohibited "use of force" under Article 2(4) of the UN Charter and the "armed attacks" to which a state has a right to respond with force in self defense under Article 51. Intrusions that constitution an illegal "use of force" but do not meet the "armed attack" threshold for triggering a legal forceful response—sometimes known as "below the threshold" cyber-operations—are extraordinarily common. Indeed, nearly all cyber-attacks by one state on another fall below the "armed attack" threshold. If states cannot legally use their right to self-defense to respond to such unlawful attacks, what can they do? There is a growing consensus that the answer can be found in countermeasures doctrine. Yet countermeasures doctrine was never intended to be applied to actions that constitute uses of force. There is good reason for this: if forceful countermeasures were allowed, there would be a serious danger that the system restricting illegal use of force would spin out of control. Improper countermeasures are inevitable, and escalation of conflict only a matter of time. This paper outlines the legal principles governing the use of force in international affairs, describes the exceptions to the broad prohibition on the use of military force, outlines the doctrine of countermeasures, and—in its key contribution to the debate—outlines reasons for concern about aggressive countermeasures. The paper concludes by briefly considering non-forceful responses that states may take in response to cyber-attacks.

Keywords: *Active defense, United Nations Charter, international law, self-defense*

Cyber-attacks have become an ever-present threat to states, individuals, and businesses throughout the world.[1] British Petroleum has reported that it faces a barrage of 50,000 attempts at cyber-intrusion a day.[2] The U.S. Pentagon has reported ten million attempts per day.[3] The U.S. National Nuclear Security Administration also records ten million attempts at hacking each day.[4] If only one out of one hundred million attacks succeeds, the national security of the United States is dangerously vulnerable.

These new threats to national security have raised deep questions about the capacity of states to protect themselves. In response, the legal framework that governs the use of force in the cyber context has been slowly taking shape. There is a growing consensus that the standard rules governing use of force in international law apply to this unconventional threat. The Tallinn Manual, now in the midst of revision and expansion, represents an extraordinary collaboration of scholars seeking to outline the specific implications of that law for cyber.[5]

An issue that has proven particularly vexing is the gap between the actions that constitute a prohibited "use of force" under Article 2(4) of the UN Charter and the "armed attacks" to which a state has a right to respond with force in self defense under Article 51. There is a well-known gap between those intrusions that are illegal and those that meet the "armed attack" threshold for triggering a legal forceful response.[6] These "below the threshold" cyber-operations, as Michael Schmitt has dubbed them, are extraordinarily common. Indeed, nearly all cyber-attacks by one state on another fall below the "armed attack" threshold.

If states cannot legally use their right to self-defense to respond to unlawful attacks below the threshold, what can they do? There is a growing consensus that the answer can be found in countermeasures doctrine. States, the argument goes, may respond in kind to an attack as long as they meet the various requirements of countermeasures doctrine—most notably that the countermeasure is proportional to the unlawful behavior that prompted it and is designed to bring the violating state back into compliance.

This paper aims to sound a cautionary note in the face of this growing consensus. It points out that countermeasures doctrine has never been applied in the use of force context and, indeed, commentary on the countermeasures doctrine makes clear that it was not intended to be applied to actions that constitute uses of force. There is, moreover, a good reason for this: if millions of "below-the-threshold" attacks are met with millions of "below-the-threshold" attacks in

1 Portions of this paper are drawn from Oona A. Hathaway et al, *The Law of Cyber-Attack*, 100 Cal. L. Rev. 817 (2012).

2 Michael Tomaso, BP Fight Off Up to 50,000 Cyber0Attacks a Day, CNBC (Mar. 6, 2013).

3 Zachary Fryer-Biggs, U.S. Military Goes on Cyber Offensive, Defense News (Mar. 24, 2012).

4 Jason Koebler, U.S. Nukes Face up to 10 Million Cyber Attacks Daily, U.S. News (Mar. 20, 2012).

5 Tallinn Manual (Michael Schmitt, ed., 2013). Its editor, Michael Schmitt, has also addressed many of the most interesting an important legal challenges relating to the application of the law of jus ad bellum and jus in bello to cyber in his own extensive writings.

6 Harold Koh, while serving as Legal Adviser for the U.S. Department of State, took the position that there was no gap. Koh stated that "the inherent right of self-defense potentially applies against any illegal use of force... There is no threshold for a use of deadly force to qualify as an 'armed attack' that may warrant a forcible response." Michael N. Schmitt, *International Law in Cyberspace The Koh Speech and Tallinn Manual Juxtaposed*, 54 Harvard Int'l L.J. 21-22 (Dec. 2012). Most scholars disagree with this view, concluding that there is, in fact, a gap between the two. *See id*.; Tom Ruys, 'Armed Attack' and Article 51 of the UN Charter: Evolutions in Customary Law and Practice 139-84 (2010). Randelzhofer shows sympathy for closing the gap between Articles 2(4) and 51 by allowing states to respond to any use of force but expresses doubt about whether that view is consistent with the Charter. A. Randelzhofer, Article 51, in B. Simma et al, eds., The Charter of the United Nations: A Commentary, Vol 1 (2002), at pp. 791-92.

response, there is a serious danger that the system restricting illegal use of force will spin out of control. Improper countermeasures are inevitable, and escalation of conflict only a matter of time.

This paper proceeds in four parts. First, it briefly outlines the legal principles governing the use of force in international affairs. Second, it describes the exceptions to the broad prohibition on the use of military force. Third, it outlines the doctrine of countermeasures. Fourth—in its central contribution to the debate—the paper explains the reasons for concern about aggressive countermeasures. It concludes by briefly considering non-forceful responses that states may take in response to cyber-attacks.

1. GOVERNING LEGAL PRINCIPLES: PROHIBITION ON USE OF FORCE AND INTERVENTION IN INTERNAL AFFAIRS

Article 2(4) of the U.N. Charter provides that member states "shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations."[7] This prohibition is complemented by a customary international law norm of non-intervention, which prohibits states from interfering in the internal affairs of other states.[8] The International Court of Justice ("ICJ") has held that, where the interference takes the form of a use or threat of force, the customary international law norm of non-intervention is coterminous with Article 2(4).[9]

The precise scope of the international prohibition on the threat or use of force has been the subject of intense international and scholarly debate. Weaker states and some scholars have argued that Article 2(4) broadly prohibits not only the use of armed force, but also political and economic coercion. Nonetheless, the consensus is that Article 2(4) prohibits only armed force.[10]

Discussions about cyber-attacks have the potential to reignite debates over the scope of Article 2(4).[11] Because it is much less costly to mount cyber-attacks than to launch conventional

7 U.N. Charter art. 2, para. 4.

8 *See* G A. Res. 37/10, U.N. Doc. A/RES/37/10 (Nov. 15, 1982); G.A. Res. 25/2625, U N. Doc. A/RES/25/2625 (Oct. 24, 1970).

9 Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), 1986 I.C.J. 14, para. 209 (June 27) ("[A]cts constituting a breach of the customary principle of non-intervention will also, if they directly or indirectly involve the use of force, constitute a breach of the principle of non-use of force in international relations."). It is possible, however, that to the extent cyber-attacks do not constitute a use of force, they may nevertheless violate the customary international law norm of non-intervention, as discussed below.

10 Daniel B. Silver, *Computer Network Attack as a Use of Force Under Article 2(4) of the United Nations Charter*, in Computer Network Attack and International Law 73, 80–82 (Michael N. Schmitt & Brian T. O'Donnell eds , 2002). The principal arguments for the prevailing view are: (1) that Article 2(4) was conceived against a background of efforts to limit unilateral recourse to armed force, not economic and political coercion; (2) that the *travaux preparatoires* show that the San Francisco Conference rejected a proposal that would have extended Article 2(4) to include economic sanctions; and (3) that the ICJ has held that financing armed insurrection does not constitute force, indicating that other economic measures that are even less directly related to armed violence would not constitute prohibited force either. Id. at 81. There remains some ambiguity, however, as to the extent to which Article 2(4) prohibits non-military physical force, such as flooding, forest fires, or pollution. *Id*. at 82–83.

11 *See* Matthew C. Waxman, Cyber-Attacks and the Use of Force: Back to the Future of Article 2(4), 36 YALE J. INT'L L. 421, 458-59 (2011).

attacks, and because highly industrialized states are generally more dependent upon computer networks and are more vulnerable to cyber-attacks, cyber-attacks may prove to be a powerful weapon of the weak. This change in the cost structure of offensive capabilities may both increase the likelihood of cyber-attacks and change the political valence of different interpretations of Article 2(4)'s scope. Stronger states may begin to favor more expansive readings of Article 2(4) that prohibit coercive activities like cyber-attacks.[12]

Cyber-attacks may also violate the customary international law norm of non-intervention, as defined by a growing record of state practice and *opinio juris*. First, states generally do not engage in cyber-attacks openly, but rather try to hide their responsibility by camouflaging attacks through technical means[13] and by perpetrating the attacks through non-state actors with ambiguous relationships to state agencies.[14] As Thomas Franck has observed, "[l]ying about facts . . . is the tribute scofflaw governments pay to international legal obligations they violate."[15] In other words, the very fact that states attempt to hide their cyber-attacks may betray a concern that such attacks may constitute unlawful uses of force. Second, when states acknowledge that they have been victims of cyber-attack, they and their allies tend to denounce and condemn the attacks.[16] Third, in its common approach to cyber-defense, NATO has indicated that cyber-attacks trigger states parties' obligations under Article 4 of the North Atlantic Treaty,[17] which applies only when "the territorial integrity, political independence or security of any of the Parties is threatened."[18] The invocation of this provision strongly suggests that NATO member states believe that cyber-attacks violate the customary norm of non-intervention or a related international law norm.[19] Still, as the next Section explains, the fact that a cyber-attack is unlawful does not necessarily mean that armed force can be used in response.

2. EXCEPTIONS FOR COLLECTIVE SECURITY AND SELF□DEFENSE

Article 2(4)'s blanket prohibition on the non-consensual use or threat of force is subject to two exceptions: actions taken as part of collective security operations and actions taken in self-defense.

12 Walter Sharp has advocated that the United States make precisely this kind of strategic interpretive move, arguing that a broad array of coercive cyber-activities should fall within Article 2(4)'s prohibition. Walter Gary Sharp, Sr., CyberSpace and the Use of Force 129–33 (1999).

13 *See* Matthew J. Sklerov, *Solving the Dilemma of State Responses to Cyberattacks A Justification for the Use of Active Defenses Against States Who Neglect their Duty to Prevent*, 201 Mil. L. Rev., Fall 2009, at 1, 74–75.

14 *See, e.g.*, Jeffrey Carr, Inside Cyber Warfare 176 (2010), at 29 ("Hacking attacks cloaked in nationalism are not only not prosecuted by Russian authorities, but they are encouraged through their proxies, the Russian youth associations, and the Foundation for Effective Policy.").

15 Thomas M. Franck, *Legitimacy After Kosovo and Iraq*, *in* International Law and the Use of Force at the Turn of Centuries: Essays in Honour of V. D. Degan 69, 73 (Vesna Crnić-Grotić & Miomir Matulović eds., 2005).

16 *See, e.g.*, Ian Traynor, *Russia Accused of Unleashing Cyberwar to Disable Estonia*, Guardian, May 16, 2007, http://www.guardian.co.uk/world/2007/may/17/topstories3.russia (detailing the reactions by Estonian, EU, and NATO officials to a cyber-attack on Estonia).

17 *NATO Agrees Common Approach to Cyber Defence*, Euractiv.com (Apr. 4, 2008), http://www euractiv.com/en/infosociety/nato-agrees-common-approach-cyber-defence/article-171377.

18 North Atlantic Treaty, art. 4, Apr. 4, 1949, 63 Stat. 2241, 34 U.N.T.S. 243.

19 NATO has not endorsed the view that cyber-attacks rise to the level of armed attacks justifying self defense. *See NATO Agrees Common Approach to Cyber Defence*, *supra* note 17.

The first exception falls under Article 39 of the U.N. Charter. Article 39 empowers the Security Council to "determine the existence of any threat to the peace, breach of the peace, or act of aggression, and [to] make recommendations, or decide what measures shall be taken . . . to maintain or restore international peace and security."[20] The Security Council may employ "measures not involving the use of armed force"[21] and authorize "action by air, sea, or land forces."[22] Collective security operations under Article 39 can be politically difficult, however, because they require authorization by the often deadlocked or slow-moving Security Council.

The second exception to Article 2(4) is codified in Article 51, which provides that "[n]othing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs."[23] Lawful self-defense can be harder to define and identify than lawful collective security operations. Indeed, in many armed conflicts, both sides claim to be acting in self-defense, and the international debates tend to focus on factual and political disputes rather than legal doctrine.[24] It is clear, however, that the critical question determining the lawfulness of self-defense is whether or not an armed attack has occurred. A cyber-attack must rise to the level of an armed attack for a state to lawfully respond under Article 51.[25]

In scholarly debates over the application of *jus ad bellum* to cyber-attacks, three leading views have emerged to determine when a cyber-attack constitutes an armed attack that triggers the right of armed self-defense: the instrument-based approach, the target-based approach, and the effects-based approach.[26] Scholarly judgment has largely coalesced around the effect-based approach.[27] In essence, that approach holds that an attack is judge by its effects. For example, Daniel Silver, former General Counsel of the CIA and National Security Agency, argues that the key criterion determining when a cyber-attack constitutes an armed attack is the severity of the harm caused. A cyber-attack justifies self-defense "only if its foreseeable consequence is to cause physical injury or property damage and, even then, only if the severity

20 U.N. Charter art. 39.

21 *Id*. art. 41.

22 *Id*. art. 42.

23 *Id*. art. 51. For example, the White House's recent cyberspace strategy paper includes the right of self-defense as one of the norms that should guide conduct in cyberspace. International Strategy for Cyberspace, White House 5 (May, 2011), [hereinafter White House Cyberspace Strategy] available at http://www.whitehouse.gov/sites/default/files/rss_viewer/ international_strategy_for_cyberspace.pdf. at 10.

24 Christine Gray, International Law and the Use of Force 95–96 (2d ed. 2004).

25 *See, e.g.*, International Strategy for Cyberspace, White House 5 (May, 2011), [hereinafter White House Cyberspace Strategy] *available at* http://www.whitehouse.gov/sites/default/files/rss_viewer/international_strategy_ for_cyberspace.pdf, at 14 ("When warranted, the United States will respond to hostile acts in cyberspace as we would to any other threat to our country. All states possess an inherent right to self-defense, and we recognize that certain hostile acts conducted through cyberspace could compel actions under the commitments we have with our military treaty partners.").

26 Once a state has been the victim of an armed attack, a further question arises as to against whom the state can respond. Where the armed attack is perpetrated by a state, this question is easily answered—self-defense may be directed against the perpetrating state. However, cyber-attacks may be perpetrated by non-state actors or by actors with unclear affiliations with state security agencies. Although some scholars argue that cyber-attacks (and conventional attacks) must be attributable to a perpetrating state in order for the victim state to take defensive action that breaches another state's territory, others—drawing on traditional jurisprudence on self-defense—argue that states possess the right to engage in self-defense directly against non-state actors if certain conditions are met. See Jordan J. Paust, *Self-Defense Targetings of Non-State Actors and Permissibility of U.S. Use of Drones in Pakistan*, 19 J. Transnat'l L. & Pol'y 237, 238–39 (2010) ("The vast majority of writers agree that an armed attack by a non-state actor on a state, its embassies, its military, or other nationals abroad can trigger the right of self-defense addressed in Article 51 of the United Nations Charter, even if selective responsive force directed against a non-state actor occurs within a foreign country.").

27 *See* Hathaway, et al, *supra* note 1.

of those foreseeable consequences resembles the consequences that are associated with armed coercion."[28] Under this test, a cyber-attack on the air traffic control system causing planes to crash would be regarded as an armed attack, because it is foreseeable that such an attack would cause loss of life and substantial property damage. But a cyber-attack on a website or mere penetration of a critical computer system generally would not, unless it caused physical injury or property damage. A cyber-attack on financial systems presents a harder case for this approach—the analysis would depend on whether the attack was found to have caused substantial "property damage." This effects test defines a small core of harmful cyber-attacks that rise to the level of an armed attack.[29] It also focuses the armed attack analysis on a limited set of criteria—particularly severity and foreseeability.[30]

The effects test solves the problem of how to judge the severity of a cyber attack. But it leaves intact the problem of a gap between the uses of force that constitute a violation of Article 2(4) and armed attacks sufficient to give rise to the right to respond with force under Article 51. Indeed, the "armed attack" is linguistically distinct from several other related terms in the U.N. Charter and has been interpreted to be substantively narrower than them.[31] The ICJ has indicated that cross-border incursions that are minor in their "scale and effects" may be classified as mere "frontier incident[s]" rather than "armed attacks."[32] Instead, to be armed attacks sufficient to justify a response under Article 51, attacks must be of sufficient gravity to constitute "most grave forms of the use of force."[33] Where they may not resort to defensive force under Article 51 (because an attack does not arise to the level of an "armed attack"), states may be permitted to respond with retorsions or non-forceful countermeasures within carefully proscribed legal limits.[34]

28 Silver, *supra* note 10, at 90–91. It is important to note that the purpose of the attack is already accounted for in the definition of cyber-attack recommended herein: the attack must have been committed for a political or national security purpose. Therefore a cyber-attack that has unforeseen national security consequences would not be considered a cyber-attack, much less cyber-warfare.

29 *Id*. at 92.

30 The Department of Defense has signaled its approval of this approach. *See* Office of Gen. Counsel, Dep't of Def., An Assessment of International Legal Issues in Information Operations (1999), *reprinted* in Computer Network Attack and International Law 459, 484–85 [hereinafter DOD Memo] (Michael N. Schmitt & Brian T. O'Donnell eds., 2002), at 483 (arguing "the consequences are likely to be more important than the means used," and providing examples of cyber-attacks that would cause civilian deaths and property damage).

31 *See* Yoram Dinstein, Computer Network Attacks and Self-Defense, *in* Computer Network Attack and International Law 99, 100–01 (Michael N. Schmitt & Brian T. O'Donnell eds., 2002).

32 Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), 1986 I.C.J. 14, ¶ 195 (June 27); *cf*. Definition of Aggression, G.A. Res. 29/3314, Annex, art. 2, U N. Doc. A/RES/29/3314 (Dec. 14, 1974) [hereinafter Definition of Aggression] (determining that "[t]he first use of armed force by a State in contravention of the Charter shall constitute *prima facie* evidence of an act of aggression although the Security Council may . . . conclude that a determination that an act of aggression has been committed would not be justified in the light of other relevant circumstances, including the fact that the acts concerned or their consequences are not of *sufficient gravity*" (emphasis added)) . Scholars generally agree that there is a gap between the prohibition on the use of force and the right of self-defense. *See, e.g*., Dinstein, *supra* note 31, at 99, 100–01.

33 Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), 1986 I.C.J. 14, ¶ 191 (June 27).

34 Retorsions are lawful unfriendly acts made in response to an international law violation by another state; countermeasures are acts that would be unlawful if not done in response to a prior international law violation. U.N. Int'l Law Comm'n Draft Articles on Responsibility of States for Internationally Wrongful Acts, Rep. of the Int'l Law Comm'n, U.N. GAOR, 53d Sess., Supp. No. 10, U.N. Doc. A/56/10 (2001), at 31, 80 [hereinafter Draft Articles]. *See* DOD Memo, *supra* note 30 ("If the provocation is not considered to be an armed attack, a similar response will also presumably not be considered to be an armed attack.").

Until recently, forceful countermeasures were generally regarded as outside the countermeasures regime. As the next section explores, however, that consensus has begun to crumble as a growing number of voices have called for forceful countermeasures for cyber.

3. COUNTERMEASURES

The customary international law of countermeasures governs how states may respond to international law violations that do not rise to the level of an armed attack justifying self-defense—including, implicitly, cyber-attacks. The Draft Articles on State Responsibility define countermeasures as "measures that would otherwise be contrary to the international obligations of an injured State *vis-à-vis* the responsible State, if they were not taken by the former in response to an internationally wrongful act by the latter in order to procure cessation and reparation."[35]

The international law of countermeasures does not define when a cyber-attack is unlawful—indeed the Draft Articles do not directly address cyber-attack at all. The law simply provides that when a state commits an international law violation, an injured state may respond with a countermeasure.[36] As explained above, some cyber-attacks that do not rise to the level of an armed attack nonetheless violate the customary international law norm of non-intervention.[37] These violations of international law may entitle a harmed state to use countermeasures to bring the responsible state into compliance with the law.

The Draft Articles lay out the basic customary international law principles regulating states' resort to countermeasures.[38] The Draft Articles provide that countermeasures must be targeted at the state responsible for the prior wrongful act and must be temporary and instrumentally directed to induce the responsible state to cease its violation.[39] Accordingly, countermeasures cannot be used if the international law violation has ceased. Countermeasures also can never justify the violation of fundamental human rights, humanitarian prohibitions on reprisals, or peremptory international norms, nor can they excuse failure to comply with dispute settlement procedures or to protect the inviolability of diplomats.[40]

35 Draft Articles, *supra* note 34, at 128. Traditionally, these acts were termed "reprisals," but this report follows the Draft Articles in using the more modern term "countermeasures." Reprisals now predominantly refer to forceful belligerent reprisals. *Id*.

36 States thus resort to countermeasures at their own risk. If the use of countermeasures does not comply with the applicable international legal requirements, the state may itself be responsible for an internationally wrongful act. *Id*. at 130.

37 *See* Hathaway et al, *supra* note 1.

38 Countermeasures are distinct from retorsions. Retorsions are acts that are unfriendly but lawful, such as limiting diplomatic relations or withdrawing from voluntary aid programs, and they always remain a lawful means for a State to respond to a cyber-attack or other international legal violation.

39 Draft Articles, *supra* note 34, at 129. Accordingly, the law of countermeasures does not specify how states may respond to international law violations by non-state actors. However, international law violations by non-state actors often lead to international law violations by states. For example, if a non-state actor launches an attack on state A from state B's territory and state B is unwilling or unable to stop it, state B may violate an international law obligation to prevent its territory from being used for cross-border attacks. *See, e.g.*, Corfu Channel Case (U.K. v. Albania) (Merits), 1949 I.C J. 4, 22 (Apr. 9) (holding that states are obligated "not to allow knowingly its territory to be used for acts contrary to the rights of other States"). In the cyber-attack context, a state may commit an international law violation by allowing harmful cyber-attacks to be launched from its territory. *See* Sklerov, *supra* note 13, at 62–72.

40 Draft Articles, *supra* note 34, at 131.

Before resorting to countermeasures, the injured state generally must call upon the responsible state to cease its wrongful conduct, notify it of the decision to employ countermeasures, and offer to negotiate a settlement.[41] However, in some situations, the injured state "may take such urgent countermeasures as are necessary to preserve its rights."[42] Countermeasures need not necessarily be reciprocal, but reciprocal measures are favored over other types because they are more likely to comply with the requirements of necessity and proportionality.[43] Under the customary law of countermeasures, an attacking state that violates its obligation not to intervene in another sovereign state through a harmful cyber-attack may be subject to lawful countermeasures by the injured State.

A rising number of institutions and scholars have left the door open to active countermeasures in response to illegal cyber-attacks. In this view, countermeasures might go beyond "passive defenses," such as firewalls, that aim to repel cyber-attacks, and constitute "active defenses," which attempt to disable the source of an attack.[44] Active defenses—if properly designed to meet the requirements of necessity and proportionality—might be considered a form of "reciprocal countermeasures," in which the injured state ceases obeying the same or a related obligation to the one the responsible state violated (in this case, the obligation of non-intervention).

Before a state may use active defenses as a countermeasure, however, it must determine that an internationally wrongful act caused the state harm and identify the state responsible, as well as abide by other restrictions.[45] The countermeasures must be designed, for example, to induce the wrongdoing state to comply with its obligations. The Draft Articles also have detailed provisions regarding when acts committed by non-state agents may be attributed to a state—for instance, when the state aids and assists the act with knowledge of the circumstances.[46] Countermeasures must also be necessary and proportional. Though there is no requirement that countermeasures are taken in relation to the same or a closely related obligation, the Commentary notes that necessity and proportionality will be more likely to be satisfied if they are.[47]

While countermeasures provide states with a valuable tool for addressing cyber-attacks that do not rise to the level of an armed attack, countermeasures are far from a panacea. Even putting to one side concerns about legality, there are practical challenges to an active countermeasures regime. First and foremost, cyber countermeasures require the identity of the attacker and the computer or network from which the attack originates to be accurately identified. Second, in order for a countermeasure to be effective, the targeted actor must find the countermeasure

41 *Id*. at 135.
42 *Id*.
43 *Id*. at 129.
44 In 2011, the Department of Defense has made clear that it employs such "active cyber defense" to "detect and stop malicious activity before it can affect DoD networks and systems." U.S. Dep't of Def., Department of Defense Strategy for Operating in Cyberspace 2 (July 2011) [hereinafter Dod Strategy], at 7; see Comm. on Offensive Info. Warfare, Nat'l Research Council of the Nat'l Acads., Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities 38 (William A. Owens et al. eds., 2009) [hereinafter NRC REPORT], at 142-49 (outlining possible "active responses" to cyber-attacks); Jay P. Kesan & Carol M. Hayes, Mitigative Counterstriking: Self-Defense and Deterrence in Cyberspace, 25 Harv. J. L. & Tech 415 (2012) (arguing that "permitting mitigative counterstrikes in response to cyberattacks would be more optimal" than the current passive regime). *Cf.* Tallinn Manual; Michael N. Schmitt, "Below the Threshold" Cyber Operations: The Countermeasures Response Option and International Law, 54 V. J. I. L. __ (forthcoming 2014).
45 Draft Articles, *supra* note 34, at 129–34.
46 *Id*. at 65.
47 Commentaries to the draft articles on Responsibility of States for internationally wrongful acts (adopted by the International Law Commission at its 53rd Session) (2001), at 327 [hereinafter ILC Commentaries].

costly—ideally costly enough to cease its unlawful behavior. If the target can easily relocate its operations across national boundaries, as is often possible in the cyber-context, the countermeasure may not impose a significant cost on the actor responsible for the attack. For this reason, countermeasures are likely to be more effective against state actors and less effective against non-state actors. Finally, it can be difficult to design a countermeasure that targets only the actor that perpetuated the legally wrongful attack. In particular, a countermeasure that disables a computer or network may very well cause harm to those who have little or nothing to do with the unlawful attacks. This could have the perverse effect of making the state injured by the original attack a perpetrator of an unlawful attack against those who simply happen to share a network with the actor that generated the original attack or whose computer was being used as a pawn to carry out attacks without their knowledge or acquiescence. Together these challenges can lead a system that relies too heavily on active countermeasures from spinning out of control.

4. THE DRAWBACKS AND DANGERS OF DEVELOPING AN AGGRESSIVE COUNTERMEASURES REGIME

The rising chorus of voices in favor of an active countermeasures regime has thus far not taken full account of the potential drawbacks and dangers of such a regime. In this section, I outline both the legal concerns and policy concerns regarding active countermeasures. My hope is that this will give pause to those advocating an expansive countermeasure regime and encourage some careful thinking in the future about the appropriate limits on active countermeasures.

First, the legal constraints. Those who favor application of countermeasures as a means of addressing the gap between Article 2(4) and 51 often turn to the International Law Associations Draft Articles on State Responsibility as the source of authority on countermeasures. They point, in particular, to Article 49, which outlines the "object and limits" of countermeasures.[48] As described in the previous section, this Article establishes that an injured state may take countermeasures against a State that is responsible for an internationally wrongful act in order to induce the non-complying state to come into compliance.

But often overlooked in this discussion is the Article that follows immediately after Article 49. Article 50—"Obligations not affected by countermeasures"—outlines a series of constraints on countermeasures. Of particular importance to cyber is the first, which provides that "Countermeasures shall not affect . . . the obligations to refrain from the threat or use of force as embodied in the Charter of the United Nations."[49] Furthermore, Article 59 reaffirms that, "These articles are without prejudice to the Charter of the United Nations."[50]

The commentaries on the Draft Articles further reinforce that the Articles apply only to "non-forcible countermeasures."[51] It expressly notes that it "excludes forcible measures from the ambit of permissible countermeasures under chapter II."[52] Moreover, it notes:

48 Draft Articles, *supra* note 34, at art. 49.
49 Draft Articles, *supra* note 34, at art. 50 (a).
50 *Id*. at art. 59.
51 ILC Commentaries, *supra* note 47, at 327.
52 ILC Commentaries, *supra* note 47, at 334.

> The prohibition of forcible countermeasures is spelled out in the Declaration on Principles of International Law concerning Friendly Relations and Cooperation among States in accordance with the Charter of the United Nations, by which the General Assembly of the United Nations proclaimed that "States have a duty to refrain from acts of reprisal involving use of force." The prohibition is also consistent with prevailing doctrine as well as a number of authoritative pronouncements of international judicial and other bodies.[53]

The implications for active countermeasures against cyber-attacks should be obvious. If a cyber-attack constitutes a "use of force" in violation of Article 2(4)—and this is the source of their international wrongfulness—then an active countermeasure that utilizes similar technology to "hack back" is, presumably, also a "use of force." If that is the case, then the ILC Draft Articles and Commentaries would seem to prohibit such countermeasures—at least any countermeasures comparable to the act that prompted the response.

The *Tallinn Manual* experts and Mike Schmitt struggle admirably with these issues.[54] The *Tallinn Manual* experts were unable to decide even how to determine when a cyber-attack constituted an illegal use of force, much less what responses were permissible for those uses of force that fall in the gap between Article 2(4) and 51. Schmitt, writing separately, notes this lack of agreement. He identifies a minority view "that forceful countermeasures reaching the level of use of force are appropriate in response to an internationally wrongful act that constitutes a use of force, but remains below the armed attack threshold,"[55] pointing to a separate opinion by Judge Simma in the *Oil Platforms* case that some read to endorse forceful countermeasures.[56] Read in context, however, the opinion—which was, after all, the opinion of a single judge—does not stand for the proposition that forceful countermeasures are permitted. Instead, it simply makes the commonsense observation that "a State may of course defend itself" even against uses of force that do not amount to an armed attack, but such defense is subject to limits of "necessity, proportionality, and immediacy in a particular strict way."[57]

There is little legal support for the proposition that countermeasures doctrine provides a legal end-run around the prohibition on the use of force in Article 2(4) of the UN Charter. The leading authorities on countermeasures have affirmed that the UN Charter prohibitions are unaffected by the doctrine of lawful countermeasures. A state that counterstrikes or "hacks back" is therefore in violation of Article 2(4) of the UN Charter. It is true that the (now) victim state will not have the legal right to respond with force in self defense under Article 51, but the "hack back" (or "mitigative attack," as one article puts it[58]) is illegal nonetheless. Indeed, as a matter of international law, it is just as illegal as the attack that prompted it.

Is there is a class of cyber-attacks that do not amount to a "use of force" but constitute a violation of a customary norm of non-interference in a sovereign state that would give rise to a right to active cyber-defense? Again, the legal grounds for such a right to active cyber-defense are extremely weak. Those who hold that there is a right to non-interference distinct

53 ILC Commentaries, *supra* note 47, at 334.
54 Schmitt, *supra* note 44, at 16-19; Tallinn Manual, *supra* note 5, r. 48-52.
55 Schmitt, *supra* note 44, at 16.
56 Schmitt, supra note 44, at 16. Oil Platforms (Iran v. U.S.), 2003 I.C.J. 161 (Nov. 6), Separate Opinion of Judge Simma, ¶ 14.
57 *Oil Platforms*, Separate Opinion of Judge Simma ¶ 14.
58 Kesan & Hayes, *supra* note 44, at 469 ("Reflecting attacks back or initiating a new attack could, under the proper circumstances, both be considered mitigative counterattacks.").

from the prohibition on use of force often cite the *Nicaragua* case, where the International Court of Justice explained that the principle of state sovereignty "forbids all States or groups of States to intervene directly or indirectly in the internal or external affairs of other States."[59] A cyber attack could violate the right to non-interference, the argument goes, and therefore constitute internationally wrongful act that would trigger a right to respond with a non-forceful countermeasure (including a similar cyber attack). As yet, however, the norm of non-intervention likely remains too ill defined to support such a claim. It is far from clear that there is, indeed, a norm of non-intervention distinct from the prohibition on use of force in the UN Charter. Even were the norm better defined, cyber-attacks would be a poor fit. According to the *Nicaragua* case, the norm protects states from interference in "matters in which each State is permitted, by the principles of State sovereignty, to decide freely."[60] A cyber attack is generally not intended to "coerce" in this way.

There are important policy reasons for the legal limits on forceful countermeasures. There is a reason that the UN Charter does not permit states to respond with force to every single illegal use of force—in particular, to those uses of force that do not arise to the "most grave" level sufficient to amount to an "armed attack" and trigger Article 51. It is this: The gap between Article 2(4) and Article 51 prevents an endless process of retaliations for small offenses—a process that could, indeed is likely, to spin out of control over time. The gap between 2(4) and 51 puts some play in the joints, requiring states to absorb low-level uses of force without immediately responding in kind.

When considering the wisdom of continuing to observe this force gap, it is important to remember that cyber does not operate in isolation. If the legal principle were established that forceful countermeasures are permitted in cyber, there would be no reason not to apply those same principles outside the cyber context. If a state may respond to a use of force that does not rise to an armed attack with a use of force of its own in cyber, this could effectively eliminate the generally well-accepted gap between "use of force" under Article 2(4) and "armed attack" in Article 51. As a consequence, any use of force could provoke a forceful response. At stake, therefore, is not simply the capacity to respond to cyber-attacks, but the rules that govern the use of force in the international legal system more generally.

Likewise, there are good policy reasons to be wary of endorsing an expansive norm of non-interference that might give rise to a right to engage in active countermeasures. An expansive norm of non-interference could have far-reaching ramifications for other bodies of law. For example, if states have a right to demand non-interference by other states—and have a right to respond with countermeasures against those that do not observe this limit on interference—that might lead to countermeasures for a wide range of extraterritorial activities. Affected activities might include state funding for non-governmental organizations in other countries or extraterritorial application of commercial law (for example, anti-trust law and intellectual property law). It is important that lawyers and policymakers be careful not to create bigger problems in other areas of international law when trying to solve the threshold problem in cyber by engaging in over-interpretation of broadly applicable legal principles.

59 Nicaragua, ¶ 205. The Court continued: "A prohibited intervention must accordingly be one bearing on matters in which each State is permitted, by the principle of State sovereignty, to decide freely. One of these is the choice of a political, economic, social and cultural system, and the formulation of foreign policy. Intervention is wrongful when it uses methods of coercion in regard to such choices, which must remain free ones." Id.

60 Nicaragua, ¶ 205.

CONCLUSION: NON-FORCEFUL RESPONSES TO CYBER-ATTACKS AND A CALL FOR COLLABORATION

The argument made thus far may seem overly rigid and legalistic. Indeed, the prohibition on forceful countermeasures in cyber may appear absurd, effectively blessing illegal uses of force that stay just within the artificial line where a "use of force" crosses over into an "armed attack." But it is important to remember that even if force may not be used in response to an illegal use of force, states are not left powerless in the face of cyber-attacks. States that are subjected to an illegal use of force may respond with economic, diplomatic, or political sanctions—including asset freezes, trade sanctions, withdrawal of cooperation, travel bans, and banking restrictions—none of which are subject to limits under the UN Charter.[61] Customary countermeasures are limited to the suspension of international obligations, must be proportional, generally are "in kind"—involving like action for like action—and cannot be taken by third parties. Economic, diplomatic, and political sanctions are not subject to these same constraints (though they may be subject to independent legal constraints). As a result, sanctions can offer a wider range of options for responding to an unlawful action by a state—particularly an unlawful use of force—than do countermeasures.

States may also respond more directly with non-forceful cyber-measures. These might include some activities that have at times been classified as "active responses" to cyber-attacks—internal notification (notifying users, administrators, and management of the attacked entity), internal response (taking action to defend the system such as blocking certain IP addresses, creating an air gap), and external cooperative responses (including coordinated law enforcement and upstream support to internet service providers).[62] It may also include elements of non-cooperative information gathering and even traceback.

Collaboration between technical experts and international lawyers could be especially fruitful in drawing the line between cyber-responses that constitute uses of force and those that do not. Projecting satellite signals and sound waves into the sovereign space of another country do not constitute "uses of force." Nor does gathering satellite imagery—even very detailed imagery—or reporting activities of international news media, even state-run or state-funded news media, such as the BBC. Some of the more intrusive forms of intelligence gathering are also not restricted by international law, though the precise bounds of the international legal limits on such activities is a point of some contention.[63] The question that technical experts, collaborating with lawyers, could answer is what defensive cyber-measures are functionally similar to these well-accepted activities and which step over the line into use of force.

61 For more on what I call "outcasting," see Oona A. Hathaway & Scott J. Shapiro, Outcasting: Enforcement in Domestic and International Law, 121 Yale L. J. 252 (2011).

62 *See* NRC REPORT, *supra* note 44, at 148-49.

63 Compare 1 Oppenheim, International Law 862 (H. Lauterpacht ed , 8th ed. 1955) (asserting that peacetime intelligence gathering "is not considered wrong morally, politically or legally . . . ."), *and* Geoffrey B. Demarest, Espionage in International Law, 24 DENV. J. INT'L L. & POL'Y 321 (1996) (concluding that "peacetime espionage has always been seen as an issue of domestic law," and therefore not governed by international law), *with* Quincy Wright, Espionage and the Doctrine of Non-Intervention in Internal Affairs, in Essays on Espionage and International Law 3, 12 (Roland J. Stranger ed., 1962) (raising concerns that intelligence gathering may transgress the territorial integrity and political independence of a country, in violation of the UN Charter). It is clear that states may punish captured spies. They do not receive prisoner of war status or any of the immunities due to combatants in an armed conflict.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Artificial (Intelligent) Agents and Active Cyber Defence: Policy Implications

Caitríona H. Heinl
Research Fellow
Centre of Excellence for
National Security (CENS)
S. Rajaratnam School of International Studies
Singapore

Abstract: This article examines the implications of employing artificial (intelligent) agents for active cyber defence (ACD) measures, in other words proactive measures, in the context of military and private sector operations. The article finds that many complex cyber-related challenges are solved by applying artificial intelligence (AI) tools, particularly since intelligent malware and new advanced cyber capabilities are evolving at a fast rate and intelligent solutions can assist in automation where pre-fixed automation designs are insufficient. Intelligent agents potentially underpin solutions for many current and future cyber-related challenges and AI therefore plays a possible role as one of a number of significant technical tools for ACD. However, this article considers that although such advanced solutions are needed, it finds that many technical and policy-related questions still surround the possible future consequences of these solutions, in particular the employing of fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. While these AI tools and ACD actions might be technologically possible, the article argues that a number of significant policy gaps arise such as legal question marks, ideological and ethical concerns, public perception issues, public-private sector ramifications, and economic matters. It highlights several areas of possible concern and concludes that it is important to examine further the implications of these rapidly evolving developments. Finally, the article provides several policy options as a start so as to begin responsibly shaping the future policy landscape in this field.

Keywords: *artificial intelligence, artificial (intelligent) agents, active cyber defence, autonomy*

1. INTRODUCTION

Given that current cyber defence measures, in particular passive cyber defences, are inadequate for increasingly sophisticated threats, many argue for proactive measures to be taken. This

article therefore examines the implications of employing artificial (intelligent) agents for active cyber defence (ACD) measures in the context of military and private sector operations.

The article finds that many cyber-related challenges are solved by applying artificial intelligence (AI) tools, particularly since intelligent malware and new advanced cyber capabilities are evolving at a rapid rate. Employing AI techniques and intelligent solutions for the purposes of dealing effectively with complex cyber-related threats is then best explained by the ability of these technologies to assist in automation since pre-fixed automation designs are insufficient. Intelligent agents potentially underlie solutions for many current and future cyber-related challenges and AI therefore plays a possible position as one of a number of significant technical tools for ACD.

However, this article argues that although such advanced solutions are required, many technical questions and uncertainties still surround the possible future consequences of their use, most particularly for the employing of fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. Therefore, while numerous AI applications are already in use for cyber-related issues, this article suggests that the potential policy implications of a number of emerging and proposed techniques including possible disruptive technologies now require serious consideration. Although these AI tools and ACD actions might be technologically possible, the article considers that there are a number of serious legal implications, ideological and ethical concerns, public perception issues, public-private sector ramifications, and economic matters that could arise. It finds that to date, insufficient widespread attention has been paid in the public policy domain to many of these gaps in policy. The article concludes that there is a significant time-sensitive need to commence an in-depth further examination and serious public discourse on these issues in order to develop the future policy landscape, and finally, it provides several possible policy options that could be considered.

The article is organised as follows:

- Section 2 explores the core background concepts of artificial intelligence.
- Section 3 outlines cyber-related challenges for which AI solutions could be effectively employed.
- Section 4 considers active cyber defence and the possible roles of AI.
- Section 5 examines potentially successful emerging AI technologies.
- The final section discusses several possible policy implications based on the findings of this article and provides a number of policy recommendations.

2. BACKGROUND: CORE AI CONCEPTS

AI or computational intelligence is generally defined as technology and a branch of computer science that develops intelligent machines and software. It is regarded as the study of the design of intelligent agents where an intelligent agent is a system that perceives its environment and takes actions to maximise its chances of success. Intelligent agents are software components with features of intelligent behaviour such as (at a minimum) pro-activeness, the ability to communicate, and reactivity (in other words the ability to make some decisions and to act).[1]

1 Enn Tyugu, "Command and Control of Cyber Weapons", *4th International Conference on Cyber Conflict*, Tallinn, 2012.

Additionally, AI may be described as the automation of activities such as decision-making, problem solving, learning, and the study of the computations that make it possible to perceive, reason, and act. It can assist planning, learning, natural language processing, robotics, computer vision, speech recognition, and problem solving that requires large amounts of memory and processing time. And while AI may be considered as a science for developing methods to solve complex problems that require some intelligence such as making the right decisions based on large amounts of data, it may also be viewed as a science that aims to discover the essence of intelligence and develop generally intelligent machines.[2] General intelligence is predicted by some to come into being by 2050, possibly leading to singularity, in other words the technological creation of intelligence superior to human intelligence. Approaches for improving machine intelligence are progressing in areas such as the expression of emotion, language interaction, as well as face recognition and forecasts suggest that they will be "interim substitutes" before direct machine intelligence is realised but for now a further maturation of AI techniques and technologies is required.[3]

Several examples of AI in use include Deep Blue (IBM's chess playing computer), autonomous vehicles that drive with traffic in urban environments[4], IBM's Watson (the computer system that can answer natural language questions), and the X-47 robotic aircraft which recently landed autonomously.[5] In addition, although not readily apparent to those working outside the field, many AI technologies such as data mining or search methods are part of everyday use. This phenomenon, where a technique is not considered as AI by the time it is used by the general public, is described as the "AI effect". It is a particularly significant concept in that public perception of what constitutes AI as well as acceptance of these tools, especially the more advanced future tools, could play an important role in the shaping of future policies. Some well known examples of the AI effect include Apple's Siri application which uses a natural language user interface to answer questions and make recommendations, Google's new Hummingbird algorithm which makes meaning of the search query for more relevant "intuitive" search results, and Google's self-driving cars.

Employing AI technologies and techniques for the purposes of cybersecurity, cyber defence (or cyber offence) and ACD is currently best explained by the ability to assist in automation. Many contend that automation is essential for dealing effectively with cyber-related threats and that many cyber defence problems can only be solved by applying AI methods. Intelligent malware and new advanced cyber capabilities are evolving rapidly, and experts argue that AI can provide the requisite flexibility and learning capability to software.[6] Intelligent software is therefore being increasingly used in cyber operations and some argue that cyber defence systems could be further adaptive and evolve dynamically with changes in network conditions

2 Enn Tyugu, "Artificial Intelligence in Cyber Defense", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.

3 Development, Concepts and Doctrine Centre (DCDC), UK Ministry of Defence, *Strategic Trends Programme Global Strategic Trends – Out to 2040*, 4th ed., January 2010.

4 Defense Advanced Research Projects Agency (DARPA), United States, "DARPA Urban Challenge", http://archive.darpa.mil/grandchallenge/, November 2007.

5 Alessandro Guarino, "Autonomous cyber weapons no longer science-fiction", Engineering and Technology Magazine, Vol 8 Issue 8, http://eandt.theiet.org/magazine/2013/08/intelligent-weapons-are-coming.cfm, 12 August 2013.

6 Tyugu, Artificial Intelligence in Cyber Defense.

by implementing dynamic behaviour, autonomy, and adaptation such as autonomic computing or multi-agent systems.[7]

3. CYBER-RELATED CHALLENGES: AI SOLUTIONS

Although many AI methods are currently available for cyber defence, there is still an identified need for further advanced solutions, intelligent decision support, automated knowledge management and rapid situation assessment[8] for the more complex cyber-related problems. In short, reports state that intelligent systems and networks, even self-repairing networks, could increase resilience in the longer term.[9] Pre-fixed automation designs are not sufficiently effective against evolving cyber incidents for instance. New vulnerabilities, exploits and outages can occur simultaneously and at any point in time,[10] and experts contend that it is difficult for humans to effectively handle the sheer volumes of data and speed of processes without high degrees of automation - very fast, if not automated, reaction to situations, comprehensive situation awareness, and a handling of large amounts of information at a rapid rate to analyse events and make decisions is therefore considered necessary.[11]

A recent United States Department of Defense report[12] explains that the identification of operationally introduced vulnerabilities in complex systems is extremely difficult technically, and "[i]n a perfect world, DoD operational systems would be able to tell a commander when and if they were compromised, whether the system is still usable in full or degraded mode, identify alternatives to aid the commander in completing the mission, and finally provide the ability to restore the system to a known, trusted state. Today's technology does not allow that level of fidelity and understanding of systems." The report then outlines the need for the development of capacity to conduct "many, potentially hundreds or more, simultaneous, synchronized offensive cyber operations while defending against a like number of cyber attacks". For now however, it describes system administrators as inadequately trained and overworked, a lack of comprehensive automation capabilities to free personnel for serious problems, and an inadequate visibility into situational awareness of systems and networks. In addition, systems such as automated intrusion detection, automated patch management, status data from each network, and regular network audits are currently unavailable.

7 Igor Kotenko, "Agent-based modelling and simulation of network cyber-attacks and cooperative defence mechanisms", St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, available at: http://cdn.intechopen.com/pdfs/11547/InTech-Agent_based_modeling_and_simulation_of_network_infrastructure_cyber_attacks_and_cooperative_defense_mechanisms.pdf, 2010.

8 Tyugu, Artificial Intelligence in Cyber Defense.

9 DCDC, *Global Strategic Trends*.

10 Beaudoin, Japkowicz & Matwin, "Autonomic Computer Network Defence Using Risk State and Reinforcement Learning", Defense Research and Development Canada, 2012.

11 Tyugu, Artificial Intelligence in Cyber Defense.

12 Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, *Resilient Military Systems and the Advanced Cyber Threat*, United States Department of Defense, Defense Science Board, January 2013.

Intelligent agents and AI-enhanced tools potentially play a significant role by underpinning solutions for several, if not most, of these problems as well as the following cyber-related challenges:[13]

- The need for continual collection, comprehensive understanding, analysis and management of large amounts of dynamic data, in other words knowledge management, from a plethora of sources and devices to develop actionable intelligence.
- Insufficient pattern recognition and behavioural analysis across different data streams from many channels.
- Lack of visibility of the complete conditions of the IT environment, and insights into possible threats and systems compromise in real time.
- Non-identification of unusual behaviour, systems and network traffic, in other words anomalies, and unusual user behaviour to spot insider threats and internal misuses.
- The need for comprehensive knowledge of the threats for decision support and decision-making.
- Intrusion detection.
- Situational awareness and continual monitoring so as to detect and mitigate attacks.
- Harnessing of information to prevent, detect and even "predict" (or rather foresee) attacks.
- Insufficient passive defences and resilience of systems to attacks.

Lastly, one of the core challenges facing nations and corporations today includes the difficulties in identifying, training and retaining skilled individuals and general consensus currently holds that the numbers working in this area need to markedly increase. However, recent defence reports from the U.S. now identify that there is a "burnout factor beginning to exhibit itself"[14] among the current cyber workforce. Therefore, although increasing the number of "cyber warriors" might alleviate the current cybersecurity skills gap to a certain degree, AI and advanced automation of particular tasks could be highly beneficial over the longer term. Furthermore, strains on labour and financial resources might be alleviated. This issue therefore requires serious consideration and further concrete analysis, especially in light of future expected trends in demographics, which according to some defence officials will work against several countries.[15]

4. ACTIVE CYBER DEFENCE AND INTELLIGENT AGENTS

But this virtual version of vigilante justice is fraught with peril….[16]

13 General information from: Security for Business Innovation Council, "Getting Ahead of Advanced Threats: Achieving Intelligence-Driven Information Security", RSA Sponsored Report, 2012; and Mirko Zorz, "Complex security architectures and innovation", http://www.net-security.org/article.php?id=1692&p=1, 29 March 2012.

14 Under Secretary of Defense, *Resilient Military Systems*.

15 William Lynn III, former United States Under Secretary of Defense, "2010 Cyberspace Symposium: Keynote – DoD Perspective", 26 May 2010.

16 Gregory Zeller, "Cyber warriors eye move to 'active defense'", Long Island Business News, 25 February 2013.

Current defence measures are not considered as prepared for the limitless ways to attack a network,[17] and many argue that passive defence alone may not be sufficient.[18] Arguments are therefore being made for policy makers and network defenders to incorporate lessons such as "the best defence includes an offence", in other words active cyber defence. William Lynn III, former United States Under Secretary of Defense, argues for instance[19] that in cyber, offence is dominant and "we cannot retreat behind a Maginot Line of firewalls" - defences should therefore be dynamic and responses at network speed as attacks happen or before they arrive. Corporations and government bodies are beginning to use ACD techniques more frequently, and this section therefore explores those aspects of ACD where AI could play a role as one of a number of technical tools in the ACD toolbox.

Although there is no universal definition for the term, for the purposes of this article ACD is understood to entail proactive measures that are launched to defend against malicious cyber activities. According to a recent CNAS analysis[20] on ACD options available to the private sector, one of the few formal definitions is found within the United States 2011 Department of Defense Strategy for Operations in Cyberspace: "DoD's synchronized real-time capability to discover, detect, analyze, and mitigate threats and vulnerabilities. It builds on traditional approaches to defending DoD networks and systems, supplementing best practices with new operating concepts. It operates at network speed by using sensors, software, and intelligence to detect and stop malicious activity before it can affect DoD networks and systems. As intrusions may not always be stopped at the network boundary, DoD will continue to operate and improve its advanced sensors to detect, discover, and mitigate malicious activity on DoD networks."

The CNAS analysis lays out a framework (adapted in Figure 1 below) to show that it is at the Delivery phase, during the Cyber Engagement Zone, that employing ACD techniques becomes most significant, in other words when the defender can take the initiative. However, organisations are often unaware of a compromise until the Command and Control (C2) phase when installed malware communicates outside the organisation under attack. Under this analysis, three ACD concepts are identified for responding to an attack: detection and forensics, deception, and attack termination. For detection, a number of ACD techniques to detect attacks that circumvent passive defences may be used, and once information is gathered it can inform the company's response decisions. Detection can be by way of local information gathering using ACD techniques within the organisation's networks, or by what is known as remote information gathering where an organisation may gather information about an incident outside its own networks (by for example accessing the C2 server of another body and scanning the computer, by loading software, removing or deleting data, or stopping the computer from functioning). For attack termination, ACD techniques can stop an attack while it is occurring by, for instance, preventing information from leaving the network or by stopping the connection between the infected computer and the C2 server. More aggressive actions could include "patching computers outside the company's network that are used to launch attacks, taking

17 David T. Fahrenkrug, Office of the United States Secretary of Defense, "Countering the Offensive Advantage in Cyberspace: An Integrated Defensive Strategy", *4th International Conference on Cyber Conflict*, Tallinn, 2012.

18 Porche, Sollinger & McKay, "An Enemy Without Borders", U.S. Naval Institute Proceedings, October 2012.

19 Lynn, 2010 Cyberspace Symposium.

20 Irving Lachow, "Active Cyber Defense: A Framework for Policymakers", Center for a New American Security, February 2013.

control of remote computers to stop attacks, and launching denial of service of attacks against attacking machines."

While ACD actions such as deploying honeypots, actively tracking adversaries' movements, using deception techniques, watermarking documents and terminating connections from the C2 node to infected computers do not seem to be illegal, the CNAS study concludes that there is an absence of clear national and international law for some actions, particularly remote information gathering and some of the more aggressive actions. In effect, ACD options that involve retaliation or "hacking back" are generally considered illegal (whether the ACD response is before, during or after an incident) since attempts are made to access the systems of another organisation without permission so as to access or alter information on the C2 server or computers. The study further finds that it is unclear whether accessing the C2 server of another organisation could violate privacy laws and expose a company to civil actions as well as criminal prosecution. In addition, if an organisation is in another jurisdiction, a company could possibly violate that country's national laws, even if not violating its own. It is also unclear whether a company could legally patch the C2 server of another organisation since it would entail altering or deleting information on its computers. Finally, when the C2 server is not directly connected to the adversary but "several hops away", not only is it technically challenging to find the source of the attacks but the company tracing the sources could violate its own national laws, those of multiple other jurisdictions, and international laws such as the Budapest Convention on Cybercrime.

FIGURE 1: CYBER KILL-CHAIN (ADAPTED FROM LACHOW, "ACTIVE CYBER DEFENSE: A FRAMEWORK FOR POLICYMAKERS", CNAS, 2013)

Phase	Description
Reconnoiter/ Reconnaissance	Adversary researches, identifies and selects its targets.
Weaponise	Adversary couples malware with a delivery mechanism, often using an automated tool.
-	Cyber Engagement Zone:
Deliver	Adversary transmits weaponised payload to the target through emails or websites for example.
Exploit	Malware delivered to the target is triggered when the user takes an ac ion such as opening email attachments or visiting an infected site.
Install	The malware infects the user's system. It may hide itself from malware detection software on that system.
Command and Control (C2)	The malware sends an update on its location and status to a C2 server, often through encrypted channels that are hard to detect.
Act	The malware takes actions for the adversary such as exfiltrating, altering or destroying data.

This framework is a helpful tool to clarify when AI techniques might play a significant role. For instance, the time between an attack and systems compromise can often take minutes yet it could take months to discover the breach.[21] AI techniques could therefore be of particular value in these earlier phases of the Cyber Engagement Zone. They can assist earlier detection of compromise and provide situational awareness. In particular since active defence demands high levels of situational awareness to respond to the threat of intrusion.[22] They can also assist information gathering and decision support. Deception techniques such as proposals for experimental frameworks for autonomous baiting and deception[23] of adversaries could also be useful.

However, although these ACD concepts are technologically possible, there is legal uncertainty and it is therefore unclear whether AI tools could (or should) be used as possible ACD techniques. Before employing these tools for ACD actions, legal certainty should therefore be sought so that existing laws are not violated, even where it might be argued that the law is "grey" or national and international law is unclear.

5. CYBER GAME CHANGERS: EMERGING EFFECTIVE INTELLIGENT AGENTS & AI COMBINED WITH OTHER DISCIPLINES

While numerous AI applications such as neural networks, expert systems, intelligent agents, search, learning, and constraint solving are in use for several cyber-related challenges, a number of emerging and proposed intelligent agent hybrid technologies and techniques require further research and consideration (for example, agent-based distributed intrusion detection and hybrid multi-agent/neural network based intrusion detection). Most particularly, the policy ramifications of possible future tools that combine AI technologies with other disciplines should be seriously analysed since these tools could prove to be disruptive technologies and cyber game changers if successfully developed in the medium to long term. Further research should therefore be conducted in the near term on the consequences of their possible development.

A recent analysis of the future strategic context for defence to 2040 by the Development, Concepts and Doctrine Centre (DCDC) of the UK Ministry of Defence[24] states that advances in robotics, cognitive science coupled with powerful computing, sensors, energy efficiency and nano-technology will combine to produce rapid improvements in the capabilities of combat systems. The report explains that advances in nanotechnology will underpin many breakthroughs and that developments in individual areas are likely to be evolutionary. However, developments may be revolutionary where disciplines interact, such as the combination of cognitive science and ICT, to produce advanced decision-support tools. Furthermore, according to this report, research on mapping or "reverse engineering" the human brain will likely lead to development of "neural models" and this combined with other systems such as sensors may provide human like qualities for machine intelligence. The simulation of cognitive processes using AI is likely

21 Costin Raiu, Kaspersky Labs, "Cyber Terrorism – An Industry Outlook", Cyber Security Forum Asia, 03 December 2012.
22 Fahrenkrug, Countering the Offensive Advantage.
23 Bilar & Saltaformaggio, "Using a Novel Behavioural Stimuli-Response Framework to Defend against Adversarial Cyberspace Participants", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.
24 DCDC, *Global Strategic Trends*.

to be focused in the short term on probability and pattern recognition and in the longer term to aid knowledge management and support decision-making.

In light of several conclusions within the DCDC report,[25] and for the purposes of this article, the possible future consequences of the following disciplines and technologies should be seriously considered from a policy perspective:

- *Quantum Computing:* Processing capabilities could possibly increase by 100 billion times.
- *Simulation:* Advances in mathematical modelling, behavioural science and social science will seemingly combine for more informed decision-making while advances in processing techniques and computational power will allow more comprehensive modelling and potentially enable better pattern recognition.
- *Virtual Databases:* Development of the semantic web and associated technologies will create an integrated data store with unprecedented level of access that could be exploited by reasoning techniques for more sophisticated analysis that may expose previously unseen patterns with potentially unforeseeable consequences. Sophisticated data mining tools will include automatic data reduction/filtering and automated algorithmic analysis for faster access to relevant information. "Virtual Knowledge Bases" will apparently store knowledge within large database structures in formats that intelligent software could use for improved searching, to answer questions across the whole knowledge store in near natural language form, and to issue automated situation reports on demand or in response to events to assist situational awareness.
- *Cognitive and Behavioural Science:* Certain advances such as neuro-imaging technologies may make mapping of brain activity with behaviour more reliable. Modelling techniques are likely to become more powerful and capable of more accurately understanding the complexity of human behaviour and performance which could lead to an ability to "map the human terrain".

  Advancing the field of brain sciences could open opportunities for new means to develop AI and studies are being conducted to understand the brain and how human brain function could be used as a framework for improving technologies such as cybersecurity and mobile security technologies - for example, cognitive security technology modelled after human brain function for the next generation of technology security.[26] Further, a reported new trend is the application of AI and cognitive methods in situation awareness which permits fusion of human and computer situation awareness, and supports real time and automatic decision-making.[27]

  However, commentators also contend that AI is not yet, and may never be, as powerful as "intelligence amplification", in other words when human cognition is augmented by close interaction with computers.[28] For example, after Deep Blue beat Kasparov, he tested what would happen if a machine and human chess player were paired in collaboration and found that human-machine teams, even when they did not

25 DCDC, *Global Strategic Trends*.

26 Center for Systems Security and Information Assurance, Cyber Defense and Disaster Recovery Conference 2013: Mobile Security.

27 Tyugu, Command and Control of Cyber Weapons.

28 Walter Isaacson, "Brain gain?", Book Review of Smarter Than You Think by Clive Thompson, International New York Times, 2-3 November 2013.

include the best grandmasters or most powerful computers, consistently beat teams composed solely of human grandmasters or computers.[29]

- *Autonomous Systems and Robotics:* Growth in the role of unmanned, autonomous and intelligent systems is expected. These systems could range from small sensors and personalised robots replicating human behaviour and appearance to a "cooperative plethora of intelligent networks or swarms of environmental-based platforms with the power to act without human authorisation and direction"[30] with a range of autonomy from fully autonomous to significantly automated and self-coordinating while still under high-level human command.

  Although software with intelligent agent characteristics is already in use, both technical and policy-oriented research should be further conducted on the possible consequences of employing fully autonomous intelligent agents. Autonomous intelligent agents are defined as "systems situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future - the agent is strictly associated with its environment, in other words it can be useless outside the environment for which it was designed or not an agent at all".[31]

  According to Guarino,[32] they can be purely software operating in cyberspace (computational agents) or integrated into a physical system (robotic agents) where they underpin the robot's behaviour and capabilities. Computational autonomous agents could be used for intelligence-gathering and military operations, in particular during the Reconnaissance phase for automatic discovery of vulnerabilities in target systems for example or for gathering intelligence. Autonomous agents could then develop ways to exploit these vulnerabilities and they will not need fixed and pre-programmed methods to penetrate the target system since they will analyse the target, autonomously select the points of vulnerability, and develop means to use these points so as to infiltrate the system. Currently however these capabilities are manually developed or bought on the open market since full automation of exploit development is still not widely available. Guarino continues, that although an agent's goals and targets could be pre-programmed and precisely stated to facilitate its task and to ensure legality, it could in fact occur that sometimes it might be deemed preferable to give the agent "free rein".

  The Command and Control (C2) phase therefore presents significant difficulties and warrants further attention, particularly since command and control could be hard to achieve. Experts warn that the more intelligent software becomes, the more difficult it could be to control and the C2 phase causes new threats that are difficult to avoid due to the complexity of the agents' behaviour, in particular its misunderstanding a situation, misinterpretation of commands, loss of contact and formation of unwanted coalitions, unintentionally behaving in a harmful way or its unexpected actions and unpredictable behaviour.[33]

29 Isaacson, Brain gain?
30 DCDC, *Global Strategic Trends*.
31 Alessandro Guarino, "Autonomous Intelligent Agents in Cyber Offence", *5th International Conference on Cyber Conflict*, Tallinn, 2013.
32 Guarino, Autonomous Intelligent Agents.
33 Tyugu, Command and Control of Cyber Weapons.

6. UNCERTAIN POLICY RAMIFICATIONS

To Every Man is Given the Key to the Gates of Heaven.
The Same Key Opens the Gates of Hell.[34]

These possible developments raise significant unanswered questions and concerns. At this juncture however, technical and policy-oriented solutions, at least those in the public domain, are sparse. Concrete efforts to further clarify these gaps should therefore be conducted as soon as possible, with particular focus on ideological and ethical concerns, public perception, the interplay between the public and private sectors, economic matters, and legal implications that could arise. It is pertinent that further analysis be conducted without delay so as to develop and implement, where possible, both policy-based solutions and technological safeguards from the outset.

Suffice to say that the "Internet of the Future" will not look like the Internet of today and further challenges will also include the Internet of Things and unanticipated new usages.[35] Like previous inventions, strategic reports foresee that many of these technological developments could have positive consequences, including unintended, but some could also present threats or have "catastrophic effects".[36] In particular, these reports outline[37] that reliance on AI could create new vulnerabilities that could be exploited by adversaries and there is a high chance that malicious states and non-state actors could acquire such capabilities. Further attention should therefore focus on how this threat could be thwarted and what possible technological or policy-oriented solutions could be found to mitigate malicious applications of these future tools.

Advanced intelligent systems could also challenge the interaction between automated and human components, and the complexity of controlling multiple autonomous systems and interpreting information could become extremely difficult. Forecasts suggest that those unable for these challenges may be replaced by intelligent machines or "upgraded" by technology augmentation. Autonomic defences might even be developed to take over when human judgement is deemed "too affected by emotions or information overload".[38]

A number of technical recommendations[39] so far suggested include ensuring in the design and development of new intelligent "cyber weapons" that 1) there is a guarantee of appropriate control over them under any circumstances; 2) strict constraints on their behaviour are set; 3) they are carefully tested (although thorough verification of their safety and possible behaviours is apparently difficult); and 4) the environment is restricted as much as possible by only permitting the agent to operate on known platforms. Questions such as to what extent an agent could communicate with its "base", and whether communication should be one-way (intelligence gathering from the agent for instance) or two-way in that the C2 structure could

34 Richard P. Feynman, "The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman, 1999.
35 Golling & Stelte, "Requirements for a Future EWS – Cyber Defence in the Internet of the Future", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.
36 DCDC, *Global Strategic Trends*.
37 DCDC, *Global Strategic Trends*.
38 Bilar & Saltaformaggio, Novel Behavioural Stimuli-Response.
39 Tyugu, Command and Control of Cyber Weapons.

issue instructions including target selection or self-destruct commands[40] should also be further examined. Particular attention should also be drawn to dealing with the possible cooperative behaviour of agents, in other words what is described as the "multi-agent" threat.

Tyugu[41] explains that since agents can be used most efficiently in multi-agent formations, it is expected that this will be the main form of agent application in cyber operations. They could for instance negotiate between themselves and cooperatively create a complex behaviour for achieving the general goals stated by a commander but this apparently means that the strict control of behaviour of each single agent will be weaker and it will be impossible to verify the outcome of multi-agent behaviour for all situations. He explains that unwanted coalitions could possibly occur if agents have too much autonomy in decision-making since communication between agents will only be partially visible to human controllers (Guarino argues that this could be extremely difficult to disable[42]). Technical solutions recommended for these problems so far include building safeguards such as backdoors and forced destruction into agents or self-destruction if loss of contact occurs.

Further clarity and certainty on these questions should however be sought as well as on the possible legal implications where recent analyses conclude that there is a certain amount of uncertainty. Under Guarino's analysis,[43] autonomous agents are similar to any other tool or cyber weapon employed and therefore fall under existing international law but it is unclear whether a creating state could always be held responsible if an agent exceeds its assigned tasks and makes an autonomous decision. For instance, for attribution purposes, the creators might not have known in advance the precise technique employed or the precise system targeted. Guarino therefore recommends the identification of autonomous agents, perhaps through mandatory signatures or watermarks embedded in their code, and the possible revising of international law. Lastly, if a fully autonomous agent is used as a weapon in self-defence, he also recommends that care be taken in the C2 function to clearly state the agent's targets and build in safeguards.

However, although technical safeguards such as mandatory signatures or watermarks are important recommendations, enforcing their use could prove difficult to achieve, especially in light of concerns over malicious non-state or state actors unwilling to comply with technical safeguards. Computer experts also argue that there seems to be a high risk, "too high a risk", of misfire or targeting of an innocent party due to misattribution if defensive measures are deployed with automated retaliation capability.[44] 44 Countries have now expressed concern over the challenges posed by fully autonomous lethal weapons since the May 2013 Human Rights Council.[45] A decision was also adopted in November 2013 by states party to the Convention on Conventional Weapons (CCW) to hold inaugural international discussions in May 2014 on how to address some of these challenges, including assurance of meaningful human control over targeting decisions and the use of violent force. The Campaign to Stop Killer Robots,[46] a new global campaign comprising 45 non-governmental organisations in 22

40 Guarino, Autonomous Intelligent Agents.
41 Tyugu, Command and Control of Cyber Weapons.
42 Guarino, Autonomous Intelligent Agents.
43 Guarino, Autonomous Intelligent Agents.
44 Dmitri Alperovitch, "Towards Establishment of Cyberspace Deterrence Strategy", *3rd International Conference on Cyber Conflict*, Tallinn, 2011.
45 Campaign to Stop Killer Robots, http://www.stopkillerrobots.org/2013/11/ccwmandate/.
46 Stuart Hughes, "Campaigners call for international ban on 'killer robots'", http://www.bbc.co.uk/news/uk-22250664, 23 April 2013.

countries, also recommends that states develop national policies and that negotiations should begin on a treaty to ban these weapons.

Though developing national policies is a good starting point, and while national legislation and international treaties are important, the regulating of such future developments could be difficult. An outright ban could be close to impossible to enforce while pursuing agreement by way of an international treaty could also raise its own particular difficulties. Further, not only can regulations be untimely in the context of rapid technological development but the controlling of these technological developments could be difficult, even where controls are put in place. It is safe to conclude that if a tool can be developed, it is more than likely that it will be developed. Cyber capabilities in particular are inherently difficult to prevent from being created and such regulatory solutions might not deter malicious actors. In addition, non-state actors will not necessarily feel morally or legally bound in the same way and state actors may not always play by the same "version of the rules".[47] A combination of technical and legal safeguards is required but further research is still needed to examine whether more could be done, while also ensuring that innovation is not suppressed disproportionately.

Public perception and acceptance of these technologies also requires further active attention as soon as possible since it could significantly impact the future uses of these technologies (although this might not be the case in every country). For instance, the public's understanding of AI and autonomous systems could fuel misconceptions about sci-fi doomsday scenarios. Alternatively, reports consider that concern over casualties could make these systems seem more attractive,[48] even if cyberwarfare could also lead to violent and destructive consequences.[49] Recently for example, the Campaign to Stop Killer Robots was created so as to demand a pre-emptive ban on the development, production and use of weapons capable of attacking targets without human intervention, in other words fully autonomous "human-out-of-the-loop systems". And in light of the recent privacy and security scandals, a number of advanced technologies developed by the public sector have already begun to be shelved in some countries over policy-related concerns.

To some extent, the public debate has already begun to kick off with a number of TED (Technology, Entertainment, Design) talks and sensational reporting. However, further widespread public discourse should be held and the public should be responsibly informed as soon as possible so that decisions may be made on many of these issues in an educated manner. Such proactive initiatives might go some way to ensure misperceptions are actively prevented before misunderstandings and possible negative perceptions become the norm. As the Director of DARPA (Defense Advanced Research Projects Agency) in the United States recently stated, these cutting-edge technologies will continue to be pushed and developed at an increasingly fast pace and society needs to begin making some important decisions about these questions.[50]

Where the public sector might be restrained from using some tools, it is still probable that they will eventually make their way into the commercial sector, if not already developed by the

47 Under Secretary of Defense, *Resilient Military Systems*.
48 DCDC, *Global Strategic Trends*.
49 Mariarosaria Taddeo, "An Analysis For a Just Cyber Warfare", *4th International Conference on Cyber Conflict*, Tallinn, 2012.
50 American Forces Press Services, "Director: DARPA Focuses on Technology for National Security", 15 November 2013.

private sector itself. It is therefore unclear whether the public or private sector will drive these technological developments in future. Defence reports suggest that financial constraints and reduced military budgets might further impede the public sector for instance, with particular financial strain from large weapons programmes,[51] in which case the perceived cost efficient aspects of these future technologies could make them more appealing. Further, the public sector does not always, and may not in future, match the speed of innovation in IT in the private sector. Defence officials explain that defence departments might have unique IT needs for example[52] and traditional ways of acquiring technologies which in some cases take many years. In the U.S. for instance this has traditionally taken close to seven years as compared to the development of the iphone which took two years. Lastly, while commercial off-the-shelf products could allow cost savings, security and supply problems might arise that endanger the security and availability of systems.[53]

For now, comprehensive guidelines that examine these concerns and policy gaps could greatly assist policy-makers by providing an informative and independent high-level analysis. A concrete examination of all the various scenarios that could possibly arise should be produced so that plans and strategies can be formulated now to prepare for all future expected as well as far-fetched outcomes. Care should also be taken to ensure that the policy formation process is informed by a deep technical understanding of how these technologies function, and that the public are engaged as much as possible as significant stakeholders. Currently, there is a wide gap that needs to be narrowed between the levels of understanding of those working in this field vis-à-vis policy-makers and the general public.

7. CONCLUSION

In summary, employing AI techniques and intelligent solutions for current as well as future cyber-related challenges, and in particular for active cyber defence, raises a number of significant technical questions and policy-related concerns. While advanced solutions are considered necessary, there is still much technical and policy-related uncertainty surrounding the future consequences of these tools, especially fully autonomous intelligent agents and possible disruptive technologies that combine AI with other disciplines. Several policy implications are highlighted that could perhaps arise such as legal uncertainty, ideological and ethical concerns, public perception problems, public-private sector ramifications, and economic issues. These policy gaps require even further examination and forward-looking solutions should be developed presently in order to anticipate difficulties that might arise in light of expected rapid developments in this field.

51 DCDC, *Global Strategic Trends*.
52 Lynn, 2010 Cyberspace Symposium.
53 Koch & Rodosek, "The Role of COTS Products for High Security Systems", *4th International Conference on Cyber Conflict*, Tallinn, 2012.

Chapter 2
Models of
Active Cyber Defence

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Malware is Called Malicious for a Reason: The Risks of Weaponizing Code

Stephen Cobb
Research Department
ESET North America
San Diego, USA
stephen.cobb@eset.com

Andrew Lee
Office of the CEO
ESET North America
San Diego, USA
andrew.lee@eset.com

Abstract: The allure of malware, with its tremendous potential to infiltrate and disrupt digital systems, is understandable. Criminally motivated malware is now directed at all levels and corners of the cyber domain, from servers to endpoints, laptops, smartphones, tablets, and industrial control systems. A thriving underground industry today produces ever-increasing quantities of malware for a wide variety of platforms, which bad actors seem able to deploy with relative impunity. The urge to fight back with "good" malware is understandable. In this paper we review and assess the arguments for and against the use of malicious code for either active defense or direct offense. Our practical experiences analyzing and defending against malicious code suggest that the effect of deployment is hard to predict with accuracy. There is tremendous scope for unintended consequences and loss of control over the code itself. Criminals do not feel restrained by these factors and appear undeterred by moral dilemmas like collateral damage, but we argue that persons or entities considering the use of malware for "justifiable offense" or active defense need to fully understand the issues around scope, targeting, control, blowback, and arming the adversary. Using existing open source literature and commentary on this topic we review the arguments for and against the use of "malicious" code for "righteous" purposes, introducing the term "righteous malware". We will cite select instances of prior malicious code deployment to reveal lessons learned for future missions. In the process, we will refer to a range of techniques employed by criminally-motivated malware authors to evade detection, amplify infection, leverage investment, and execute objectives that range from denial of service to information stealing, fraudulent, revenue generation, blackmail and surveillance. Examples of failure to retain control of criminally motivated malicious code development will also be examined for what they may tell us about code persistence and life cycles. In closing, we will present our considered opinions on the risks of weaponizing code.

Keywords: *malware, weaponize, malicious code, active defense, cyber conflict*

1. INTRODUCTION

On November 23 of 2013, news reports appeared stating that the National Security Agency of the United States (NSA) had installed malware on 50,000 computers around the world.[1] Three days later, Langner published a comprehensive analysis of Stuxnet.[2] Regardless of whether you agreed with all of Langner's conclusions, or regarded the reports of NSA malware deployment as fact or an erroneous allegation, these events served as a powerful reminder that the use of malicious code for nation state purposes is no longer a theoretical concern, but a present reality with serious socio-political and economic consequences. We will mention just some of these consequences as we argue that there is an urgent need for broader understanding of the merits and pitfalls of malicious code deployment, whether for cyber offense, active cyber defense, or cyber espionage, including legal and illegal surveillance for nation state or law enforcement purposes.

Numerous events over the last twenty years have demonstrated that malicious code has great potential as a means of infiltrating and disrupting digital systems of all kinds, for all manner of motives. Online markets now exist within which criminals and countries alike can acquire all of the means necessary for a malware campaign. With access to malware now easier than ever, the use of malicious code for either active defense or direct offense holds great fascination for nation states. Commercial suppliers are emerging to meet the demand, such as KEYW and Endgame.[3] Yet the literature of cyber conflict frequently notes that the deployment of malicious code by nation states is problematic.[4]

Unfortunately, detailed descriptions of the exact nature of the problems posed by weaponizing code are hard to find, a situation that we consider to be a problem in itself because it tends to create the impression that the objections to malware deployment are addressable. In turn, this could lead to the assumption that deployment of malicious code by nation states is inevitable. In the context of human conflict, to ascribe inevitability to an act that in reality requires a conscious decision is to court danger. Nation states can chose not to deploy malicious code and we will argue that more of them may make that choice if the problems inherent in malicious code deployment are better understood.

Clearly, more light must be shed on these issues at all levels, from the citizenry to the military, to the body politic. In this paper we elucidate the problems inherent in malicious code deployment by nation states and law enforcement agencies by first reviewing a list of reasons for thinking that a "good virus" is a bad idea. However, we distinguish the idea of a good virus designed to perform acts widely seen as beneficial, like backing up databases or patching systems, from code written to perform acts that benefit the deployer to the detriment of the target. We propose the term "righteous malware" for the latter. We also propose that any plans to deploy righteous malware be checked against the list of objections to good viruses, and then further evaluated relative to addition considerations that we present.

1 "NSA infected 50,000 computer networks with malicious software," NRC, Nov. 23, 2013. Available: http://www nrc.nl/nieuws/2013/11/23/nsa-infected-50000-computer-networks-with-malicious-software

2 R Langner, "To Kill a Centrifuge: A technical analysis of what Stuxnet's creators tried to achieve," Nov. 2013. Available: http://www.langner.com/en/resources/papers

3 M Riley and A Vance, "Cyber Weapons: The New Arms race," BusinessWeek, Jul. 20, 2011. Available: http://www.businessweek.com/magazine/cyber-weapons-the-new-arms-race-07212011 html

4 Tallinn Manual, p.53

After considering the possible benefits of righteous malware we will conclude with an attempt to understand why some people still favor deploying malware in spite of longstanding objections from those who deal with malware on a daily basis.

2. DEFINING MALWARE AND MOTIVES

For a working definition of malicious code we thought it fitting to use the one provided by the National Security Agency of the United States (NSA) in its 2007 publication: *Guidance for Addressing Malicious Code Risk*.[5] We note that this document borrows from the Committee for National Security Systems (CNSS) Instruction 4009 *National Informational Assurance (IA) Glossary*,[6] signed in 2006 by Lieutenant General Michael Hayden. The entry for malicious code reads: "software or firmware intended to perform an unauthorized process that will have adverse impact on the confidentiality, integrity, or availability of an IS [Information System]."

The NSA document goes on to clarify that malicious code includes both unauthorized software that has an adverse effect, and authorized software that, when used improperly, has an adverse effect, noting: "This may include software in which exploitable faults have been intentionally included." Clearly, this view of malicious code encompasses logic bombs and backdoors coded into software and firmware during design and development, as well as the more commonly discussed phenomena such as viruses, worms, and Trojans. One could argue that it also includes causing industrial control software to increase the speed of an electric motor, such as you might find in a centrifuge.

The meat of the NSA's guidance on malware is found in the section headed "Malicious Code in the Software Life Cycle" which reviews threats, vulnerabilities, and mitigation strategies across the seven life cycle stages listed in Table I.

TABLE 1: THE SOFTWARE LIFE CYCLE IN SEVEN STAGES

1. Acquisition
2. Requirements
3. Design
4. Construction.
5. Testing
6. Installation (delivery, distribution, installation)
7. Maintenance (operation, maintenance, and disposal)

What is striking about this table is that the general public, and possibly too many information and communication technology (ICT) professionals, think of malicious code as being a stage seven problem. Despite this popular perception of malware as something inserted into systems after they are installed, for the purposes of this paper we will use malware to refer to all

5 *Guidance for Addressing Malicious Code Risk*, NSA, 2007.
6 *National Informational Assurance (IA) Glossary*, CNSS National Security Systems Instruction 4009, 2006.

malicious code, not least because the NSA itself is alleged to have deployed backdoors in hardware, presumably at stage three or four.[7]

The idea of code that automatically inserts itself into a computer system at stage seven has been around almost as long as computers themselves. We refer to the concept of the "good virus," sometimes referred to as the "beneficial virus," self-replicating which does something positive, like encrypt files or patch code, in a fully automated and unsupervised manner.[8] However, both goodness and benefit are in the eye of the beholder, or in this case, in the opinion of the system owner on which the automated code is running. If you discern an unauthorized process on your network and find that its function is to email all of your engineering drawings to another country you are not likely to call it good or beneficial, in your opinion it is malicious.[9] Of course, the recipient of your drawings may find the arrangement beneficial and consider the code that delivers them to be good, even though it is, by all definitions, malware.

For this reason we introduce a new term to assist in the discussion of malware used for allegedly legitimate purposes: righteous malware. The following definition of righteous malware adds the aspect of motive to the purpose of the code: software or firmware deployed with intent to perform an unauthorized process that will impact the confidentiality, integrity, or availability of an information system to the advantage of a party to a conflict or supporter of a cause. We use the terms conflict and cause to distinguish righteous malware from malicious code that is motivated purely by financial gain. The party might be a person or group of persons, such as a nation state or agent thereof, or non-state actors, or even so-called hacktivists. What they have in common is the belief that their use of malware is justified, despite the fact that owners of systems and data impacted by the code are unlikely to agree.

While the concept of righteous malware is very different from that of good viruses, we assert that the persistent allure of the latter contributes to the persistence of the notion that malware can be deployed in a controlled manner to achieve, at least in the eyes of the deployer, beneficial results, such as hindering the process of enriching uranium that might be used to build nuclear weapons.

3. THE GOOD VIRUS PROBLEM

The allure of using self-replicating computer code to perform beneficial tasks dates back at least as far as the 1980s when it was explored by Dr. Fred Cohen.[10] Some early virus writing efforts were inspired by this concept.[11] Unfortunately, the results ranged from annoying to expensive. However, the idea of beneficial viruses has proved surprisingly immune to discouragement,

7 T. Simonite, "NSA's Own Hardware Backdoors May Still be a "Problem from Hell", Oct. 8, 2013. Available: http://www.technologyreview.com/news/519661/nsas-own-hardware-backdoors-may-still-be-a-problem-from-hell/

8 C. Peikari, "Fighting Fire with Fire: Designing a "Good" Computer Virus," Informit, Jun. 2011. Available: http://www informit.com/articles/article.aspx?p=337309&seqNum=2

9 R. Zwienenberg, "ACAD/Medre.A 10000's of AutoCAD files leaked in suspected industrial espionage," We Live Security, Jun. 21, 2012. Available: http://www.welivesecurity com/2012/06/21/acadmedre-10000s-of-autocad-files-leaked-in-suspected-industrial-espionage

10 F. Cohen, "Computational Aspects of Computer Viruses," *Computers & Security*, 8, 1989, pp. 325–344.

11 For example, the 1982 Xerox Worm designed to enable distributed computation, see D. Harley, R. Slade, et al, *Viruses Revealed*, Osborne/McGraw-Hill, 2006, p. 56.

prompting antivirus researchers to make repeated public statements of the problem in an effort at dissuasion, most notably in 1994, when Vesselin Bontchev, then a research associate at the Virus Test Center of the University of Hamburg, published an article titled: *Are "Good" Computer Viruses Still a Bad Idea?*[12]

Despite the many changes in the technology landscape that have occurred in the two decades since that paper was published, it is still a useful starting point for understanding objections to the deployment of malware. We think that a review of problems with the release of self-replicating code that was created to do good makes a convenient starting point for assessing the virtue of employing any kind of code designed to execute without permission or through deception.

One reason to use Bontchev's list is that it summarizes extensive input from a group of antivirus experts. Bontchev asked participants in VirusL/comp.virus,[13] an electronic forum dedicated to discussions about computer viruses, to list all the reasons why they thought the idea of a "beneficial" virus was flawed. From their responses Bontchev produced "a systematized and generalized list of those reasons" of which there were twelve, grouped into three categories: technical, ethical and legal, and psychological. The reasons are presented in Table II.

TABLE 2: REASONS WHY GOOD VIRUSES ARE A BAD IDEA

Technical Reasons	
Lack of Control	Spread cannot be controlled, unpredictable results
Recognition Difficulty	Hard to allow good viruses while denying bad
Resource Wasting	Unintended consequences (typified by the "Morris Worm")
Bug Containment	Difficulty of fixing bugs in code once released
Compatibility Problems	May not run when needed, or cause damage when run
Effectiveness	Risks of self-replicating code over conventional alternatives
Ethical and Legal Reasons	
Unauthorized Data Modification	Unauthorized system access or data changes illegal or immoral
Copyright and Ownership Problems	Could impair support or violate copyright of regular programs
Possible Misuse	Code could be used by persons will malicious intent
Responsibility	Sets a bad example for persons with inferior skills, morals
Psychological Reasons	
Trust Problems	Potential to undermine user trust in systems
Negative Common Meaning	Anything called a virus is doomed to be deemed bad

We recommend that anyone considering the deployment of malicious code, either for offense or active defense, use this table as a basic checklist of concerns that need to be addressed (a more advanced checklist will be supplied later).

12 V. Bontchev, "Are 'Good' Computer Viruses Still a Bad Idea?" Proc. EICAR'94 Conf., pp. 25-47.

13 Virus-L and comp.virus were a mailing list and online forum respectively, now archived at Google group, located at https://groups.google.com/forum/#!forum/alt.comp.virus

Consider a scenario in which a nation state is considering deployment of a virus designed to analyze cyber attacks against the deployer's systems, then identify the systems that are the source of the attack, and attempt to disable those systems in a counter attack.[14]

How does this plan measure up to the checklist? Frankly, we see problems in all twelve areas but will highlight just a few. Firstly, we doubt that such a program could be written in a way that would: rule out unanticipated actions that interfered with the attack code control mechanisms (Lack of Control); and prevent unanticipated and harmful reactions in all systems traversed during or after the counter attack (Compatibility Problems). We further doubt that this code could achieve its objective without detection, which would result in it being blocked by commercial antivirus programs (Recognition Difficulty Problem[15]).

While legal niceties (Unauthorized Data Modification) and excessive use of resources (Resource Wasting) may not bother the nation state behind the counter attack code, these are issues that may bother its citizens if the program comes to light. Spending taxpayer money to create code which is quickly co-opted by criminals to attack taxpayers (Possible Misuse) is also likely to be very unpopular. Of course, if the makers of the code solve all of these problems and achieve a successful deployment that defeats a serious attacker, the project may appease criticism in the area of Responsibility. However, a lack of success could undermine confidence in technology (Trust Problems) and lead to economic contraction.[16] Clearly the road to successful malware deployment is fraught with problems, as many failed malicious code campaigns attest.[17]

Of the above problems, the one that seems to have received the most attention in the literature of cyber conflict is control. However, even the most compelling examination of whether or not adequate levels of control over malware are achievable acknowledges that controls cannot prevent all problems: "Despite the care with which cyber weapon controls may be developed, there is always the possibility of undesired effects such as affecting the wrong target. The ability to control malware is only as good as the intelligence informing its development".[18]

A large part of that intelligence involves knowing the environment in which your malware will seek to achieve its righteous ends. Yet this process may not be able to fully anticipate every eventuality. What if the target changes some of the software or hardware it is running just moments before or after the malware is deployed? Is the malware going to be smart enough to detect such changes and shut itself down? When you look at the experience of the commercial software industry, which conducts a massive amount of pre-launch product testing, you see that every product launch plan invariably includes support staff and engineers standing by to deal with the inevitable problems that simply could not be predicted.

14 A scenario akin to the anti-viral virus referenced by Enn Tyugu, "Command and Control of Cyber Weapons," 4th International Conference on Cyber Conflict, NATO CCDCOE, 2012, p. 334.

15 Despite headlines to the contrary, commercial antivirus products frequently detect, identify, and block previously unknown malware, including that deployed by government entities. See R. Lipovsky, "German Policeware: Use the Farce…er, Force…Luke," We Live Security, Oct. 10, 2011. Available: http://www.welivesecurity.com/2011/10/10/german-policeware-use-the-farce-er-force-luke/

16 S. Cobb, "NSA and Wall Street: online activity shrinks, changes post-Snowden," We Live Security, Nov. 4, 2013. Available: http://www.welivesecurity com/2013/11/04/nsa-wall-street-online-activity-shrinks-post-snowden/

17 S. Cobb, "When malware goes bad: an historical sampler," We Live Security, Nov. 31, 2013. Available: http://welivesecurity.com/2013/11/30/when-malware-goes-bad-an-historical-sampler

18 D. Raymond, G. Conti, et al, "A Control Measure Framework to Limit Collateral Damage and Propagation of Cyber Weapons," 5th International Conference on Cyber Conflict, 2013.

We realize that proponents of righteous malware could counter this analysis by asserting the following: If anything goes wrong it will not be a problem because nobody will know it was us. This assertion reflects a common misunderstanding of the attribution problem, which is defined as the difficulty of accurately attributing actions in cyber space. While it can be extremely difficult to trace an instance of malware or a network penetration back to its origins with a high degree of certainty, that does not mean "nobody will know it was us." There are people who know who did it, most notably those who did it. If the world has learned one thing from the actions of Edward Snowden in 2013, it is that secrets about activities in cyber space are very hard to keep, particularly at scale, and especially if they pertain to actions not universally accepted as righteous.

Before moving on from the good virus checklist we should note that attribution was not listed as a problem for "the beneficial virus" back in 1994. After all, many virus writers proudly claimed their creations precisely because they thought they had created something beneficial (or at least functional with no intentional ill effects). Only when illegal activities rose to the fore as the primary motive for virus writing did malicious code attribution become an issue, initially for purposes of prosecution. Attribution becomes a critical issue when malicious code is used for cyber espionage or cyber attack, although it may not perceived to be a problem by those who deploy malware for motives they deem righteous. Responsibility for malware can be plausibly denied (with varying degrees of success, see Mandiant report[19]), or it can be tacitly acknowledged if you want to make a point (as may have been the case with Stuxnet[20]). And, indeed, there are good reasons why an agency involved in such attacks might wish to claim responsibility for an attack, though this is something of a double-edged sword. The fact remains that the perpetrators know who they are, and one day they may talk.

4. RIGHTEOUS MALWARE

Of course, a lot has changed in the two decades since Bontchev's paper laid out the reasons why a consensus of antivirus researchers think a virus designed with the best of intentions is a bad idea. As active participants in the antivirus community, we have not observed any change in that consensus over the years and we have heard the reasons against intentional malware deployment reiterated many times, yet we continue to see malware intentionally released into the wild with what its deployers believe to good intentions, such as waging "war on terror" and "war on drugs".[21]

One development we have observed over the last twenty years is an increase in the use of malicious code that is not self-replicating and therefore, one could argue, not subject to all of the problems ascribed to viruses and worms. We will concede that deploying righteous malware that is designed to work without self-reproductive abilities will address some of the problems we have listed, but this design choice also limits the capabilities of the malware. Furthermore, it does not mean that the malware will not be reproduced, either inadvertently (for example, when

19 Mandiant, "APT1: Exposing One of China's Cyber Espionage Units," Feb. 2013. Available: http://intelreport.mandiant.com/Mandiant_APT1_Report.pdf

20 R Langner, ibid, p.16

21 Reuters, "U.S. directs agents to cover up programs used to investigate Americans," Aug. 5, 2013. Available: http://www.reuters.com/article/2013/08/05/us-dea-sod-idUSBRE97409R20130805

an infected system is cloned or archived), or intentionally (by someone who has discovered it and wants to re-use it).[22]

Another change in recent years has been the growth of criminal enterprises founded on the exploitation of all kinds of malicious code. There is now a well-established system of markets in which to buy and sell all of the components necessary to carry out a malware campaign, from system infection through to mule services for turning purloined data into cash.[23] Davison of labor and specialization have enabled advances in efficiency and expertise not seen when a malware campaign has to be conducted end-to-end by a single campaigner (known in the last century as simply a virus writer).[24]

The rapid evolution of a market-based malware industry has turned the Possible Misuse problem identified in 1994 into an Inevitable Misuse problem today. It is not an exaggeration to say that the efforts by nation states to develop righteous malware fuel the criminal enterprise of malware production, delivery, and exploitation, to say nothing of making a market in zero day vulnerabilities.[25] Even when code itself is not re-used, techniques observed in weaponized malware may be quickly appear in criminal malware. For example, the creators of Stuxnet are widely considered to be pioneers in the use of stolen code-signing certificates to facilitate the spread of malware.[26] Today, the practice is mainstream and found in malware targeting the financial assets of consumers and corporations around the world.[27] Stuxnet also highlighted the benefits of modular malware design in which an existing infection could be enhanced with additional capabilities. Today, all the best banking malware sports a modular framework able to accept new tasking, leveraging the investment in infection to maximize returns.[28] Having a hard time recruiting money mules to convert stolen banking credentials into cash? Push a distributed denial of service (DDoS) module to your network of compromised machines and rent them out.

There may be an even bigger re-use problem. We are not experts in military history, doctrine, or philosophy, so we are unaware of the correct word for the following category of weapons: the ones you deliver to your enemies in re-usable form. Examples we can think of are rocks, arrows, throwing spears, and non-returning boomerangs. These weapons are delivered intact, available for re-use by the recipients, assuming they, the recipients and the weapons, are not too badly damaged by the act of delivery. Whatever the correct term for this ancient category of weapon, we think it includes the most modern of weapons, righteous malware. In fact, it is

22 R. Langner, ibid, p. 20.

23 B. Krebs, "The value of a hacked email account," Krebs on Security, Jun. 2013. Available: http://krebsonsecurity.com/2013/06/the-value-of-a-hacked-email-account

24 S. Cobb, "The Industrialization of Malware: One of 2012's darkest themes persists," We Live Security, Dec. 31, 2012. Available: http://www.welivesecurity.com/2012/12/31/the-industrialization-of-malware-one-of-2012s-darkest-themes-persists

25 Tom Simonite, "Welcome to the Malware-Industrial Complex," MIT Technology Review, Feb. 13, 2013. Available: http://www.technologyreview.com/news/507971/welcome-to-the-malware-industrial-complex

26 Tom Simonite, "Stuxnet Tricks Copied by Computer Criminals," MIT Technology Review, Sep. 12, 2012. Available: http://www.technologyreview.com/news/429173/stuxnet-tricks-copied-by-computer-criminals

27 J. Boutin, "Code certificate laissez-faire leads to banking Trojans," We Live Security, Feb. 21, 2013. Available: http://www.welivesecurity.com/2013/02/21/code-certificate-laissez-faire-banking-trojans. Also R. Lipovsky, "Back to School Qbot, now Digitally Signed," We Live Security, Sep. 7, 2011. Available: http://www.welivesecurity.com/2011/09/07/back-to-school-qbot-now-digitally-signed

28 ESET, "Hesperbot: A New Advanced Banking Trojan in the Wild," Sep. 9, 2013. Available: http://www.welivesecurity.com/wp-content/uploads/2013/09/Hesperbot_Whitepaper.pdf

perhaps true to say that righteous malware is unique in that you are giving away your weapons, tactics, and designs, simply by using them.[29]

Almost by definition, righteous malware is code that you deliver to the victim/target in working order, whether via email, browser exploit, USB key, firmware update, or embedded chipset. This raises the very real possibility that the recipient can discover the code, reverse engineer it, and use it against you. As Rustici has pointed out, the practical impossibility of knowing whether or not this has happened is just one of many ways in which cyber weapons differ from conventional weapons.[30] For example, satellite imagery cannot provide you with an early warning of a cyber attack. Your adversary cannot be seen marshaling cyber weapons on your borders, not least because there are no borders in cyberspace.

Ascertaining the cyber capabilities of potential adversaries is a non-trivial task further complicated by globally dispersed non-state actors and an international sub-culture of hackers for hire and malicious code delivery systems for purchase or rent. There is also a risk of tremendous inequality in targets. Take for instance, a terrorist group operating a malware network from an undeveloped or chaotic country with the intention of attacking infrastructure in a developed nation. The group may feel it has little to lose if it deploys righteous malware that provokes a cyber response. Is there enough digital infrastructure in their country for a retaliatory cyber-attack to have a punishing affect. Not only that, but when dealing with people who have little interest in preserving their own lives or the lives of others, cyber capabilities may not offer much deterrence.[31]

While there has been extensive discussion of cyber conflict relative to theories and codes of war, much of it directed at a goal we support, limiting the use of cyber weapons, we argue that righteous malware has already created fallout, at a level we can ill afford to ignore. Three months after the press started reporting on the Snowden papers, we asked a representative sample of American adults who use the Internet how the revelations had affected their sentiment, in general and with respect to specific aspects of Internet usage. About one in five agreed with this statement: "Based on what we have learned about government surveillance I have done less banking online."[32] A similar percentage said they were less inclined to use email. We found that 14% had cut back on online shopping.

Whether this sentiment will lead to an ecommerce contraction remains to be seen. Our subjects said they were cutting back, not cutting off the Internet. We do not know if doubts will persist, but bear in mind that this sentiment was assessed before people heard about the following, all of which would tend to further exacerbate the problem: the NSA's mapping of Americans' social contacts, capturing of their address books and contact lists, hacking into connections

29 A. Anghaie, "STUXNET: Tsunami of Stupid or Evil Genius?" Infosec Island, Jun. 1, 2012. Available: http://infosecisland.com/blogview/21507-Stuxnet-Tsunami-of-Stupid-or-Evil-Genius.html

30 R. Rustici, "Cyberweapons: Leveling the International Playing Field," Parameters, Vol. XLI, No. 3, Autumn 2011. U.S. Army War College, p. 32.

31 A. Lee, "Cyberwar: Reality, Or A Weapon of Mass Distraction?" Proceedings Of the 22nd Virus Bulletin International Conference, 2013, pp. 292 - 300.

32 S. Cobb, "Survey says 77% of Americans reject NSA mass electronic surveillance, of Americans," We Live Security, Oct. 29, 2013. Available: http://www.welivesecurity.com/2013/10/29/survey-says-77-of-americans-reject-nsa-mass-electronic-surveillance-of-americans

between data centers owned by Yahoo and Google, and infecting 50,000 systems with righteous malware. All indications are that new and equally unsettling revelations will continue well into 2014.[33]

One more survey finding that should be cause for concern is that half of respondents said that they were now less likely to trust technology companies such as Internet service providers and software companies. One way to look at that number is as an erosion of public trust in the very entities to which people normally turn for help in securing their systems and protecting their digital domains. Ironically, the source of mistrust is the other place that people turn for protection: the government.

Trust in the very software that is designed to defeat malicious code has also been shaken. In October of 2013, a coalition of digital rights organizations and academics published an 'open letter' asking for clarification on vendor policies regarding cooperation with government agencies and/or law enforcement using state-sponsored Trojan code.[34] Historically, there is no evidence that any antivirus company had ever collaborated with any nation state or law enforcement agency to further the spread of righteous malware. The letter demonstrates the corrosive effect that revelations of government malware deployment can have on both trust and common sense. Several antivirus companies responded by pointing out they had already refuse to give passes to righteous malware.[35] Others pointed to their exposure of righteous malware in the past, and the improbability than any such software could be "allowed" by the antivirus industry.[36]

One term that keeps occurring to us as we look at the effect of righteous malware deployment on our industry and on the wider economy, is attrition. We fear that nations are at a tipping point, the downside of which is a slow but steady erosion of that essential building block of prosperous societies: trust. Malware of any kind eats away at trust in networked systems, the very systems that form the critical infrastructure and industrial fabric of developed countries. We are already seeing righteous malware deployment eroding trust in the institutions that deliver and defend that infrastructure.

While each new development of malicious code is met with new security measures, and the network continues to function for most people most of the time, each new round of attack and counter-measure further encumbers the technology and reduces its potential to deliver the continued productivity gains upon which much future economic growth is predicated. Not only that, but savvy operators will find other channels to avoid detection, while millions of the innocent will have their privacy and security compromised.

33 S. Cobb, personal notes on Gen. R. Hayden's comments to The Ecommerce Summit, San Diego, Nov. 23, 2013.

34 J Leyden, "Antivirus bods grilled: Do YOU turn a blind eye to government spyware, The Register," Nov. 5, 2013. Available: http://www.theregister.co.uk/2013/11/05/av_response_state_snooping_challenge

35 M. Hypponen, "F-Secure Corporation's Answer to Bits of Freedom," News rom the lab, Nov. 6, 2013. Available: http://www.f-secure.com/weblog/archives/00002636.html

36 R. Marko, A. Lee, et al, "ESET response to Bits of Freedom open letter on detection of government malware," We Live Security, Nov. 11. Available: http://www.welivesecurity.com/2013/11/11/eset-response-to-bits-of-freedom-open-letter-on-detection-of-government-malware/

5. THE BENEFITS OF RIGHTEOUS MALWARE

Whether used for offense or active defense, malicious code can boast numerous advantages, in its own right or relative to conventional weapons. Malicious code is an essential component of cyber weaponry, which is envisioned by Rustici as leveling the international playing field.[37] We examine these benefits and counter some of the arguments against deployment of righteous malware listed in the preceding section.

A. Less deadly than kinetic weaponry

The argument has been made that using code instead of kinetic weapons is more humane.[38] Cyber-attacks, if used carefully, certainly seem as if they could provide tactical advantage in ways that are not physically harmful and that do not require troop deployments.[39] If one nation state is convinced it has to take action against another, surely it is better to threaten, or execute, an attack on the networked systems of its opponent, where the effects may range from inconvenient to life-threatening, but stop short of deadly force. The demoralizing effect of sustained inconvenience, like intermittent malware-induced power outages, should not be under-estimated. However, as Rustici has pointed out, the benefits of weaponized code in this context do not accrue equally to all nations.[40] In fact, they stack the cards against developed nations whose greater reliance on cyber everything leaves them most vulnerable to such attacks, and in favor of less cyber reliant nations that nevertheless have rich traditions of learning and innovation.

B. Works well for espionage

Undoubtedly, malware can greatly facilitate espionage. Electronic espionage can definitely strengthen a nation's hand against its enemies and appears to be less encumbered by international treaties and norms governing nation state behavior. However, espionage is not without political and economic risks for the countries that engage in it, as the world discovered in 2013. We do not know if revelations of large-scale electronic spying, including widespread use of righteous malware, will have long term negative effects on nations, or the commercial entities perceived to be enablers of this activity. We remain alert to signs of economic contraction, retaliatory network Balkanization, or other potential ill effects. The apparent impact of being seen as an enabler on the price of shares in Cisco, shown in Figure 1, is a useful visual reminder.[41]

37 R. Rustici, ibid, p. 32.

38 D. Denning, "Obstacles and Options for Cyber Arms Controls," Paper presented at Arms Control in Cyberspace Conference, Berlin, Jun. 2001: "instead of dropping bombs on an enemy's military communication systems, for example, cyber forces could take down the system with a computer network attack, causing no permanent damage and no risk of death or injury to soldiers or civilians. The operation would be more humane and should be preferred over more destructive alternatives."

39 J. Andress and S. Winterfeld, *Cyber Warfare Techniques, Tactics and Tools for Security Practitioners*, Syngress, 2011.

40 R. Rustici, ibid.

41 D. Meyer, "Cisco's gloomy revenue forecast shows NSA effect starting to hit home," Gigaom, Nov. 14, 2013. Available: http://gigaom.com/2013/11/14/ciscos-gloomy-revenue-forecast-shows-nsa-effect-starting-to-hit

FIGURE 1: THE NOVEMBER 14 "NSA EFFECT" ON CISCO STOCK

C. Less expensive that physical options

Nation states have surely asked this question: Why spend billions to arm our country with sophisticated kinetic weapons and the trained soldiery needed to deploy them, when we can obtain malware-based cyber weapons for mere millions? Unfortunately, the allure of lower prices, particularly in terms of human cost, evaporates when cyber weapons are examined from a technical perspective. Many fall short of traditional definitions of weaponry and into various categories of strategic support for kinetic warfare, such as disabling or disrupting key infrastructure as an adjunct or precursor to kinetic attack.[42]

One can argue that the issue of cyber war is more properly considered an issue of security: systems security, network security, and due diligence on part of its operators. The majority of security breaches today—be they commercial, consumer, or military—are as a result of systems failure and human error, and the legal responses considered should perhaps be limited to such.[43] This problem lends itself to a situation of diminishing return, escalating cost and a strengthened enemy. It may be possible to gain a brief advantage initially, but this is soon lost if the enemy increases his own security posture in response.

Consider the case of Estonia, which came under digital attack in 2008. The damage was certainly quantifiable, but the end result was, that, paradoxically, the confident, even defiant, response by the Estonian government, and the prompt support lent by the North Atlantic Treaty Organization and the European Union, may have left Estonia in a stronger technical, political, and moral position after the attacks than before.[44] Therefore, expense cannot only be measured in development and deployment cost, but also in reputational and operational cost. There is also an issue of attribution. It stands to reason that one would want an enemy to recognize that an attack has been carried out, certainly if the purpose is deflection of further kinetic activity due to a show of strength. Strategically, this would require exposure (as perhaps is the case with the

42 A. Lee, ibid.

43 T. Guo. "Shaping Preventive Policy in "Cyber War" and Cyber Security: A Pragmatic Approach" *J. Cyber Sec. Info. Sys*. 1-1 14 (2012). Available: http://works.bepress.com/tony_guo/2.

44 T. C. Wingfield, "International Law and Information Operations," in *Cyberpower and National Security*, F. D. Kramer, H. S. Starr, & K. L. Wentz (Eds., pp. 525-542). Washington DC: Potomac Books, 2009.

US and Israel claiming responsibility for the Stuxnet malware). Such exposure though, raises the stakes, creating an arms race.

Eventually, we may reach equilibrium, where we understand that use of our own cyber-weaponry will result in an equally destructive response from our enemy. The 'nightmare' scenario is one where our 'enemy' has less to lose in terms of connected infrastructure, a strong defensive posture, and an advanced cyber weapons deployment capability. This will certainly be a costly situation for an attacker.

6. CONCLUSIONS

We see many problems with, and arguments against, the deployment of malicious code by anyone for any purpose. We have shown that many of these objections have been raised before. We have discussed additional risks, some of which have recently been demonstrated in world events. We also note additional objections and obstacles from those seeking to understand the relationship between cyber weapons and concepts like the laws of armed conflict (LOAC),[45] *jus in bello*,[46] and *jus ad bello*.[47] Review of these is beyond the scope of this paper but we have included them in our summary table of questions to ask before proceeding with the deployment of righteous malware, Table III.

TABLE 3: CONSOLIDATED LIST OF RIGHTEOUS MALWARE QUESTIONS TO ASK

Control	Can you control the actions of the code in all environments it may infect?
Detection	Can you guarantee that he code will complete its mission before detection?
Attribution	Can you guarantee that he code is deniable or claimable, as needed?
Legality	Will the code be illegal in any jurisdictions in which it is deployed?
Morality	Will deployment of the code violate treaties, codes, and other international norms?
Misuse	Can you guarantee that none of the code, or its techniques, strategies, design principles will be copied by adversaries, competing interests, or criminals
Attrition	Can you guarantee that deployment of the code, including knowledge of the deployment, will have no harmful effects on the trust that your citizens place in its government and institutions including trade and commerce.

Note that we are talking about objections to the deployment of righteous code, not its development. Detailed discussion of this important distinction is beyond the scope of this paper.

Frankly, we do not anticipate the imminent outbreak of outright cyber war, but we do anticipate that righteous malware will continue to be a serious problem. As one of the authors has previously stated: "Cyber-attack capabilities, then, seem most likely to be useful in the future

45 J. Healey, "When 'Not My Problem' Isn't Enough: Political Neutrality and National Responsibility in Cyber Conflict," 4th International Conference on Cyber Conflict, NATO CCDCOE, 2012.

46 R. Fanelli and G. Conti, "A Methodology for Cyber Operations Targeting and Control of Collateral Damage in the Context of Lawful Armed Conflict," 4th International Conference on Cyber Conflict, NATO CCDCOE, 2012. p. 327.

47 Reese Nguyen, "Navigating Jus Ad Bellum in the Age of Cyber Warfare," 101 Cal. L. Rev. 1079 (2013). Available at: http://scholarship.law.berkeley.edu/californialawreview/vol101/iss4/4

precisely in the same ways as they are being used now: causing temporary and generally non-injurious disruption to systems, whether to embarrass, shame or disrupt organizations, or to steal useful information, and perhaps prevent or delay technological progress."[48] To this we would now add the risks of economic contraction and trust erosion that come from secret cyber operations, including the use of righteous malware, being made public.

Finally, we need to ask why nation states and law enforcement agencies persist in the deployment of righteous malware. Do those who are in a position to approve such deployments still think the potential benefits outweigh the risks? Naturally, we would argue that the risks have not been fully appreciated, a recurring problem in information assurance if risk assessment methodologies developed in simpler times are applied to rapidly evolving technology. When assessing the location for a proposed data center you can use historical tables to put a number on the likelihood of floods, high winds, and other threat events. But what about assessing risks to systems on which novel attacks are possible? Just because a country has never experienced a particular type of attack, such as malicious code damaging a critical infrastructure, does not mean the probably of this happening in the future is zero. Indeed, it is entirely possible and efforts are underway in many countries to defend against such eventualities.

The best place to find an explanation of why a government that openly acknowledges its vulnerability to cyber attack would simultaneously engage in cyber attack, as the US arguably has done, may be the Gerras critical thinking model,[49] which has already been applied to an exploration of the prudent limits of automated cyber attack.[50] The model is apt because it derives from a military setting and the two entities most heavily involved in decisions to deploy righteous malware in the US are, at the time of writing, under military command. We fear that one or more of the nine common logical fallacies enumerated by Gerras could lead to a damaging weaponized code deployment of which the right questions were not asked or critically answered.

Acknowledgments

The authors acknowledge the assistance of fellow ESET researchers including Lysa Myers, David Harley, Aryeh Goretsky, and Righard Zwienenberg.

48 A. Lee ibid

49 S. Gerras, "Thinking Critically About Critical Thinking: A fundamental guide for strategic leaders," Carlisle Barracks: U.S. Army War College, Department of Command, Leadership, and Management, August 2008. Available: http://www.au.af.mil/au/awc/awcgate/army-usawc/crit_thkg_gerras.pdf

50 J. Caton, "Exploring the Prudent Limits of Automated Cyber Attack." 5th International Conference on Cyber Conflict, NATO CCDCOE, 2013.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Changing the game: The art of deceiving sophisticated attackers

Nikos Virvilis
Cyber Defence and Assured Information Sharing
NATO Communications and Information Agency
The Hague, Netherlands

Oscar Serrano Serrano
Cyber Defence and Assured Information Sharing
NATO Communications and Information Agency
The Hague, Netherlands

Bart Vanautgaerden
Consultant for NATO Office of Security, InfoSec
NATO HQ
Brussels, Belgium

Abstract: The number and complexity of cyber-attacks has been increasing steadily in the last years. Adversaries are targeting the communications and information systems (CIS) of government, military and industrial organizations, as well as critical infrastructures, and are willing to spend large amounts of money, time and expertise on reaching their goals. In addition, recent sophisticated insider attacks resulted in the exfiltration of highly classified information to the public. Traditional security solutions have failed repeatedly to mitigate such threats. In order to defend against such sophisticated adversaries we need to redesign our defences, developing technologies focused more on detection than prevention. In this paper, we address the attack potential of advanced persistent threats (APT) and malicious insiders, highlighting the common characteristics of these two groups. In addition, we propose the use of multiple deception techniques, which can be used to protect both the external and internal resources of an organization and significantly increase the possibility of early detection of sophisticated attackers.

Keywords: *Advanced persistent threat, deception, insiders, honeypot, honey net, honey tokens*

1. INTRODUCTION

In the last decade, there have been a large number of advanced, well-orchestrated cyber-attacks against industry, military and state infrastructures. The main goal of most of these attacks is the exfiltration of large amounts of data. For example in 2006, China was accused of downloading

10 to 20 terabytes of data from the US NIPRNet[1] Military Network [1], and in 2008 a USB drive was deliberately left in the parking lot of a US Department of Defense facility in the Middle East for the purpose of subsequently infecting a laptop computer connected to the United States Central Command, resulting in the exfiltration of sensitive information [2]. In 2010 "Operation Aurora" targeted more than 20 organizations including Google, Adobe, Symantec and US defence contractors [3]. Furthermore, cyber-attacks intended to cause physical destruction have been known to occur [4].

While it is believed that these attacks were originated by different threat actors, they share certain common features and some of them have been categorized as advanced persistent threats. The term "advanced persistent threat" (APT), coined by the US Air Force in 2006[2], is not strictly defined and loosely covers threats with a number of characteristics in common. The definition of APT given by the National Institute of Standards and Technology (NIST) [5] is:

> *"An adversary with sophisticated levels of expertise and significant resources, allowing it through the use of multiple different attack vectors (e.g. cyber, physical, and deception) to generate opportunities to achieve its objectives, which are typically to establish and extend its presence within the information technology infrastructure of organizations for purposes of continually exfiltrating information and/or to undermine or impede critical aspects of a mission, program, or organization, or place itself in a position to do so in the future; moreover, the advanced persistent threat pursues its objectives repeatedly over an extended period of time, adapting to a defender's efforts to resist it, and with determination to maintain the level of interaction needed to execute its objectives."*

In addition, organizations face the always present threat of malicious insiders, a clear example of which is Edward Snowden, who recently downloaded 50,000 to 200,000 classified documents belonging to the US National Security Agency [6]. This incident arose shortly before Bradley Manning was convicted and sentenced to 35 years in prison in connection with the largest data leak in US history [7].

The ability of current security solutions to address such attackers has been questioned openly [8] [9] [10] [11], with authors stating that prevention techniques (e.g. network-intrusion prevention and antivirus products), and especially those focused on signatures, will never be able to successfully address sophisticated attacks.

The shortcomings of signature-based detection are well accepted, and the research community has focused on the use of anomaly-based detection systems. However, the effectiveness of such systems has also been challenged. Sommer and Paxson [12] describe anomaly detection as flawed in its basic assumptions. Research relies on the belief that anomaly detection is suitable for finding new types of attacks, however it is known that machine learning techniques are best suited to finding events similar to ones seen previously. Therefore, these approaches show promising detection possibilities for specific (training) data sets, but are subject to serious operational limitations.

1 Non-classified Internet Protocol Router Network

2 It was initially used as a generic term to describe intrusions without disclosing the classified threat name [32].

APTs use unique attack vectors and custom-built tools tuned for the particular target, making detection very challenging whether either signature or anomaly detection techniques are used. In this context, deception techniques are valuable for monitoring enterprise networks and identifying attack preparation and subsequent exploitation.

We present in this paper: (a) a comparison of APTs and malicious insiders, highlighting the common characteristics of these two attacker groups and suggesting that malicious insiders should be considered a subcategory of APTs, and (b) a proposal for the use of multiple deception techniques, such as social network avatars, fake (honey token) Domain Name System (DNS) records, and HTML comments – none of which, to the best of our knowledge, has been proposed before – that can significantly increase the likelihood of early detection in every phase of an attack's life-cycle.

The remainder of the paper is structured as follows: Section 2 presents related work. Section 3 focuses on the similarities between APTs and malicious insiders, as we believe that both can be treated in the same way for the purpose of detecting sophisticated attacks. In Section 4, we propose a number of deception techniques for protecting both the Internet-facing and the internal assets of an organization. Conclusions and further work are reported in Section 5.

2. RELATED WORK

Decoys, a popular strategy long used in warfare, played an important role during the Second World War [13] and the Cold War [14]. Decoys are also an integral part of electronic warfare strategies [15], however they are rarely used in the cyber domain. The first general reference to cyber decoys is attributed to Clifford Stoll, who describes them in his 1989 novel 'The Cuckoo's Egg' [16]. More than 10 years later, Spitzer described mechanisms for the detection of insider attacks using honeypots [17] and honey tokens, which share similar characteristics with honey files, as described in [18] and [19].

Elsewhere, honeypots [20] [21] have been proposed for attack detection [22] [23], including detection and analysis of botnets/worms, while honey nets [24] have been proposed as an effective means for the classification of network traffic and the detection of malicious users on Wi-Fi networks [25].

Honey files that include beacon signaling are discussed by Bowen et al. [26], who propose an architecture for monitoring multiple system events, including user interactions with a set of previously marked honey files. Similar work was pursued by Whitham [27], who introduced canary files, which have similar characteristics to honey files. Most of the published work concentrates on the creation and distribution of "perfectly believable" honey files [28], which contain certain properties that make them indistinguishable from real files to malicious users and at the same time are enticing enough to attract attention. Finally, researchers have also proposed embedding, in legitimate documents, code that will be automatically executed when the files are opened and will initiate a connection to a monitoring server [29] to provide a means of detecting unauthorized access.

To the best of our knowledge, there has been no research on the use of deception techniques for the detection of advanced persistent threats (APT).

3. ADVANCED PERSISTENT THREATS AND INSIDERS

The definition of a malicious insider based on Silowash et al. [30] is:

> *"... a current or former employee, contractor, or business partner who meets the following criteria:*
>
> - *has or had authorized access to organization's network, system, or data*
> - *has intentionally exceeded or intentionally used that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems."*

The motives of insiders vary, and can be based on revenge or can be financial, ethical or political [31].

APTs and malicious insiders share specific characteristics that markedly differentiate them from traditional (e.g. opportunistic) attackers:

- Their attacks require detailed planning [32], and are spread over a long period of time in an effort to evade detection. Insiders have a potential advantage over APTs in planning their attack, as they may be aware of existing security controls. This is very likely if an insider holds a privileged position (e.g. an administrator is expected to have knowledge of the deployed security mechanisms and potentially has the access rights to control them, while a less privileged user would not [32]). Nevertheless, experience has shown that APTs have also managed to reach their goals while evading detection without prior knowledge of the infrastructure [3].
- Both groups are willing to spend a substantial amount of time exploring all possible attack paths for reaching their goals, including social engineering and deception [32]. APT groups tend to have teams of highly skilled individuals with access to important resources (financial, technical, intelligence). Malicious insiders, although they work mostly alone, as in the case of Manning and Snowden, might also have well developed technical skills.
- Both are interested in maintaining access to the penetrated infrastructure and continuing the exfiltration of data for as long as possible.

The main difference between the two types of attackers is that malicious insiders have by definition authorized access to the infrastructure and potentially even to the servers storing sensitive information (e.g. file servers, database servers), while APTs need to gain unauthorized access.

APTs and insider threats are currently considered to be two different threat groups. However, given the known instances of APT groups blackmailing or bribing an insider to perform a malicious action on their behalf [33], we strongly believe that malicious insiders should be regarded as a subset of advanced persistent threats.

Robust models have been proposed for the detection of insider threats [34], however they assume that the malicious insider(s) will perform the entire attack life-cycle on their own (information gathering, exploitation, exfiltration). Yet, in the Stuxnet case [33], a malicious insider was used only to deliver the payload, while the rest of the exploitation was performed in an automated way. Such an attack strategy, which combines APT with the insider element, poses a serious challenge for insider threat detection models.

Taking into consideration the substantial resources available to APT groups [35], we can expect similar attacks to occur in the future, and thus we strongly believe that further research is necessary to augment the detection capabilities of such models against combined insider-APT attacks.

4. DECEPTION TECHNIQUES

Detection of network-based security threats can significantly increase the likelihood of detecting APT and insider attacks by monitoring the operational networks/infrastructure as well as the unused IP address space (“darknets”) [36]. The APT attack life-cycle [37] consists of several stages: attack preparation and initial compromise, establishing a foothold, escalation of privileges, internal reconnaissance, exploitation of systems and exfiltration of data. For the sake of simplicity in this paper, we group these stages into two general phases: attack preparation (information gathering), and exploitation and data exfiltration.

A. Phase 1: Attack preparation (information gathering)

The initial step of an APT attack is the preparation phase, in which perpetrators gather as much information as possible about their target. Identification of the operating system, third-party software and publicly accessible services (e.g. web servers, mail servers) of the organization is crucial for planning a successful attack. Information related to the security solutions in use (intrusion-detection and intrusion-prevention systems, endpoint protection, data leakage prevention) is also important for the attackers to have, as it allows them to test their tools and techniques in advance.

An additional element of the preparation phase is collection of information about employees, their positions in the organization, their skills and their connections with other employees. Using such information, APTs can create highly targeted spear-phishing campaigns. For example, if an attacker has identified an employee working in the human resources (HR) department as well as his supervisor, he can send a spoofed email from the email address of the supervisor to the employee, asking him to review an attached file (e.g. a curriculum vitae). The attachment can be a malicious Word or PDF file that when opened will execute the attacker’s payload. The fact that the email originates from a person known to the victim significantly increases the likelihood of its being accepted as legitimate.

In order to address this first phase of the attack life-cycle, we propose the following deception techniques.

1) DNS honey tokens
DNS honey tokens are proposed as a complementary technique to honeypots.

Because attackers will try to identify Internet-facing systems/services belonging to the organization, defenders can deploy honeypots spread over the unused public IP range of the organization. Based on the fact that these systems will not be publicly listed (e.g. not returned as part of a search query with a link to the organization's web site), a connection attempt could be due to: (a) user error (mistyping an IP address), (b) an automated attack such as a worm randomly scanning the IP address space to find vulnerable hosts to compromise, or (c) an attacker trying to identify all publicly accessible systems and services of the organization.

However, the use of honeypots generates a substantial amount of noise owing to the vast number of automated attacks on the Internet [38]. In addition, it can be difficult to differentiate between an automated non-targeted attack and a targeted one.

We propose a technique that is simpler to implement than honeypots and will significantly limit the number of false positives occurring. It consists of inserting fake DNS records (a type of honey token) in the DNS servers.

Attackers are very likely to use "brute force" for common subdomains or attempt a zone transfer [39] on an organization's DNS servers to try to identify interesting resources (e.g. sub-domains, servers) as part of their information-gathering process. By creating a small number of fake DNS records on the authoritative DNS servers of the organization and configuring them to initiate an alert when these specific records are requested, defenders can receive an early warning of DNS-related information-gathering attempts against their infrastructure.

2) Web server honey tokens
The public web servers of an organization are another fruitful source of information for attackers. We propose three ways of using honey tokens to help detect malicious web-site visitors:

- Addition of fake entries in robots.txt files
- Use of invisible links
- Inclusion of honey-token HTML comments.

A robots.txt file [40] is a simple text file located in the root folder of the web server, which legitimate bots (e.g. Google bot) parse to identify which folders on the web server they should not access and index. The file is one of the first places that attackers (and automated web-vulnerability scanning tools) look for potentially sensitive directories. By including non-existing directories such as "/admin" or "/login" in the robots.txt file and monitoring for access requests to these locations, administrators can be alerted to visitors with malicious intentions.
The inclusion of invisible links (e.g. white links on white font) at random parts of the web site(s), pointing to non-existing (but interesting from the attacker's perspective) resources, can

serve a similar purpose. Although these links will be invisible to legitimate visitors, they will be detected by the crawling tools that attackers are likely to use. A request for such a fake URL should raise an alert.

A final deception mechanism, particularly useful for web sites that support authentication, is the inclusion of fake accounts in HTML comments. Legitimate users have no need to review the source code of a web page, however attackers frequently do in trying to identify vulnerabilities. The inclusion of a comment such as the following in the HTML source code of a login page is very likely to tempt the attacker to use it:

<!--test account: admin, pass: passworD123. Please remove at the end of development!-->.

Once more, an attempt to login with these credentials is a clear indication of malicious activity.

3) Social network avatars

Social networks are an invaluable source of information for attackers. In order to identify malicious activity, we propose the creation of avatars (fake personas) on the major social networks. It is important that the avatars appear to be realistic, having connections with people from both inside and outside the organization and with positions that are likely to be of interest to the attackers (e.g. HR department, financial department, developer, etc.). In addition, such avatars should have real, but very closely monitored, accounts in the organization (e.g. active directory accounts), as well as valid email addresses. Interaction with the avatars should be regularly monitored (friend requests, private messages, attachments, etc.).

External applicants interested in applying for a position in the organization may contact the human resources avatar (producing a false positive). However, because internal employees should know the correct contact details, communication between an internal employee and the avatar can be considered suspicious. Such interaction could be an indication that the employee's account has been compromised, as will be any login attempts using the avatar account(s).

B. Phase 2: Exploitation and data exfiltration

The second step of the APT life-cycle is exploitation of the target. The attackers, after gaining access to the internal network (e.g. taking advantage of 0-day vulnerabilities, social engineering, spear-phishing attack), will start the exfiltration process and try to identify (a) systems that they can compromise to be used as alternative access points to the network (in case the initial ones are detected and quarantined), and (b) systems that may contain the information they are seeking or that can help them access that information.

In order to address this phase of the attack we propose the following deception techniques.

Deception techniques for network layer defences

In a medium to large organization in which hundreds or even thousands of systems are active, identifying the location of targeted information is not a trivial task. Attackers will need to explore the network, hop between networks and exploit multiple systems. Use of darknets and

or honey nets can be invaluable in detecting such actions, as attackers may eventually access them, raising an immediate alert.

1) Darknets

A darknet, also known as a black hole, Internet sink or darkspace, is a portion of routed, unallocated IP space in which no workstations/servers or other network devices are located. Access to such regions of the network can occur by a legitimate mistake (e.g. a user mistyping an IP address), however multiple connection attempts should be considered suspicious. Monitoring such segments for connection attempts can be an easy-to-deploy and effective mechanism, however it is not guaranteed that attackers will actually access these parts of the network.

2) Honey nets

Honey nets [41] are used for monitoring larger and/or more diverse networks in which one honeypot may not be sufficient. Defenders can use honey nets to create multiple fake systems in the same IP ranges as legitimate systems/servers. An attacker who gains access to a specific network segment is very likely to access these fake systems along with the real ones. Interaction with such systems should be very closely monitored as it is a strong indication of an active attack.

Deception techniques for application layer defences

The same techniques used for detecting malicious activity on external web servers can be used for protecting internal ones. Furthermore, as the majority of organizations make use of database and files servers on their internal networks, we propose the following deception techniques for the detection of malicious activity against those servers.

1) Database server honey tokens

Use of honey tokens in the databases can be used to highlight malicious activity. For example, a number of fake patient records (with fake patient names) can be introduced in a hospital's patient database. Attempts to access such records should be considered highly suspicious. However, database auditing must be enabled for logging the queries, and this will negatively affect the performance of the database.

2) Honey files

As described in related work, a number of strategies for creating decoys (honey files) have been proposed, focusing either on the generation of perfectly believable decoys or the modification of legitimate files to include some alerting functionality. Although the practical use of perfectly believable decoys has been questioned, use of legitimate files can interfere with the operation of the organization.

We propose a combination of file system auditing and the generation of honey files with potentially interesting content for attackers (e.g. passwords.docx, new_investments.pdf, etc.). These files should be spread across the file servers of the organization and/or even workstations, however the latter will increase the number of false positive alerts [29]. In environments in which document markings are used (i.e. TOP SECRET, SECRET, etc.), those can easily be

taken advantage of for generating decoy files. For example, it is easy to mark a fake document with a classification higher than the maximum level authorized to be stored in the system. Since such a situation indicates a security infraction, all users interacting with that document should report the infraction to security, and non-reported interactions are therefore highly suspicious.

A number of detection techniques can be implemented, including:

- File system auditing [42], which will log access attempts to these files.
- Inclusion of code that when executed will report back to a monitoring server. This can be achieved by using JavaScript for PDF files, or remote images that are downloaded when the document is opened [43].
- Inclusion of bait information, such as fake credentials, that attackers may try to use.

3) Honey accounts
Creating bait accounts (such as accounts for avatars) is an additional way of detecting attackers, as any interaction (e.g. login attempts) with these accounts is a clear indication of an active attack. This could be combined with the aforementioned example of placing bait files on file servers, where a file with fake credentials (user names and passwords) could be created. An attacker who has gained access to the file is very likely to try to use these accounts to gain further access to the network and as a result will immediately raise an alert.

C. Evaluation

Preventive techniques will eventually fail against sophisticated attackers [9], thus it is critical to switch our focus to detection measures. Use of deception techniques such as those proposed will significantly increase the possibility of detecting attacks early in the attack life-cycle, allowing defenders to mitigate a threat before the attackers achieve their goals.

Although the effectiveness of such measures against insiders is open to discussion, based on the fact that insiders are likely to be aware of their use and will try to evade them, we believe that combining a number of deception techniques will make evasion very difficult, provided that it is not the insider who has implemented the deception measures.

There is a risk that the introduction of deception techniques to monitor internal assets may interfere with the normal functioning of the organization. Therefore we have focused on techniques that are non-intrusive and that will seldom result in false positives. We recommend integrating them into an anomaly-detection system [44] incorporating some additional data sources, such as HR databases (e.g. user data, leave data), access rights matrices, net-flow data, etc., as this would further increase the reliability of the detection system and limit the number of false positives occurring.

5. CONCLUSIONS AND FUTURE WORK

Insider threats and APTs have a number of characteristics in common and should be considered as a single threat type. Furthermore, current security solutions do not effectively address

sophisticated attackers. We propose the use of deception techniques as a potential solution to this multidimensional problem. Several deception techniques can be used to increase the possibility of early detection at any stage of the attack life-cycle. Furthermore, such techniques can be combined with traditional collection and correlation systems to further increase the capability to detect sophisticated attackers.

Finally, future work will focus on the improvement of existing insider threat detection models through the introduction of deception techniques.

REFERENCES

[1] R. Marquand and B. Arnoldy, "China Emerges as Leader in Cyberwarfare," *The Christian Science Monitor*, Aug. 2007.

[2] J. P. Farwell and R. Rohozinski, "The New Reality of Cyber War," *Survival Global Politics and Strategy*, vol. 54, no. 4, pp. 107–120, 2012.

[3] K. Zetter, "Google hack attack was ultra sophisticated, new details show," *Wired Magazine*, vol. 14, 2010.

[4] R. Langner, "Stuxnet: Dissecting a Cyberwarfare Weapon," *Security Privacy, IEEE*, vol. 9, no. 3, pp. 49–51, 2011.

[5] "Guide for Conducting Risk Assessments (Rev 1)," National Institute of Standards and Technology, Gaithersburg, USA, NIST Special Publication 800-30, Sep. 2012.

[6] M. Hosenball, "NSA chief says Snowden leaked up to 200,000 secret documents," Reuters, 14-Nov-2013. [Online]. Available: http://www.reuters.com/article/11/14/us-usa-security-nsa-idUSBRE9AD19B2013114.

[7] D. Nicks, *Private Bradley Manning, WikiLeaks, and the Biggest Exposure of Official Secrets in American History*. Chicago Review Press, 2012.

[8] E. Cole, *Advanced Persistent Threat Understanding the Danger and How to Protect Your Organization*. Newnes, 2012.

[9] R. Bejtlich, *The Practice of Network Security Monitoring Understanding Incident Detection and Response*. No Starch Press, 2013.

[10] N. Virvilis and D. Gritzalis, "The Big Four -- What we did wrong in Advanced Persistent Threat detection?," in *Availability, Reliability and Security (ARES), 2013 Eighth International Conference on*, 2013, pp. 248–254.

[11] N. Virvilis, D. Gritzalis, and T. Apostolopoulos, "Trusted Computing vs. Advanced Persistent Threats: Can a defender win this game?," in *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*, 2013, pp. 396–403.

[12] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Security and Privacy (SP), 2010 IEEE Symposium on*, 2010, pp. 305–316.

[13] T. Holt, *The Deceivers Allied Military Deception in the Second World War*. Simon and Schuster, 2010.

[14] G. R. Mitchell, *Strategic Deception Rhetoric, Science, and Politics in Missile Defense Advocacy*. Michigan State Univ Press, 2000.

[15] K. B. Alexander, *Electronic Warfare in Operations U.S. Army Field Manual FM 3-36*. DIANE Publishing Company, 2009.

[16] C. Stoll, *The Cuckoo's Egg Tracking a Spy Through the Maze of Computer Espionage*. New York, NY, USA: Doubleday, 1989.

[17] L. Spitzner, "Honeypots: Catching the Insider Threat," presented at the 19th Annual Computer Security Applications Conference, 2003, pp. 170–179.

[18] J. Yuill, M. Zappe, D. Denning, and F. Feer, "Honeyfiles: Deceptive Files for Intrusion Detection," presented at the Fifth Annual IEEE SMC Conference Workshop on Information Assurance, 2004, pp. 116–122.

[19] M. Bercovitch, M. Renford, L. Hasson, A. Shabtai, L. Rokach, and Y. Elovici, "HoneyGen: An automated honeytokens generator," *in Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*, 2011, pp. 131–136.

[20] N. Provos and T. Holz, *Virtual Honeypots from Botnet Tracking to Intrusion Detection*. Pearson Education, 2007.

[21] R. Joshi and A. Sardana, *Honeypots A New Paradigm to Information Security*. Science Publishers, 2011.

[22] P. Wang, L. Wu, R. Cunningham, and C. C. Zou, "Honeypot detection in advanced botnet attacks," *International Journal of Information and Computer Security*, vol. 4, no. 1, pp. 30–51, 2010.
[23] C. C. Zou and R. Cunningham, "Honeypot-aware advanced botnet construction and maintenance," in *Dependable Systems and Networks, 2006. DSN 2006. International Conference on*, 2006, pp. 199–208.
[24] O. Thonnard and M. Dacier, "A framework for attack patterns' discovery in honeynet data," *digital investigation*, vol. 5, pp. S128–S139, 2008.
[25] B. M. Bowen, V. P. Kemerlis, P. Prabhu, A. D. Keromytis, and S. J. Stolfo, "A system for generating and injecting indistinguishable network decoys," *Journal of Computer Security*, vol. 20, no. 2, pp. 199–221, 2012.
[26] B. Bowen, M. Ben Salem, A. Keromytis, and S. Stolfo, "Monitoring Technologies for Mitigating Insider Threats," in *Insider Threats in Cyber Security*, vol. 49, C. W. Probst, J. Hunker, D. Gollmann, and M. Bishop, Eds. Springer US, 2010, pp. 197–217.
[27] B. Whitham, "Canary Files: Generating Fake Files to Detect Critical Data Loss from Complex Computer Networks," presented at the Second International Conference on Cyber Security, Cyber Peacefare and Digital Forensic (CyberSec2013), Malaysia, 2013.
[28] J. A. Voris, J. Jermyn, A. D. Keromytis, and S. J. Stolfo, "Bait and Snitch: Defending Computer Systems with Decoys," 2013.
[29] M. Ben Salem and S. Stolfo, "Decoy Document Deployment for Effective Masquerade Attack Detection," in *Detection of Intrusions and Malware, and Vulnerability Assessment*, 2011, vol. 6739, pp. 35–54.
[30] G. J. Silowash, D. M. Cappelli, A. P. Moore, R. F. Trzeciak, T. Shimeall, and L. Flynn, "Common Sense Guide to Mitigating Insider Threats (4th Edition)," *Software Engineering Institute*, no. 677, 2012.
[31] B. Gellman and J. Markon, "Edward Snowden says motive behind leaks was to expose 'surveillance state'," *Washington Post*, 09-Jun-2013.
[32] J. Hudson, "Deciphering How Edward Snowden Breached the NSA," *Venafi*, 12-Nov-2013. [Online]. Available: http://www.venafi.com/deciphering-how-edward-snowden-breached-the-nsa/?goback=%2Egde_135559_member_580642620779687 1171#%21. [Accessed: 21-Nov-2013].
[33] M. Kelley, "The Stuxnet Virus at Iran's Nuclear Facility was Planted by an Iranian Double Agent," *Military & Defense*. 13-Apr-2012.
[34] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and D. Gritzalis, "An insider threat prediction model," in *Trust, Privacy and Security in Digital Business*, Springer, 2010, pp. 26–37.
[35] D. Fisher, "What have we learned: FLAME malware," *Threat Post*. 15-Jun-2012.
[36] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of Internet Background Radiation," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, New York, NY, USA, 2004, pp. 27–40.
[37] "Exposing One of China's Cyber Espionage Units," Mandiant, Feb. 2013.
[38] M. Gebauer, "Warfare with Malware: NATO Faced with Rising Flood of Cyberattacks," *Spiegel*, Mons, Belgium, 26-Apr-2012.
[39] C. Edge, W. Barker, B. Hunter, and G. Sullivan, "Network Scanning, Intrusion Detection, and Intrusion Prevention Tools," in *Enterprise Mac Security*, Springer, 2010, pp. 485–504.
[40] J. Hendler and T. Berners-Lee, "From the Semantic Web to social machines: A research challenge for AI on the World Wide Web," *Artificial Intelligence*, vol. 174, no. 2, pp. 156–161, 2010.
[41] L. Spitzner, "The Honeynet Project: Trapping the Hackers," *Security Privacy*, IEEE, vol. 1, no. 2, pp. 15–23, Mar. 2003.
[42] D. Melber, "Securing and Auditing High Risk Files on Windows Servers," *Windows Security*. 17-Apr-2013.
[43] B. Bowen, M. Ben Salem, S. Hershkop, A. Keromytis, and S. Stolfo, "Designing Host and Network Sensors to Mitigate the Insider Threat," *Journal of Security & Privacy IEEE*, vol. 7, pp. 22–29, Nov. 2013.
[44] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11994 – 12000, 2009.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

The Deployment of Attribution Agnostic Cyberdefense Constructs and Internally Based Cyberthreat Countermeasures

Jason Rivera
United States Army
Georgetown School of Foreign Service
Washington, D.C., United States
jhr47@georgetown.edu

Forrest Hare
United States Air Force
Johns Hopkins School of
Advanced International Studies (SAIS)
Washington, D.C., United States
fhare@gmu.edu

Abstract: Conducting active cyberdefense requires the acceptance of a proactive framework that acknowledges the lack of predictable symmetries between malicious actors and their capabilities and intent. Unlike physical weapons such as firearms, naval vessels, and piloted aircraft—all of which risk physical exposure when engaged in direct combat—cyberweapons can be deployed (often without their victims' awareness) under the protection of the anonymity inherent in cyberspace. Furthermore, it is difficult in the cyber domain to determine with accuracy what a malicious actor may target and what type of cyberweapon the actor may wield. These aspects imply an advantage for malicious actors in cyberspace that is greater than for those in any other domain, as the malicious cyberactor, under current international constructs and norms, has the ability to choose the time, place, and weapon of engagement. This being said, if defenders are to successfully repel attempted intrusions, then they must conduct an active cyberdefense within a framework that proactively engages threatening actions independent of a requirement to achieve attribution.

This paper proposes that private business, government personnel, and cyberdefenders must develop a threat identification framework that does not depend upon attribution of the malicious actor, i.e., an attribution agnostic cyberdefense construct. Furthermore, upon developing this framework, network defenders must deploy internally based cyberthreat countermeasures that take advantage of defensive network environmental variables and alter the calculus of nefarious individuals in cyberspace. Only by accomplishing these two objectives can the defenders of cyberspace actively combat malicious agents within the virtual realm.

All views and concepts expressed in this paper originate solely with the authors and do not represent the official positions or opinions of the U S Army, U S Air Force, or U S Department of Defense

Keywords: *active defense, attribution agnostic cyberdefense construct, internally based cyberthreat countermeasures*

1. INTRODUCTION

Thomas Hobbes, in his political text *Leviathan*, postulated that, in the absence of governance, humanity lives within a "state of nature" and that life within this state of nature is nasty, brutish, and short.[1] The text goes on to describe the development of the Social Contract—a societal construct between a ruler and the ruled in which the ruled agree to live under the laws and guidance of the ruler, as long as the ruler provides an environment in which the life, liberty, and property of the ruled are protected.[2] Today, most industrialized nations live under the safety of a social contract and are generally protected, both physically and legally, from those who wish to do harm.

Cyberspace, unlike the physical domain, is arguably still characterized by Hobbes' state of nature. While there are rules and laws that have carried over from the physical domain, they are sparingly enforced within the cyber domain. The porous borders and anonymous nature of cyberspace create an ideal environment for those with criminal intent. Although there have been a variety of collaborative efforts to construct international laws and norms to regulate cyberspace, these efforts amount to little more than an international convention; i.e., no nation or individual is forcefully obligated to abide by the laws and norms of other nations in cyberspace. Furthermore, the prevalence of the attribution problem (the difficultly of positively attributing a nefarious action in cyberspace to a specific actor) is a confounding factor that makes defensive operations increasingly complex within the cyber domain.[3] Cyberspace, therefore, is likely to remain in a state of nature for the near to medium-term future, which implies that cyberdefenders are going to have to develop creative and proactive methods to defend their networks from within.

Given the amorphous nature of cyberspace and this paper's endeavor to develop an attribution agnostic cyberdefense construct, it is imperative to put forth a definition of the nature of cyberspace. Science fiction author William Gibson first defined cyberspace in 1982 as "a consensual hallucination experienced daily by billions of legitimate operators."[4] One could argue that the vast expansion of the domain and rapid advancements in technology have rendered this idea quaint. To confront today's realities more effectively, the White House developed a definition that is used today by the U.S. government:

1 Thomas Hobbes, *Leviathan* (New York: Continuum International Publishing Group, 2005), Vol. XIII, 9.
2 Celeste Friend, "Social Contract Theory," *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/soc-cont/ (accessed Oct. 14, 2013).
3 Martin C. Libicki, *Cyberdeterrence and Cyberwar* (Santa Monica, CA: RAND Corporation), 41.
4 Dani Cavallaro, *Cyberpunk and Cyberculture Science Fiction and the Work of William Gibson* (London: The Athlone Press, 2000), ix.

> [Cyberspace is] the interdependent network of information technology infrastructures including the Internet, telecommunications networks, computer systems and embedded processors and controllers in critical industries.[5]

The above definitions make an important point very clear: cyberspace is much more than just the Internet; it is, rather, a function of infrastructure and the use of the electromagnetic spectrum, as well as the social interactions that define cyberspace activity.[6]

Based on this characterization of cyberspace, this paper will propose two theoretical shifts in the perception and engagement of cyberthreats. First, it will address the need for cyberdefenders to develop attribution agnostic cyberdefense constructs. By attribution agnostic, this paper specifically refers to the development of security mechanisms that do not rely on attribution to levy deterrent effects, increase threat-actor risk, or deliver punitive measures. It follows that the anonymous nature of the Internet implies that cyberdefenders must stop attempting to achieve attribution and instead focus on gaining a thorough understanding of the organizations they are trying to defend; only then can they engage and counter nefarious tactics that are likely to be used against the defenders. Second, this paper will propose the concept of developing internally based cyberthreat countermeasures; i.e., strategies that are specifically designed and implemented to deter, detect, and defeat network-based threats from within the friendly network's boundaries. These countermeasures must be custom tailored to the specific organization they are designed to defend and designed in such a manner that they cause a quantifiable shift in the malicious actor's calculus, thereby raising the minimum threshold that must be crossed before the actor is willing to engage in malicious online activity. If these countermeasures are successfully implemented, network defenders should be able to deter and defeat cyberthreats without needing to achieve attribution or facing the technical and legal challenges of conducting counteroffensive response measures. This paper will begin by expanding on these two theoretical shifts before it explores some real-world examples of how these theories could be deployed in network environments.

2. CYBER ACTORS, ATTRIBUTION, AND ASSOCIATED CHALLENGES

A. The Attribution-Focused Model

This section begins with the assertion that cybersecurity is inherently different from conventional security. In an effort to deter and defeat adversaries prior to the exposure of critical assets, conventional security in the physical domain is typically attribution focused and outward facing; that is, one must have a target or know what they are going to strike prior to initiating a defensive/offensive response. While there are certain parallels between the two, the cyberspace domain has characteristics that make it difficult to apply an outward-facing security framework. This brings us to the threat spectrum presented in Figure 1 which outlines seven hypothetical actor-centric threats that a commercial or government entity could face against its physical location. The likelihood of a particular actor conducting a threatening action is highest on the right side of the spectrum and lowest on the left. Conversely, the severity of a threatening action

5 The White House, *Comprehensive National Cybersecurity Initiative* (Washington, DC: National Security Presidential Directive, 2008).

6 Forrest Hare, "The Interdependent Nature of National Cyber Security: Motivating Private Action for a Public Good," *George Mason University* (2010), 13.

is highest on the left side of the spectrum and lowest on the right. This model provides a sense of predictability in terms of what threat-actors will and will not do. While it would be possible for a foreign military power to invade and occupy the sovereign territory of another country, this action is least probable. On the other end of the spectrum, delinquents and petty thieves, though a more common threat, are generally limited in terms of the damage they could inflict on a major corporation or government entity and thus can be handled in a predictable manner, given that the proper security mechanisms are in place.

FIGURE 1

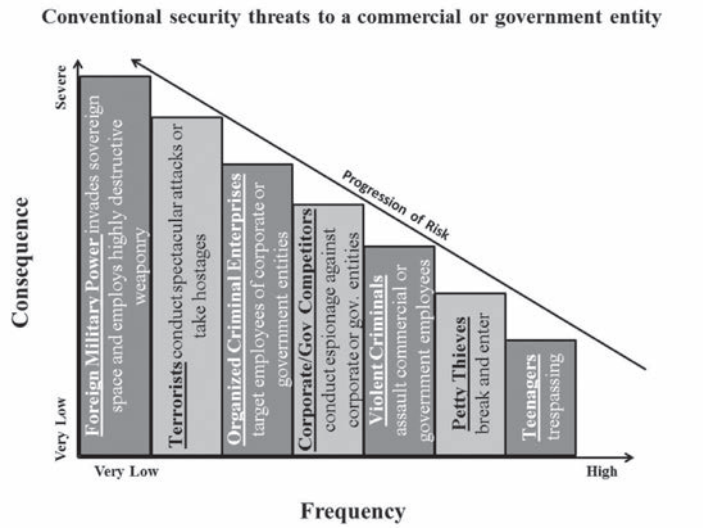
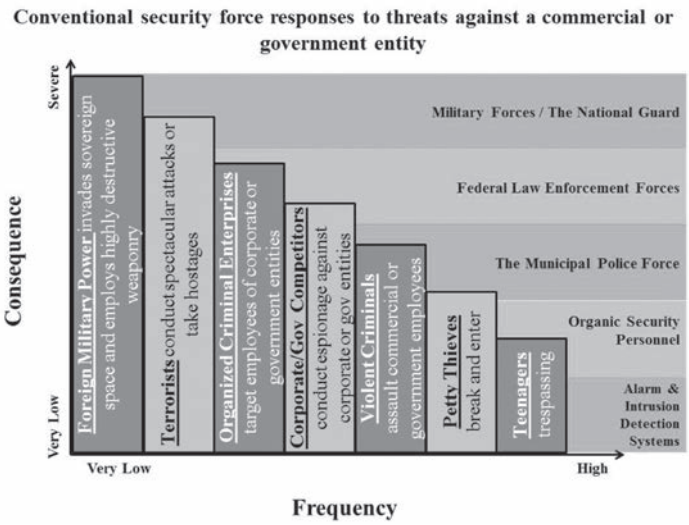


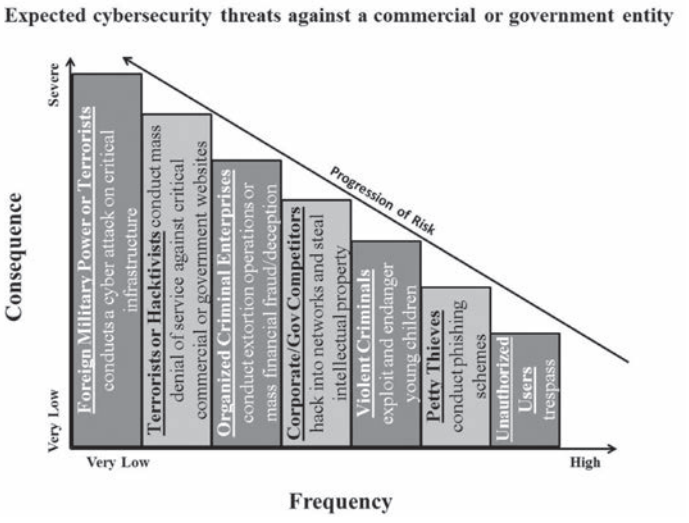
Figure 2 displays conventional responses based off attribution/identification of the nefarious actors. At the highest level of severity, friendly military forces will become involved in order to combat foreign military powers or terrorist threats, whereas low severity threats should be manageable by organic security personnel and/or intrusion-detection systems. Note that there is some level of crossover among the various security response forces, which implies a certain level of necessary cooperation. While there is sometimes friction within this system, this model is regularly adopted and employed by many industrialized nations and private-sector firms worldwide.

FIGURE 2



Naturally, as the Internet has become a more critical component in the day-to-day execution of commercial and government operations, cyberthreats also have become more prolific. In response, cyberdefense professionals have created attribution-specific threat models and defense apparatuses in a manner similar to those of the physical domain, as demonstrated in Figure 3.[7,8] Figure 3 closely resembles Figure 1 in many ways. The actors and their corresponding threats do vary slightly, but the overall threat apparatus remains largely the same.

FIGURE 3

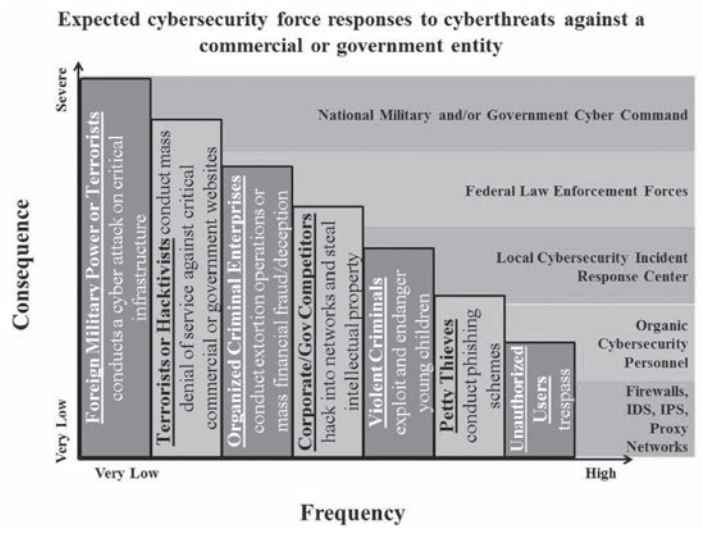


7 The threat-modeling apparatus used in this figure derives its premise from former Director of Cybersecurity Policy at the U.S. Department of Homeland Security, Mr. Andrew W. Cutts.

8 Andrew Cutts, "Warfare and the Continuum of Cyber Risks: A Policy Perspective," *The Department of Homeland Security* (2009), 3, 7.

Figure 4 follows the same security force response logic as Figure 2 and models responses in a similar escalatory manner. In this model, we expect organic cybersecurity personnel, along with various system-hardening measures such as firewalls and intrusion detection/prevention systems, to detect and defeat unauthorized users and/or petty thieves. On the opposite end of the spectrum, host-nation military/government cyber elements are expected to combat a foreign military's cyber capabilities or intrusion by terrorists. Furthermore, as shown in this model, we do not expect the friendly military force to conduct targeted operations against unauthorized users, nor do we expect foreign military powers to conduct phishing schemes or petty trespassing operations. It is at this point that an attribution-focused cybersecurity model becomes flawed, due to the asymmetric capabilities and intent as well as the requirement for attribution of actors operating in cyberspace.

FIGURE 4



B. Defensive Distortions and Critique of the Attribution-Focused Model

Within cyberspace, traditionally less powerful actors, such as unauthorized users in a sensitive network, can sometimes possess highly dangerous capabilities; this is because individual actors in the cyber domain benefit from asymmetric vulnerability relative to larger organizations such as governments or intelligence agencies.[9] Similarly, cyberspace allows foreign military powers, who are traditionally known for targeting adversarial military targets, to bypass national-level defense mechanisms and directly engage lower tier targets. This prevents cyberdefenders from accurately gauging the level of cyberthreat based on the type of aggressing actor, due to asymmetries between threat-actors and their capabilities and intent. Whereas defenders in the physical domain can reasonably assume that petty criminals do not have nuclear weapons and that foreign military powers will not rob the local McDonald's, this same categorical logic does not hold true in cyberspace. Low investment costs and low barriers to entry and exit further amplify asymmetric vulnerabilities, thereby creating defensive distortions.[10] Thus we are presented with two types of defensive distortions in cyberspace:

9 Joseph S. Nye, Jr., "Cyber Power," *Harvard Kennedy School Belfer Center for Science and International Affairs* (2010), 10.

10 Ibid., p. 13.

1. Military-grade defensive distortion: The ability of government, military, and other powerful entities to wield military-grade cyberweapons and capabilities in order to bypass a nation's national defense apparatus and interface directly with and conduct exploits against private citizens, companies, and other traditionally less defended targets.

2. Unauthorized user-access defensive distortion: The ability for an individual or small group of people to exploit the attribution problem in cyberspace and navigate through the porous portions of the cyber domain in order to conduct attacks, steal information, and/or otherwise levy threats that are typically beyond the capabilities of any one individual or small group of people within the physical domain.

The following are some historical examples of these two defensive distortions:

Unauthorized user access defensive distortions

- In 2012, Anonymous, a non-state-sponsored, loosely connected group comprised of individual hackers, managed to disrupt and degrade the websites of the U.S. Federal Bureau of Investigation, Department of Justice.[11]
- According to a Pentagon report leaked in early 2014, Edward Snowden, a lone actor and former National Security Agency contractor, downloaded 1.7 million classified intelligence files via his access to classified cyberspace networks;[12] this incident is widely considered to be the single largest breach of national security information in U.S. history.
- In 2009, a federal grand jury indicted Albert Gonzalez and two accomplices for conducting a SQL injection attack used in an international operation that compromised 134 million credit cards;[13] in late 2013, experts speculated that a network breach had occurred at Target Corp.'s point-of-sale (POS) terminals, resulting in the exposure and possible compromise of the credit and debit card information of up to 110 million customers.[14]

Military-grade defensive distortions

- Since 2006, a conventional Chinese military force known as the 2nd Bureau of the People's Liberation Army (PLA) General Staff Department's 3rd Department is reported to have targeted and compromised private-sector companies throughout the world, including at least 141 companies spanning 20 major industries.[15]

11 MSNBC.com staff and news services, "Anonymous says it takes down FBI, DOJ, entertainment sites," *NBC News Technology*, Jan. 19, 2012, http://www.nbcnews.com/technology/anonymous-says-it-takes-down-fbi-doj-entertainment-sites-117735 (accessed Oct. 15, 2013).

12 Associated Press, "Snowden obtained nearly 2 million classified files in NSA leak—Pentagon report," www *RT.com*, Jan. 9, 2014, http://rt.com/usa/snowden-downloaded-millions-documents-389/ (accessed Feb. 1, 2014).

13 Taylor Armerding, "The 15 worst data security breaches of the 21st Century," *COS Security and Risk*, Feb. 15, 2012, http://www csoonline.com/article/700263/the-15-worst-data-security-breaches-of-the-21st-century (accessed Feb. 1, 2014).

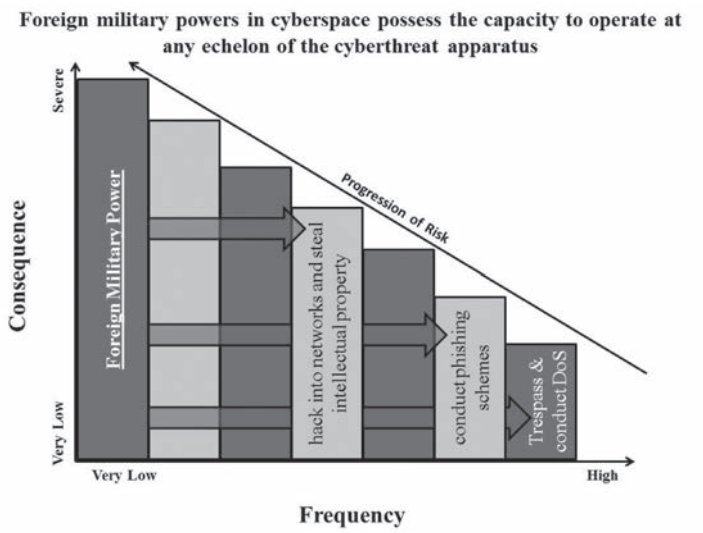
14 Tracy Kitten, "Target Breach: What Happened? Expert Insight on Breach Scenarios, How Banks Must Respond," *Bank Info Security*, Dec. 20, 2013, http://www.bankinfosecurity.com/target-breach-what-happened-a-6312/op-1 (accessed Feb. 1, 2014).

15 Why We Are Exposing APT1, "APT1: Exposing One of China's Cyber Espionage Units," *Mandiant* (2013), 6.

- From 2008 through late 2013, several media sources reported that Israel had gained access to Palestinian phone networks and demonstrated a capacity to send mass text messages to Palestinian citizens. In most cases, these text messages were used to conduct psychological operations against the Palestinian population, including one sent in 2012 that stated, "The next phase is on the way. Stay away from Hamas elements."[16] Another mass message, sent in October 2013, stated that "tunnels that were built by Hamas underground between Gaza and the Israeli-occupied territories cost millions of dollars that were supposed to be spent on the Gaza people."[17]

The above examples demonstrate the difficulties in defending cyberspace, as many malicious cyber actors successfully avoid attribution and often have the ability to circumvent traditional defensive constructs. Note in Figure 5 how a foreign military power is able to conduct cyber operations at the high-frequency end of the threat spectrum. This not only implies that powerful threats have the capacity to threaten entities that are less able to defend themselves, but also that there is a defensive distortion within the traditional national cybersecurity framework. By directly circumventing and therefore not inciting a defensive response from the friendly national military and/or government cyber force, an adversary wielding military-grade cyber capabilities is able to bring an overwhelming capacity to bear against systems that are not adequately hardened, while simultaneously operating safely below the attribution threshold necessary for a national-level response.

FIGURE 5

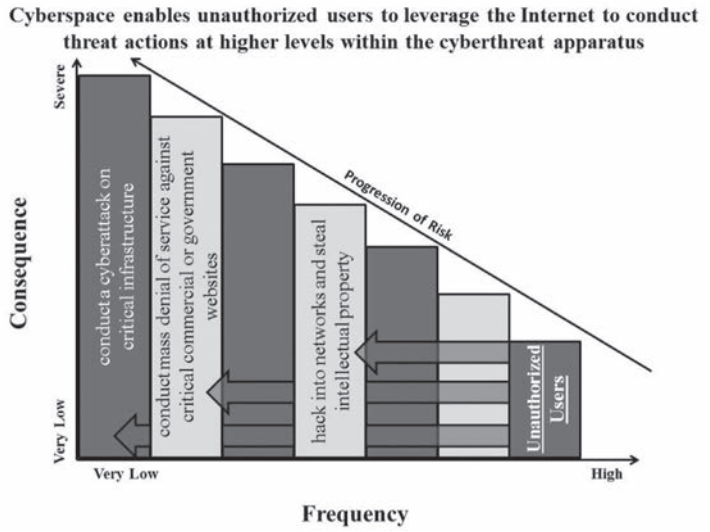


16 Lisa Goldman, "IDF sends text message to Gaza mobile phones: The next phase is on the way," *972 Mag*, Nov. 16, 2012, http://972mag.com/idf-sends-text-message-to-gaza-mobile-phones-the-next-phase-is-on-the-way/60046/ (accessed Feb. 1, 2014).

17 Associated Press, "Israeli text messages warn Gazans not to help Hamas build tunnels," *World Tribute*, Oct. 21, 2013, http://www.worldtribune.com/2013/10/21/israeli-text-messages-warn-gazans-not-to-help-hamas-build-tunnels/ (accessed Feb. 1, 2014).

On the other end of the spectrum, unauthorized users are able to wield capabilities that exceed the expectations of what traditional defensive frameworks ascribe to the individual. Figure 6 demonstrates the unauthorized user's capacity to inflict harm beyond the scope of what was possible prior to the prevalence of the Internet.

FIGURE 6



Consider a worst-case scenario, where the next insider threat is not a disenfranchised intelligence officer like Edward Snowden or Bradley Manning but a disgruntled nuclear engineer with enough computer savvy to cause a regional power crisis—or worse, a nuclear meltdown. In the cyberspace environment, unauthorized users have the ability to apply asymmetric vulnerabilities against traditionally hardened targets. Again, this implies another distortion within the traditional national cybersecurity framework, as the insider threat operates both beyond the locally emplaced defensive measures, often avoids attribution, and interfaces below the enforcement threshold of higher level cybersecurity force response entities.

The asymmetries inherent among threat-actors in cyberspace suggest the need for an attribution agnostic cyberdefense construct that focuses on the individual nature of the organization, its valuable cyberspace equities that are exposed to risk, and the organization's physical and network environment. Let us explore the development of such a construct in pursuit of the objective to implement an active, internally based defense.

3. THE ATTRIBUTION AGNOSTIC CYBERDEFENSE CONSTRUCT

An attribution agnostic cyberdefense construct (AACC) will analyze and depict the unique characteristics of an organization in a manner that enables defenders to deploy catered active defense solutions in the form of internally based cyberthreat countermeasures. Given this objective, defenders must learn to conceptualize their respective organizations and how they relate to cyberspace as a series of analytic components. The United States military community has developed a model that frames cyberspace within the context of three layers, which include the physical layer (both geographic and physical network components), logical network layer, and social layer (both persona and virtual persona components).[18] The AACC proposed in this paper derives its premise from this model and characterizes organizations as they relate to cyberspace via the following five distinct, yet related, components:

1. The Geopolitical Component: All organizations are subject to the constraints associated with their geographic locations, as well as the governing nation-state's laws and policies. This is an important factor in terms of analyzing an organization's capacity to conduct response actions in cyberspace. For example, U.S. law, per the Computer Fraud and Abuse Act, defines accessing a computer without authorization or exceeding authorized access as a criminal offense; therefore outlawing cyberspace response actions by private sector entities.[19] A commercial company in Indonesia, on the other hand, would likely face few to no repercussions for conducting aggressive response actions in cyberspace, as online law enforcement legislation in that country is virtually non-existent.[20]

2. The Physical Infrastructure Component: This component includes the physical aspects of an organization's computer infrastructure, electrical power resources, physical security layout, and public interface functionality. Physical infrastructure may include but is not limited to buildings and office space, physical domain security measures, electrical power connectivity, systems cooling, physical computing technology (hardware, servers, etc.), and communications equipment (satellite communications, VSAT dishes, telephone lines, etc.).

3. The Interface Component: This component encompasses the way an organization employs interface mechanisms to interact with cyberspace. The interface component includes the network gateway and networking identities used by organizational members. Passing through the Internet gateway can be achieved with a laptop, virtual machine thin client, smartphone, fax machine, etc. Once through the gateway, an individual assumes a virtual identity (username, email address, phone number, social media profile, etc.) to exchange information in cyberspace.

4. The Logical Network Component: This component comprises the electrons, bits and bytes, or 1s and 0s flowing to and from computer networked services using the

18 Training and Doctrine Command, "TRADOC Pamphlet 525-7-8: Cyberspace Operations Concept Capability Plan 2016-2028," *Department of the Army* (Fort Eusis, VA: GPO, 2011), 8.

19 18 U.S.C. § 1030: U.S. Code–Section 1030: Fraud and related activity in connection with computers.

20 Farisya Setiadi et al., "An Overview of the Development Indonesia National Cyber Security," *International Journal of Information & Computer Science* (2012),Vol. VI, 108.

Open Systems Interconnection (OSI) or TCP/IP layer models in terms of accurately addressing and directing the flow of information. This component is characterized by the logical connections an organization leverages to interact with cyberspace. An organization's logical network is comprised of switches, routers, various servers, firewall functions, and broadcast domains and is logically mapped via IP addressing and network routing protocol.

5. The Critical Information Component: This component comprises the societal purpose of an organization and is the most critical consideration for developing an effective cyberdefense construct. All computer networks are designed to process information, and information is, in general, processed in one of two ways.
    a. Information exchanged and processed by humans exists in the form of ideas; the most valuable ideas within an organization comprise that organization's intellectual property. Schematics, tradecraft, business strategies, formulas, and plans are some examples of intellectual property.
    b. Information exchanged and processed by machines exists in the form of protocol; the most important protocol within an organization comprises that organization's critical control systems. Electrical power switching, manufacturing processes, financial transaction systems, transportation systems, water/wastewater control systems, and temperature regulation systems are some examples of these critical control systems.

Once an organization is accurately characterized via the AACC, an appropriate internally based cyberthreat countermeasure must be developed in order to actively combat potential cyberthreats. If one thinks of the cyber domain as a fifth domain of human interactivity (the others being land, sea, air, and space), then the development of internally based cyberthreat countermeasures designed to defeat cyberthreats is a logical solution. Consider Germany's first anti-material rifle, known as the "T" Gewehr 13mm anti-tank rifle, which was developed in response to the Allies' introduction of tanks during World War I,[21] or the U.S. military's development of anti-ballistic missile technology in response to the Soviets' Intercontinental Ballistic Missiles.[22] Given historical precedence, it stands to reason that cyberdefenders should facilitate the development of internally based cyberthreat countermeasures designed to defend organizational assets from within friendly networks.

4. INTERNALLY BASED CYBERTHREAT COUNTERMEASURES

The creation of internally based cyberthreat countermeasures (IBCCs) shall be premised upon a key assumption: an adversary with malicious intent sufficiently resourced with time, capabilities, and personnel will inevitably compromise a friendly network. This assertion is reflected in the statements of leading cybersecurity experts and firms. Mandiant, a well-known cybersecurity firm credited with conducting large-scale attribution and exposure of the Chinese

21 Eric G. Berman and Jonah Leff, "Anti-Materiel Rifles," *Small Arms Survey* (2011), No. 7, 1.
22 Mark Hubbs, "Where we began—the NIKE-ZEUS Program," *Space and Missile Defense Command /Army Strategic Command* (2007), 14.

PLA Unit 61398,[23] is one of these cybersecurity firms. According to Mandiant vice president Grady Summers, "We've seen first-hand that a sophisticated attacker can breach any network given enough time and determination."[24]

Of further note, the development of IBCCs views cyberdefense as a function of environmental variables, rather than focusing solely on outward-facing measures. Consider the role environmental factors have played in history's most significant conflicts. What if, during the Battle of Agincourt in the Hundred Years' War, the French had not been canalized by dense woodlands and slowed by thick mud?[25] It is possible that the numerically superior French Army would have won the battle and perhaps even have changed the entire course of the Hundred Years' War. What would have happened during World War II if the English Channel had not separated Nazi Germany from Great Britain? It is probable that the Nazis would have used Blitzkrieg tactics to overrun British defenses, thereby negating Britain's strategic bombing campaign and preventing execution of the Allied Forces' deception plan known as Operation Fortitude,[26] which led to Allies' successful invasion of Normandy in 1944.

Cyberspace, on the other hand, is not constrained by strictly defined environmental variables and is, rather, a function of human creation and ingenuity. In the cyber domain, one can fill the English Channel with elements of danger. In cyberspace, the trees can be made denser and the mud thicker. Cyberdefense professionals are limited only by their own creativity and level of ingenuity, implying that additional attention should be focused on cyberdefense as a function of the virtual environment.

Given this supposition, this paper contends that a successful active defense will be premised on the alteration of defensive environmental variables and must be designed to deter or defeat an adversary from within; that is, such a measure must retain deterrent/defensive capacity even after the network has been compromised. An effective IBCC will have specific qualities that achieve two key functions. First, it will not be reliant upon attribution yet it will deter malicious cyber actors by affecting their cost/benefit calculus in such a manner as to raise the minimum threshold for engagement in nefarious activities. Second, it will be designed to have a negative impact on those who levy cyberthreats, even after the network has been compromised. Let us now explore two hypothetical examples of the development of IBCCs and then discuss the cost/benefit structure, including who will bear the burden of implementing such a system.

A. Example 1: The use of a counter-data strategy by a government-affiliated, private-sector organization operating in a semi-permissive environment

For this scenario, let us consider an IBCC for a corporation within the defense industrial base whose primary business function is the development, design, production, delivery, and maintenance of military weapons systems. Real-world examples of such companies include

23 Why We Are Exposing APT1, "APT1: Exposing One of China's Cyber Espionage Units," *Mandiant* (2013), 2.

24 Mandiant Press Release, "Mandiant® Releases Annual Threat Report on Advanced Targeted Attacks," *Mandiant A FireEye Company*, 2013 https://www.mandiant.com/news/release/mandiant-releases-annual-threat-report-on-advanced-targeted-attacks1/ (accessed Feb. 1, 2014).

25 Juliet Barker, *Henry V and the Battle that made England Agincourt* (New York: Little, Brown and Company, 2005), Ch. 14.

26 Ernest S. Tavares, Jr., "Operation Fortitude: The Closed Loop D-Day Deception Plan," *Air Command and Staff College* (Maxwell Air Force Base, Alabama: GPO, 2001), 1.

Lockheed Martin, Boeing, and Raytheon. In this scenario, the corporation operates within the geopolitical context of a semi-permissive cyberspace environment; that is, private organizations are authorized to conduct reasonable active defense and response actions, but not to the extent that they are violating the U.S. equivalent of the Computer Fraud and Abuse Act.

First, we must depict this organization's AACC:

1. The Geopolitical Component: A semi-permissive environment where the conduct of active defense and limited response actions are within the boundaries of the law.

2. The Physical Infrastructure Component: A highly secure office environment that is unlikely to be physically penetrated by a malicious threat; both onsite physical infrastructure and communications equipment are highly fortified to include redundancy measures and well-protected hardware/server environments.

3. The Interface Component: Most/all members of this organization will likely possess uniquely identifiable network interface personas that differentiate members from others throughout the common population. For example, company president John Doe's email address may be john.doe@CompanyName.com, thereby differentiating him from a less attributable email address such as john.doe@gmail.com.

4. The Logical Network Component: Company network and routing protocol will be restricted from the public, and secure network routing protocol will be implemented. Organization members may have tokens that allow them to tunnel into the corporate network from home, which potentially makes the system vulnerable.

5. The Critical Information Component: This organization's lifeblood is the ability to design, produce, and distribute defense systems for sale to government militaries and private security companies. Therefore, this organization's most critical information component is the intellectual property pertaining to its design and production plans for defense systems.

According to the report, "Commission on the Theft of American Intellectual Property," annual losses due to theft of intellectual property are estimated to be over $300 billion.[27] This report states further that the sectors of the economy affected most prolifically tend to be those that support U.S. national defense programs.[28] Thus, for this situation, an appropriate IBCC is one that deters the theft of intellectual property and causes harm to adversaries who successfully infiltrate friendly networks and steal intellectual property. This of course begs the question, "How does one deter or cause harm against an adversary that they cannot conduct attribution against?" This is why cyberdefense professionals should develop IBCCs based on the premises of the AACC.

An appropriate IBCC for this scenario designed to defend intellectual property is the effective use of counter-data that is carefully seeded within a friendly network via a honeynet, a network

27 The National Bureau of Asian Research, "The IP Commission Report: The Report of the Commission on the Theft of American Intellectual Property," *The National Bureau of Asian Research* (Washington, DC: GPO, 2013), 2.

28 Ibid., p. 19.

of resources designed to be compromised.[29] "Counter-data" in this paper refers specifically to one of the following:

1. Custom-designed malware/spyware seeded within a honeynet that, if exfiltrated in an unauthorized manner (i.e., network intrusion), causes direct harm against an adversary by activating a call-back module to inform law enforcement of the adversary's location, wiping the adversary's system, or opening a backdoor into the adversary's system for response actions.

2. Intentionally flawed information seeded within a honeynet that causes indirect harm against an adversary by sowing confusion, misdirection, false intent, and deception.

While a counter-data strategy comprised of custom-designed malware/spyware would have universal application, a counter-data strategy with intentionally flawed information would vary according to the particular specialty of the corporation. A defense industrial base organization working with an intelligence agency, for example, should be defended by a counter-data IBCC containing false and misleading intelligence. Organizations involved with financial institutions should use honeynets that contain counter-data that is relevant yet disadvantageous to a competing financial institution. A weapon developer's counter-data IBCC should contain erroneous blueprints, unrealistic plans, or plans that suggest the pursuit of false strategic military objectives. By using this IBCC, the cyberdefender increases the competing organization's probability of taking a strategic misstep. Facilitating such a method allows the cyberdefender to seize the initiative from those who commit intellectual property infringement by fooling them into believing they have stolen something valuable.

The IBCC described above complements the AACC, as it does not require attribution in order to induce damage against adversaries. By accurately characterizing the five components of the AACC, this countermeasure essentially defends an intellectual property oriented organization in an automated manner. It operates within the geopolitical constraints by conducting automated response actions against adversaries without going as far as to take offensive and autonomous action against intruding networks. It will possess the necessary physical infrastructure and interface components designed to make the honeynet appear as realistic as possible to the potential adversary. Similarly to government intelligence agencies' use of counterintelligence agents, intellectual property oriented organizations should employ counter-data agents in order to deploy and maintain this program. Lastly, the solution will have a logical design (believable IP addresses, appropriately routed networks, etc.) used in such a manner as to fool or at least sufficiently confuse an intruder to the point to where they are either unaware or unsure if they are obtaining intellectual property of value.

B. Example 2: The use of a "white noise" strategy by a private-sector retailer operating in a restrictive environment

For this scenario, let us consider an IBCC for a department store within the commercial retail sector, whose primary business function is the sale of tangible goods such as clothing, food, appliances, electronics, furniture, etc. Well-known real-world examples of such companies include Wal-Mart, Target, McDonald's, and Best Buy; however, we should also consider

29 Matt Walker, *All-In-One Certified Ethical Hacker* (New York: McGraw Hill, 2012), 352.

small “mom-and-pop” type stores. In this scenario, the retailer operates within the geopolitical context of a restrictive cyberspace environment; that is, private organizations are authorized to conduct active defense but not active response actions, nor any activity that would intrude on another network.

The following is this retailer’s AACC:

1. The Geopolitical Component: A restrictive environment where active defense is authorized; however, direct response actions are outside the boundaries of law.

2. The Physical Infrastructure Component: An open retail environment designed to facilitate customer service; because priority is given to the sale of retail goods, infrastructure security is not highly prioritized; communications infrastructure is primarily designed to conduct POS transactions.

3. The Interface Component: Likely only upper management will have uniquely identifiable email addresses; lower level employees (sales clerks, warehouse workers, etc.) will likely interface instead with POS machines or personal computers.

4. The Logical Network Component: In the modern era, POS machines may be connected via Wi-Fi, be cloud-based, or be centrally administered in some way or another. POS machines will likely transfer data to a back-office computer or central data-processing point for the purposes of accounting, inventory control, estimating sales trends, etc. IP address data and Internet connectivity will likely be minimally secured.

5. The Critical Information Component: The financial well-being of retailers is based on their ability to purchase goods at wholesale and sell them at a mark-up value in order to turn a profit. Therefore, a retail organization’s most critical information component is the financial transaction system that allows them to sell goods to customers and centrally manage data pertaining to POS transactions.

Recent news headlines demonstrate retail POS systems’ increased vulnerability to credit card data breach and fraud. According to LexisNexis Risk Solutions, a research-oriented firm, retail merchants paid on average 2.69 cents per dollar in 2012 and 2.79 cents per dollar in 2013 as a result of increased fraudulent use of credit cards via online transactions.[30] In addition to the rising costs of credit card fraud, research suggests that data breaches that lead to credit card fraud are increasing at an alarming rate. According to a Verizon study, over 2,500 large-scale data breaches have occurred over the nine-year period between 2004 and 2013, with 621 of those breaches occurring between 2012 and 2013, for a total of 1.1 billion compromised records.[31] In 2012, approximately 1 in 4 of these data-breach victims suffered identity theft.[32] Online vendors, who suffer the bulk of fraudulent transactions, have implemented a host of fraud-detection technologies, including IP geolocation, device fingerprinting, verification

30 LexisNexis, “2013 LexisNexis True Cost of Fraud Study,” *LexisNexis Risk Solutions* (Dayton, OH: LexisNexis, 2013), 6.
31 Verizon Risk Team, “2013 Data Breach Investigations Report,” *Verizon* (New York: Verizon, 2013), 4.
32 Ibid. 28, p. 6.

services, browser/malware tracking, rule-based filters, etc.,[33] yet these measures do not address the core problem: how do we effectively limit the breach of data in the first place?

While the online retail industry has managed to implement security measures with varied degrees of success, this does not solve the problem of data breaches; rather, it merely counters a malicious person's capacity to use fraudulent personal data to conduct online transactions. Department stores, restaurants, mom-and-pop shops, and retail stores throughout the world remain vulnerable to data breaches, due to their technical inability or lack of sufficient funds to apply high-level cybersecurity measures. Even if retail stores managed to encrypt data at POS locations, this does not change the fact that a persistent actor who is sufficiently determined can and will intercept personally identifiable information and find ways to crack the encryption. It stands to reason, then, that cyberdefense professionals must seek to drastically alter the threat environment.

Many cybertheorists have conceptualized cyberspace as a sort of environment or terrain that is governed by the laws of physics, including both its logical and physical aspects.[34] Unlike other environments, such as the land, sea, air, and space, the cyberspace environment can easily and quickly be altered by human will. Whereas a ship traveling through a narrow passage or canal is restricted to that particular body of water, human interface via the cyber domain is capable of creating new passages (links and nodes) and new ships (packets of data) at an extremely rapid rate. Given this concept, an appropriate IBCC for the defense of retail POS systems may be the introduction of "white noise" into friendly cyberspace environments.

Consider the breach that took place at Target stores in November-December 2013. Essentially, a group of individuals managed to breach Target's primary information hub, and then distributed code to POS systems and cash registers that allowed them to capture credit card data from customers.[35] Now consider the development of IBCC software that would make it so that, for every legitimate transaction that took place, the software would simultaneously fabricate 1,000 additional transactions. The aim would be that the POS system itself would be unable to differentiate between the legitimate transaction and the fabricated transactions. Each fabricated transaction would be controlled via a random data generator that combined varying sequences of the following:

1. A 16-digit credit card number
   - 9,999,999,999,999,999 possible outcomes
2. A randomly assembled combination of first name, last name, and middle initial
   - Approximately 20,360,011,698 possible outcomes[36,37]
3. An expiration date within the next four years
   - 48 possible outcomes

33 LexisNexis, "2013 LexisNexis True Cost of Fraud Study," *LexisNexis Risk Solutions* (Dayton, OH: LexisNexis, 2013), 30.

34 Gregory Rattray, *Cyberpower and National Security* (Washington, DC: Potomac Books, 2009), 255.

35 Bree Fowler, "Answers to questions about Target data breach," *The Boston Globe*, 2013 http://www.bostonglobe.com/business/2013/12/19/answers-questions-about-target-data-breach/pN7ikzJzFWYhHtsFXHISeL/story.html (accessed Feb. 7, 2014).

36 According to the U.S. Census Bureau, in the year 2000 there were 151,671 unique last names and 5,163 unique first names.

37 U.S. Census Bureau, "Genealogy Data: Frequently Occurring Surnames from Census 2000," *U.S. Census Bureau*, 2014 http://www.census.gov/genealogy/www/data/2000surnames/ (accessed Feb. 7, 2014).

4. A credit card company randomly selected from American Express, Visa, MasterCard, and Discover
    - four possible outcomes
5. A three-digit security code
    - 999 possible outcomes

When all the above factors are considered, there are approximately 3.905e+31 different possible outcomes—an astronomical figure, which implies that the probability of accidentally duplicating a real credit card is virtually zero. All transactions (both real and fabricated) would be transmitted via encrypted channels to a highly secure central processing location. The central processing entity would then cross-reference all transactions with MasterCard, American Express, Visa, and Discover databases in order to process the transactions appropriately. Real transactions would be processed as normal, and fabricated transactions would be sent to and stored in a centralized cybersecurity company database. This storage database would hold on to these fabricated transactions for a predetermined period of time. If, at some point or another, an identity thief attempted to use one of these fabricated credit cards to conduct illegitimate transactions, it would automatically be flagged in the storage database and would cue law enforcement authorities to the location of the transaction or, ideally, the location of the criminals themselves.

C. Costs, Benefits, and Bearing the Burden

The implementation of IBCCs requires expending resources on secondary defense efforts. In addition to maintaining current outward-facing cybersecurity efforts, IBCCs require the allocation of potentially substantial resources to conduct defense and deterrence from within the network. The amount of resources allocated for this effort will be situationally dependent. For example, it would behoove a major firm whose main asset is intellectual property to bear the burden of implementing an IBCC by hiring one or more full-time counter-data strategists to manage their deception program. This individual would be required to have both cybersecurity and traditional counterintelligence-like traits, which suggests that firms will be required to pay a premium for both skillsets. Firms employing the white-noise IBCC, on the other hand, would likely bear the burden of implementing an IBCC by paying a premium on installing and maintaining the defense mechanism, as opposed to paying the salary of a full-time individual. Large computer security firms such as McAfee, Kaspersky, Symantec, and others are capable of implementing such an IBCC today, given currently available technology. Major firms, like Target, would likely be more than willing to bear such costs, whereas small companies would be able to band together to share the maintaining an IBCC. Additional cost-sharing structures could include customers, business partners (such as credit companies), and, potentially, national governments who are responsible for shouldering the costs of national security.

Because the benefits to be gained from implementing IBCCs are not always realized by a private firm directly, there would be a role for national governments to adjust the load-sharing appropriately. However, considering the magnitude of loss that companies regularly face due to data breaches and intellectual property theft, firms that successfully implement IBCCs may be able to limit their losses due to fraudulent activity and enjoy the benefits of long-term loss reduction, in terms of their liability due to identify theft, their reduced losses from intellectual property theft, and lower cost of customer/product remediation measures.

5. CONCLUSION

This paper outlines the need for cyberdefenders to construct frameworks that proactively define an organization's characteristics and conduct environmentally oriented cyberdefense measures. By acknowledging the asymmetries between actors and their capabilities and intent within the cyber domain, cyberdefenders can free themselves from the biases that security professionals have developed as a result of operating within a conventional threat environment. The Internet's history and current events demonstrate that cyberspace yields asymmetric advantages to those who leverage intrusive capabilities. This paper therefore surmises that network defenders must secure friendly networks by using attribution agnostic cyberdefense constructs and designing internally based cyberthreat countermeasures that take advantage of network environmental variables in order to deter and defeat nefarious cyber actors.

The Internet was initially designed to be a collaborative domain characterized by the free sharing of ideas. Unfortunately, the lack of security mechanisms implemented within the initial design has created opportunities for malicious individuals to exploit other people. The framework proposed in this paper, while by no means a comprehensive solution, represents the aggressive mindset that cyberdefenders must develop if they want to combat threats in cyberspace. Like the creation of countermeasures in the physical domain, it is not merely suggested but imperative that network defenders shift to an aggressive mindset and apply energy and resources to create IBCCs within friendly network domains.

Chapter 3
Cyber Situational Awareness

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Dynamic Cyber-Incident Response

Kevin Mepham
PhD Research Student,
Defence and Cyber Security Research Group,
Brunel University
London, UK
kevin mepham@brunel.ac.uk

Panos Louvieris
Defence and Cyber Security Research Group,
Brunel University
London, UK
panos.louvieris@brunel.ac.uk

Gheorghita Ghinea
Defence and Cyber Security Research Group,
Brunel University
London, UK
george.ghinea@brunel.ac.uk

Natalie Clewley
Defence and Cyber Security Research Group,
Brunel University
London, UK
natalie.clewley@brunel.ac.uk

Abstract: Traditional cyber-incident response models have not changed significantly since the early days of the Computer Incident Response with even the most recent incident response life cycle model advocated by the US National Institute of Standards and Technology (Cichonski, Millar, Grance, & Scarfone, 2012) bearing a striking resemblance to the models proposed by early leaders in the field e.g. Carnegie-Mellon University (West-Brown, et al., 2003) and the SANS Institute (Northcutt, 2003). Whilst serving the purpose of producing coherent and effective response plans, these models appear to be created from the perspectives of Computer Security professionals with no referenced academic grounding. They attempt to defend against, halt and recover from a cyber-attack as quickly as possible. However, other actors inside an organisation may have priorities which conflict with these traditional approaches and may ultimately better serve the longer-term goals and objectives of an organisation.

Shortcomings of traditional approaches in cyber-incident response and ideas for a more dynamic approach are discussed including balancing the requirements to defend against an incident with those of gaining more intelligence about an attack or those behind it. To support this, factors are described which have been identified as being relevant to cyber-incident response. These factors were derived from a literature review comprising material from academic and best-practice sources in the computer security, intelligence and command and control fields.

Results of a PhD research survey conducted across military, government and commercial organisations are discussed; this assesses the importance of the aforementioned factors. The surveyed participants include (but were not limited to) respondents from areas such as Intelligence and Operations, as well as the more conventional computer security areas.

Situational awareness and decision-making aspects of incident response are examined as well as other factors such as intelligence value, intelligence gathering, asset value, collaboration and Intelligence Cycle factors.

Keywords: *Cyber Incident Response Active Passive Risk*

1. INTRODUCTION

In recent decades technology has changed rapidly, especially in the Information Technology (IT) area; in a drive for efficiency and cost-saving organisations and governments have become increasingly-dependent upon IT and its supporting infrastructure. In recent years this transformation has also led to an increasing dependence upon the Internet by critical or important infrastructure. However, the other side of the coin is that this evolution has led to an increased exposure to exploitation or compromise by those with hostile intent as traditionally closed networks or systems have become more accessible. Despite this rapidly-evolving environment and associated risks, to all intents and purposes standard computer security incident response models, have remained largely unchanged since the 1990s. Furthermore, much of the research which contributed to the production or revision of these models has been called into question. In a review of 90 works which claimed to employ quantified investigation and analysis of security, it was discovered that the validity of the majority of these works was questionable when used in the perspective of an operational setting (Verendel, 2009).

This research investigates factors which may influence Cyber-Incident Response from the perspective of a wider-affected audience in order to produce a more dynamic and stakeholder-independent Cyber Incident Response model. It attempts to do this by taking into account the strategic and wider priorities of an organisation and also considers intelligence gathering and sharing priorities as part of incident response. Although not yet at an experimental stage in the research, evaluation of the identified factors by international communities from within and outside the core Cyber-Security areas have already confirmed the requirement for changes to the current models. This has been deduced from both discussion and by the statistical analysis of their responses collected as part of a research survey discussed in this paper.

2. RELATED WORK

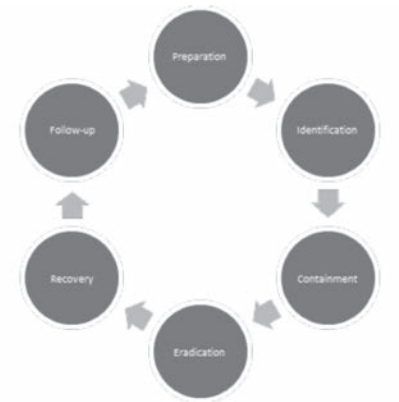
As part of the research, a cross-domain literature review was carried out; this covered not only the core CIS/Cyber Security field but also areas such as Military Intelligence, Command and Control (C2) and Human Factors issues. The aim of this review was to identify significant independent variables defining the problem domain of Cyber Incident Response including parallels from other domains outside of the Cyber Security field. In parallel to the literature review, participation in Multi-National Experiment 7 (MNE7), an experiment intended to capture the important factors related to preservation of access to the Global Commons (air, sea, space and cyber), led to the identification of factors deemed to influence the effectiveness of Cyber Situational Awareness; a key component of effective Cyber Incident Response.

A. Literature Review

The literature review was approached from two perspectives. The first was a practitioner's perspective looking at the best-practice documents from Cyber Security and associated fields. The second was the academic perspective where research was already busy identifying gaps and shortcomings within the field. Both of these perspectives were then drawn together to identify a consolidated list of the existing factors influencing Cyber Incident Response as well as missing factors which could be utilized in future models. These perspectives and factors are described in the subsequent paragraphs.

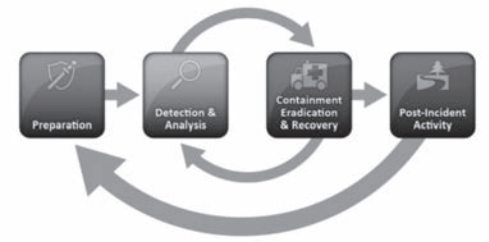
Traditional cyber incident response, even from the early days of widespread computer use, tended to take an approach of detecting an incident and then trying to halt, contain or mitigate it followed by a recovery phase to restore normal operation. Post-incident analysis was then used to identify potential improvements to the infrastructure and processes (if necessary). This approach is best illustrated utilising the SANS Institute Model (Northcutt, 2003) which added more detail to the cycle in 2003 (Figure 1).

FIGURE 1 - SANS INSTITUTE INCIDENT RESPONSE CYCLE 2003 (NORTHCUTT, 2003)



Although some evolution has taken place, even the most recent iterations of the best-practice processes still broadly cover the same issues, for example the latest guidance (Cichonski, Millar, Grance, & Scarfone, 2012) published by NIST (Figure 2), establishes the incident response process as an inner circle with "lessons learned" (post-incident activity) providing the feedback to improve the infrastructure and processes (preparation).

FIGURE 2 - NIST SPECIAL PUBLICATION 800-61 INCIDENT HANDLING PROCESS (CICHONSKI, MILLAR, GRANCE, & SCARFONE, 2012)



This perspective is also echoed in international standards, for example the international Information Security Management standard ISO27001 advocates the Deming Cycle (Calder & Watkins, 2008). This standard advises that Information Security (and consequently Cyber-Security) can be divided into the phases of Plan, Do, Check and Act. Within the live incident response environment this is reduced to the "Do", deploy the sensors and implement planned measures; "Check", look for incidents by monitoring the information sources that have been deployed; Act, respond to detected incidents or identified shortcomings. Outside of this shortened cycle the planning takes place to improve the longer term protection of the information and infrastructure. However, all of these cycles are based around the core tenets of preserving the Confidentiality, Integrity and Availability of these protected assets. Whilst understandable from a Cyber Defence perspective, there are also other communities impacted by Cyber Incidents.

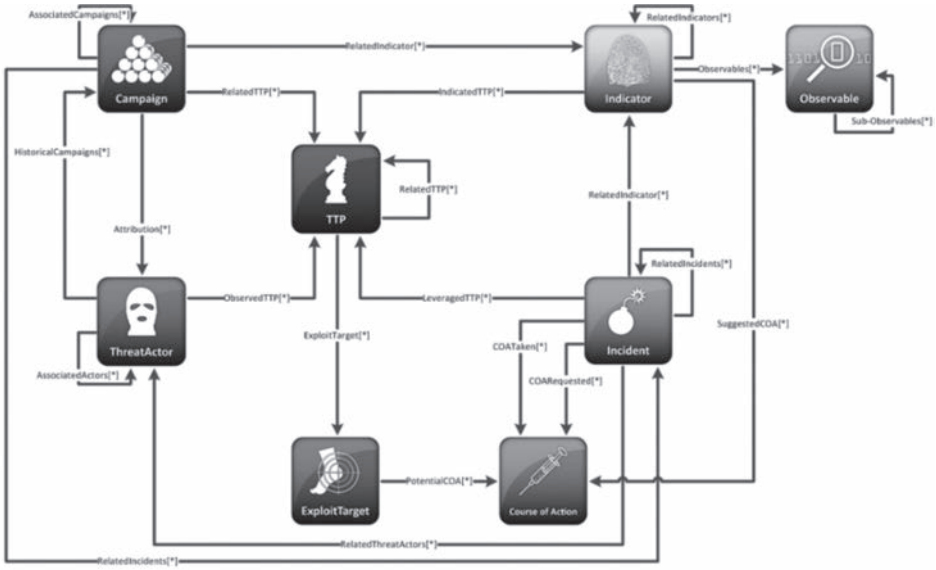
Looking at cyber incidents from a Military/Business Intelligence perspective, the Intelligence Cycle lens can be applied. The Intelligence Cycle (MoD, UK, 2011) has some similarities to the traditional Incident Response cycles (commonly having the phases Planning and Direction, Collection, Analysis or Processing and Dissemination), however, there are also some contrasts. Intelligence work by its nature is designed to gather information about potential adversaries as well as understanding this in the context of own and partner capabilities and objectives; as Sun Tzu (Tzu, 2011) is reputed to have stated "know the enemy and know yourself, in a hundred battles you will never be in peril". This emphasis on "knowledge of the enemy" puts the Intelligence community at odds with the Cyber-Defence community as Intelligence gathering is not a natural partner of preserving Confidentiality. However, this is not an insurmountable problem providing that the priorities can be put in context as will be discussed later.

In the UK, joint doctrine (MoD, UK, 2011) talks about "Inform", which is defined as "the ability to collect, analyse, manage and exploit information and intelligence to enable information and decision superiority" i.e. this equates to the "Disseminate" of the Intelligence Cycle. In traditional Cyber-Incident Response the collection and analysis is only traditionally carried out up to the point where the incident is thwarted and in the post-incident analysis; at this point the incident has been resolved or averted and there is nothing more to gain in terms of intelligence value (or to disseminate in order to improve infrastructure or intelligence). Combined with the increasing difficulty of maintaining a credible honeynet or honeypot solution (Rowe, 2006); (Wang, Wu, Cunningham, & Zou, 2010) where information has traditionally been gathered to provide Cyber intelligence, this leads to the danger of information starvation for those trying to assess some of the key Cyber Intelligence requirements such as attacker identity, motivation, ultimate target, attack methods, attacking resources, attack goal. The lack of this type of intelligence (especially for novel attacks or unknown attackers) will undoubtedly lead to a reduced ability to defend in the longer term.

With reference to Situational Awareness, this requirement for Cyber-Intelligence is indirectly reinforced by Endsley's model (Endsley, 1995); in this model "Long term memory stores" are seen to inform "expectations". In turn expectations inform the three identified stages of situational awareness: perception, comprehension and projection. This approach infers that without the information (or intelligence) in the long term memory stores the expectations will not be optimally informed, thereby depriving the decision maker of the best situational awareness. This introduces the concept of not only utilising static intelligence but also using this to predict future events to enhance decision-making.

Taking this prediction thread further, as early as 2000, the importance of usable intelligence in a cyber-environment was recognised (Yuill, et al., 2000). In this research a military intelligence type process to enhance the effectiveness of intrusion detection and the subsequent incident response was proposed. At that time, prior to the introduction of the SEI State of the Practice process (Killcrece, Kossakowski, Ruefle, & Zajicek, 2003), Yuill et al considered standard incident-response process to be attack repair, neutralization and containment (ARNC). However, by providing positive identification of the attacker (using part of a proposed technique referred to as Cyber-Intelligence Preparation of the Battlespace (C-IPB)), likely compromised devices (LCDs) could also be identified based on models of the attacker and the infrastructure. This information could then be used to produce two types of estimate for Courses of Action (COA) by the attacker: possible and likely i.e. the notion of predicting cyber-incident progress was proposed. From these estimates, further monitoring could be more targeted and incident-response measures more relevant. The C-IPB process is summarised in four steps: define the battlespace (define the boundaries of the infrastructure), describe the battlespace effects (evaluate the infrastructure and its influence on attack and defence), evaluate the threat (assess attacker capabilities and intent) and determine the threat's COA and infrastructure LCDs. At that time, the cyber-intelligence was broken down into: what the attacker has done (executed action), capabilities, personal traits and intentions. However, whilst the principles remain sound there has been significant development in the types of information that are relevant to capturing threats and attacks such as those described in the Structured Threat Information eXpression (STIX) community-driven standard (Barnum, 2012) maintained by MITRE Corporation. This standard is directly related to another standard maintained by MITRE Corporation, Trusted Automated Exchange of Indicator Information (TAXII) which is designed to allow collaboration between Cyber-entities to exchange threat intelligence.

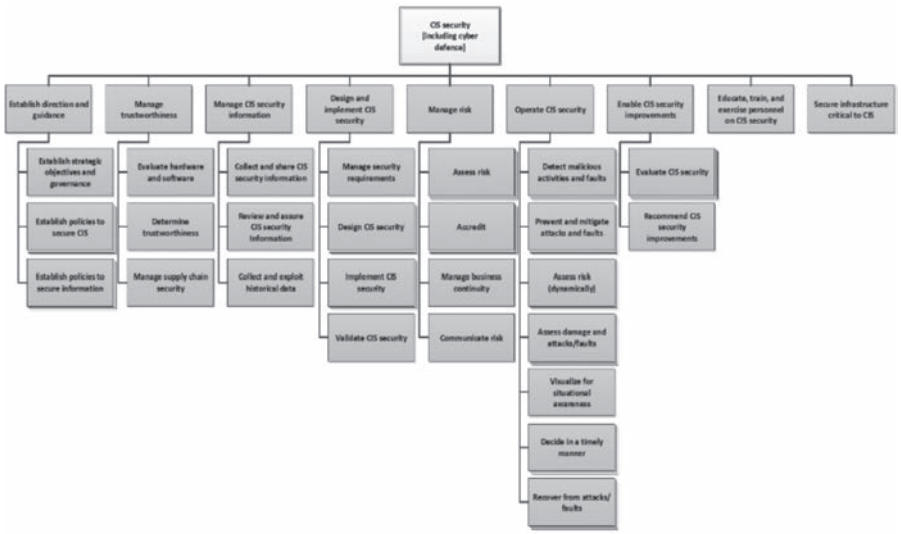
FIGURE 3 - MITRE CORPORATION STRUCTURED THREAT INFORMATION EXPRESSION (STIX) (BARNUM, 2012)



STIX, provides identification of each of the information components illustrated in Figure 3 by a number of variables. Utilising these it attempts to achieve the following four use case goals: analyse cyber threats; specify indicator patterns for cyber threats; manage cyber response threat activities and the sharing of cyber-threat information.

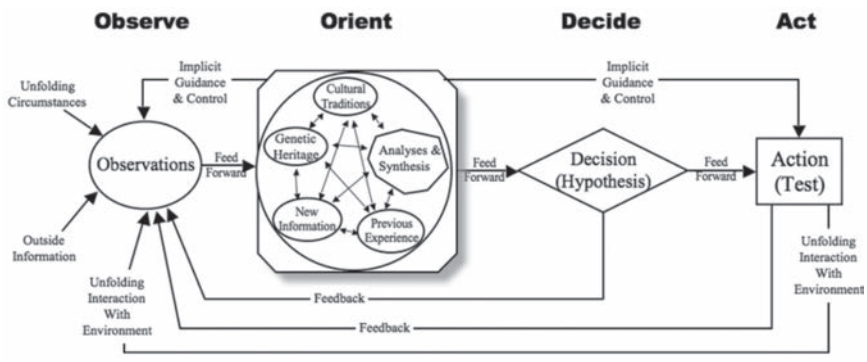
The combination of several elements from the approaches in the previous paragraphs can also be found in a NATO framework document (Hallingstad & Dandurand, 2011) this document (produced with cooperation from several NATO member nations participating in a NATO-led research task group) is summarised in a top-level diagram (Figure 4) which also includes the incident-response processes. This framework was broad enough to cover areas of interest, not only to the Cyber-Defence community but also for senior decision makers and Intelligence community. Whilst explaining the more obvious issues of making sure that the appropriate sensors and trained personnel are in place to allow incidents to be detected, it also covered areas such as ensuring that risks are owned and managed and that trustworthiness of hardware, personnel and partners is addressed. Interestingly, the quandary of whether to stop interesting attacks or to monitor them to gain intelligence is also discussed briefly within the document. Information sharing with regard to CIS security incidents is also identified as a relevant issue in this framework; the importance of this is confirmed by the international work that has taken place in recent years such as Multi-National Experiment 7 – Access to the Global Commons (MNE7), and continues to take place at the moment in the Multinational Capability Development Campaign (MCDC) Cyber Implications for Combined Operational Access (CICOA) 2013-2014.

FIGURE 4 - NC3A CIS SECURITY FRAMEWORK



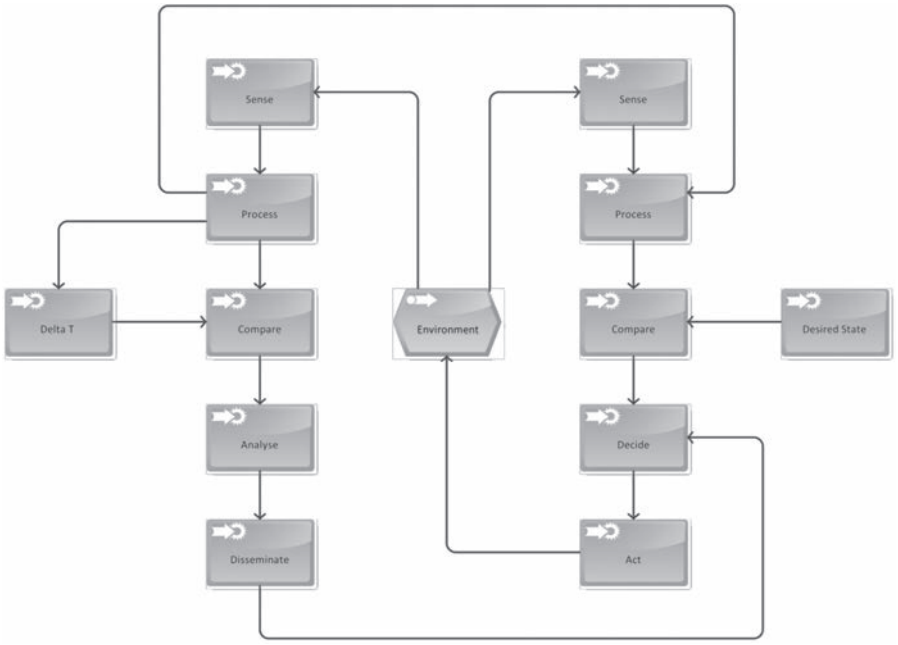
Ultimately, whichever response is chosen to a cyber-incident by the empowered decision-maker, it has to be timely enough to be able to influence the outcome. This is best summarised by the model proposed by Colonel John Boyd USAF (Orr, 1983). The model of Observe (monitor the enemy's actions), Orient (work out possible actions and consequences based on the observations of the enemy and knowledge of your own capabilities), Decide (choose a course of action), Act (carry it out), otherwise known as the OODA loop was designed to describe how to gain superiority in air combat. By completing an OODA loop more quickly than an adversary, the adversary would not be able to react in time to gain air superiority. In Figure 5, this is shown as not only a single uni-directional loop (as illustrated by several interpretations of the model), but also a series of inner feedback loops which influence the observation and consequently orientation, decision-making and subsequent action. Although originally intended to reflect air combat, it has since been recognised that this has wider application for strategy in both military and commercial contexts. This is also pertinent in the context of Cyber-Incident response where, for the advanced attacker, they are often able to respond quickly to any mitigation or actions carried out by the defender. If this response is achieved inside the defending OODA loop they then gain "cyber superiority".

FIGURE 5 - COLONEL JOHN BOYD USAF'S OODA LOOP (ORR, 1983)



A further development of the OODA loop was proposed to describe a Command, Control, Communication and Intelligence (C3I) model (Figure 6) which explicitly includes a simulation/ prediction function (Lawson, 1980).

FIGURE 6 - C3I PROCESS MODEL (LAWSON, 1980)



In this model, the Intelligence aspect can be seen on the left hand side of the model (with Delta T representing a time difference) and the Command and Control (C2) aspect on the right (the communication would be in the sensing and dissemination). Effectively, this creates two unidirectional OODA loops, one for Intelligence and one for C2 (although the right-hand side could also be representative of the conventional incident-response cycle. In the right-hand side, 'sense' equates to 'observe'; 'process' and 'compare' equate to 'orient(ate)' the current situation compared to the desired situation; 'decide' and 'act' then influence the environment which is then reassessed. In the left-hand loop (which feeds into the decision-making process of the right-hand loop), analysis is carried out with respect to time which allows some prediction of the direction of the environment; this is then fed into the decision-making to allow more informed actions to be taken rather than relying upon a static snapshot of the environment.

However, in the context of cyber-incident response, the "Desired State" could be replaced with "normal" state to reflect normal infrastructure operation whilst the left-hand side assesses whether the environment is moving away from or towards this state over time. This is a good demonstration of situational awareness; if used in a military decision-making process, the sensors would provide Intelligence information (rather than data) which is then used with expert knowledge or systems to provide a prediction of the future infrastructure state based on monitored behaviour over time.

Ultimately, the literature review confirmed that Cyber-Intelligence is an essential aspect of Cyber-Incident response; modelling of cyber-incidents to provide prediction/projection of the future path of an incident is also important in providing optimal situational awareness and

that different stakeholders impacted by a cyber-incident can have a different perception of the priorities which may not be aligned with organisational goals. When combining these findings with established models from other areas such as the Command and Control and Intelligence areas it can be surmised that further evolution of Cyber Incident Response is necessary to best serve organisational aims.

B. Contribution of MNE7 to this Research

As previously mentioned, the MNE7 Campaign was conducted at the same time that the literature review was carried out. This experiment brought together a rare collection of professionals from governmental, military, commercial and academic areas from both inside and outside the core cyber security areas. Participation in the collaborative cyber-situational awareness track allowed the opinions of an expert community to be gauged and the same community also provided significant feedback on the pilot questionnaire, where the water was being tested with regard to potential gaps in the existing Cyber Incident Response models and processes. However, one of the strongest messages to come across from this community is that everybody can see the benefits of collaborating by sharing incident information, but in practice they are reluctant to do it. Despite this, given trustworthy filtering of information and a mechanism to establish sufficient trust between partners, collaboration can prove invaluable in enhancing situational awareness. In the context of this research, information received from collaboration is viewed as one of many information sources.

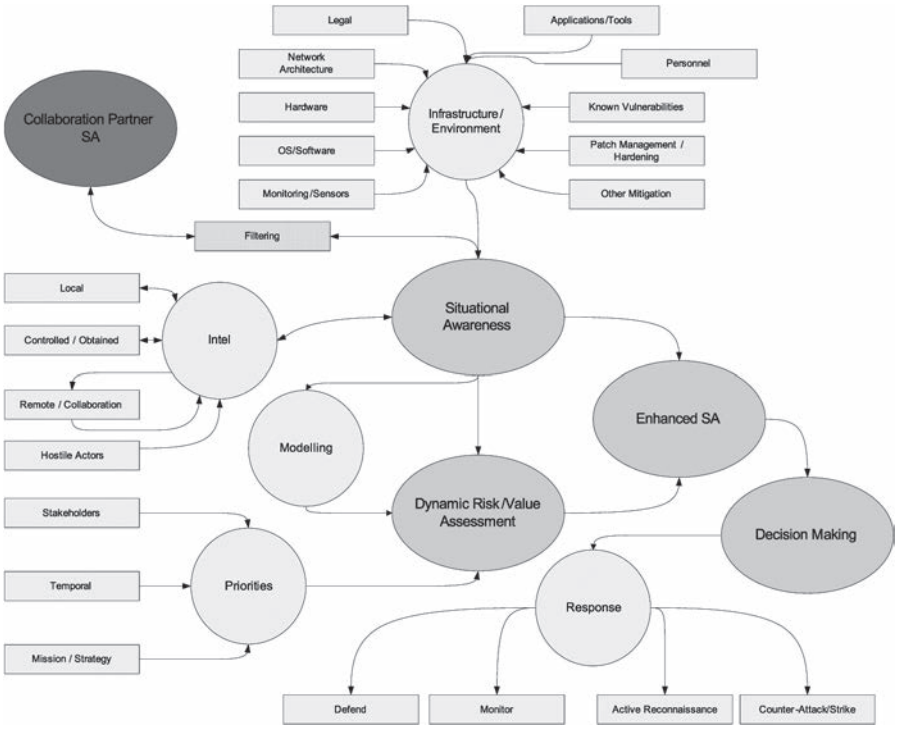
3. METHODOLOGY

A limited pilot survey was carried out with participants from international military, commercial and governmental cyber security communities to evaluate the initially identified variables from the communities and the literature review. Utilising principal component analysis and Varimax rotation (described in more detail later) an initial attempt was made to group some of the identified factors. Whilst not strictly observing the identified grouping, as the results were not statistically significant at that time (due to the sample size) this provided a suitable discussion point within these communities to sharpen the areas of focus for the remaining portion of the literature review and subsequent surveys. However, this focusing of the initial evaluation of these variables, discussions within expert communities and the remainder of the initial literature review led to the production of an initial model which has also been used as a starting point to describe the contribution of cyber to the operational planning process by the technical strand of MCDC-CICOA.

This initial model shown in Figure 7 (which combines process, functions and infrastructure) attempted to describe the interaction between infrastructure and what is described here as static situational awareness i.e. the impact of an incident on the defending environment as it is now, utilising the existing intelligence. This static situational awareness is then used as an input to dynamic risk and value assessment, where, based on the current known situation, modelling of an attack is attempted. This utilises the known vulnerabilities and paths through the infrastructure with the available attack intelligence which is then combined with the assessments by the different stakeholders for that point in time of the value of the threatened assets (recognising that different stakeholders may well place different priorities on the same

asset). The output of this process would be “balance of equities” information to be provided to the key decision maker together with the static situational awareness in order to provide them with enhanced situational awareness. This information would allow them to choose the optimum response in order to meet the organisational goals; examples of these described by the response options (without reference to legal constraints) are to defend the attacked assets via passive means, gather more intelligence about an attack or attacker (via passive means) or use active means to pacify attacker infrastructure or gather more intelligence about the attacker. Referring back to the OODA loop, this whole process needs to be completed before the attacker has a chance to detect and respond to any actions taken by the defenders in order to gain an advantage over the attacker.

FIGURE 7 - INITIAL MODEL



Utilising this initial model and the literature review as a starting point, a new large-scale survey was produced to evaluate the importance of identified factors in providing effective Cyber-Incident Response; this not only included respondents from the Cyber-Security communities, but also other communities involved with and impacted by cyber-incidents such as Military/ Business Intelligence, Operations, Communications Information Systems Management and other support areas. The questions assessed not only the opinions of the participants as to the importance of the identified factors affecting cyber incident response but also how these factors were viewed in their communities and organisations. The survey was conducted using a

7-point Likert scale for each of the assessed variables in order to achieve an appropriate degree of granularity in the results; to date, a combined total of 186 professionals from the identified communities have participated in the survey.

4. RESULTS AND ANALYSIS

From the results to date, there has been a striking difference in opinion between individuals in all communities and their perception of their organisations' opinions. This assessment was confirmed by paired t-tests where all 30 variables were found to have significant results. From the results it appears that individuals across the communities tend to place more importance on the identified factors than their organisations or communities. A good example of this can be seen in the response to Configuration Management (CM) where almost half of the participants assessed that effective CM was essential to provide optimal Cyber-Incident Response (Figure 8) whereas in their communities and organisations just over 10% of the participants (Figure 9) believed that their communities and organisations found CM to be essential. Other notable examples of this phenomenon were reflected in the use of automatic tools for intelligent data reduction, sensors for monitoring at all levels, timeliness and reliability of data and to a lesser extent areas such as environmental conditions that analysts work in.

FIGURE 8 – CONFIGURATION MANAGEMENT: INDIVIDUAL RESPONSE

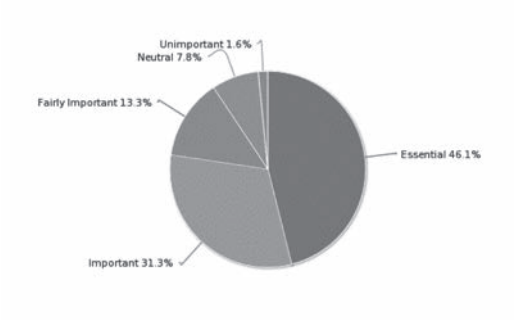
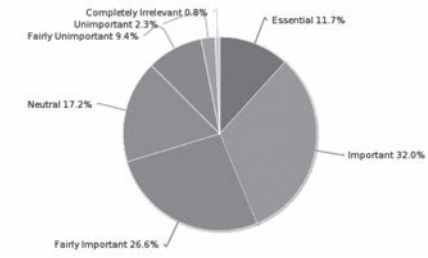


FIGURE 9 – CONFIGURATION MANAGEMENT: ORGANISATION RESPONSE



As expected, there are also significant differences in the importance placed on assigning a value to intelligence regarding the attackers and attacks between the communities. This

is demonstrated below in the contrasting opinions on the importance of placing a value on Intelligence information as part of the Cyber-Incident response process (Figures 10 and 11).

FIGURE 10 - IMPORTANCE OF INTELLIGENCE VALUE: INTELLIGENCE PROFESSIONALS

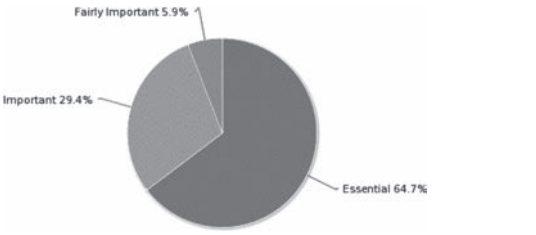
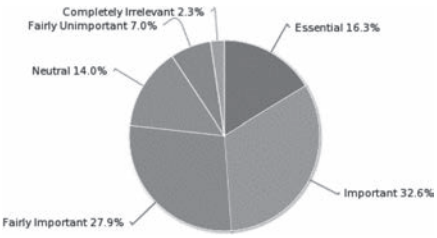


FIGURE 11 - IMPORTANCE OF INTELLIGENCE VALUE: IA/SECURITY PROFESSIONALS



However, some unexpected differences of opinion were also identified across the communities, even relating to the importance of stakeholders being able to assess the value of assets from different perspectives (Figures 12 to 15). In this example, it might be assumed that the CIS/ Engineering communities believe that they already know the priority of the assets that they maintain so it is not essential to have the functional owner's perspective.

FIGURE 12 - IMPORTANCE OF ASSESSING STAKEHOLDER VALUES: IA/SECURITY COMMUNITY

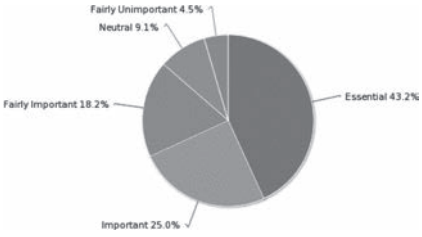


FIGURE 13 - IMPORTANCE OF ASSESSING STAKEHOLDER VALUES: OPERATIONS COMMUNITY

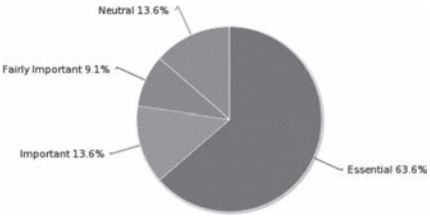


FIGURE 14 - IMPORTANCE OF ASSESSING STAKEHOLDER VALUES: IT/ENGINEERING COMMUNITY

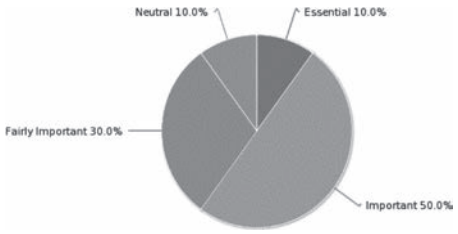
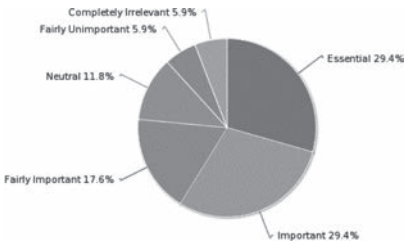


FIGURE 15 - IMPORTANCE OF ASSESSING STAKEHOLDER VALUES: INTELLIGENCE COMMUNITY



However, when the survey was initially produced, a set of 30 variables were identified which might be considered important to Cyber Incident Response and as can be seen from the draft model, this creates an almost unmanageable model from a conceptual point of view. In order to simplify this, a series of statistical processes were run to try and reduce the number of variables (i.e. to check for significant correlation between similar factors in order to merge them as a single variable) and these are summarised in the subsequent tables. Not only does this allow simplification of the model but also makes experimentation more realistic (as too many variables will make it almost impossible to test all inter-relationships and assess their significance on the measured output variables).

For the first time (as far as can be determined) factor analysis was carried out to determine key areas of importance in the cyber incident response process. This was achieved by analysing the results obtained from the communities of interest (from the survey) using principal axis factoring and Varimax[1] rotation. This dimension reduction process allows correlated variables to be grouped into common components or factors and those which are orthogonal to them are grouped into separate factors. From the sample size, it is suggested (Hair, Black, Babin, & Anderson, 2014) that a factor loading of more than 0.50 be used in order to achieve power level of 80%. Utilising this process (using the SPSS software package), the following factors were identified from the data sources:

i) Sensors (monitoring of operating system logs, network sensor logs, application logs etc).
ii) Collaboration (both inbound and outbound SA collaboration with trusted partners).
iii) Information Credibility (accuracy, timeliness and reliability of information).

1 Created by Henry F Kaiser in 1958

iv) Incident Discrimination (analyst experience and automated tools to reduce the "noise" of routine events).

TABLE 1- PRINCIPAL COMPONENT ANALYSIS OF INTELLIGENCE SOURCES

	Component			
	Sensors	**Collaboration**	**Credibility**	**Discrimination**
OS Monitoring	.85			
App Monitoring	.72			
Hardware Mon	.71			
Network Mon	.69			
Collaboration In		.87		
Collaboration Out		.83		
Accuracy			.75	
Timeliness			.73	
Reliability			.50	
Automated Tools				.80
Analyst Experience				.73

These variables were then grouped together to create a process that for the purposes of the model will be called Intelligence Gathering. Utilising a series of similar reductions using the same Varimax process, the rest of the variables were grouped together to create a number of functions to form the basis for a new model. These processes then become:

i) Intelligence Gathering: the gathering of information from relevant sources with the appropriate credibility including collaboration information received from partners.
ii) Static Impact Evaluation: the immediate assessment of the relevance of the attack at that point in time given the received intelligence and the known configuration of the infrastructure.
iii) Dynamic Risk and Value Assessment (DRVA): the relative values of the "at risk" assets from the perspectives of different stakeholders combined with their exposed known vulnerabilities and the known attacks. In this function an intelligence value is also calculated for the information that may be gained by responding in an "unconventional" manner. The organisational goals are also taken into account in creating this assessment for both the asset and intelligence values.
iv) Modelling: this is the prediction of the future path of the attacks based on known attack patterns, attackers, exposed vulnerabilities and asset values. Combined with the output of the DRVA this provides the decision maker with optimal enhanced situational awareness.
v) Decision: based on the modelling, the DRVA and the static impact evaluation, the responsible decision maker takes the organisational goals into account before deciding on a course of action. They are provided with a number of response options (which may be reduced by their legal and organisational constraints): these options are:

a. A conventional response, i.e. defend against the attack via conventional means (for example blacklists, IPS, etc).
b. Passive monitoring response, i.e. observe but show no reaction at all to the incident (as though it was undetected) in order to gain intelligence.
c. Active intelligence gathering, i.e. actively reconnoitre the attacking infrastructure by any means possible in order to gain intelligence but without intentionally causing disruption to the attacking infrastructure.
d. Cyber strike, neutralise the attacking infrastructure via any available Cyber means.

5. CONCLUSIONS

By analysing the relevant literature it is concluded that the traditional responses to Cyber-Incidents and the implementation of these models are not meeting the requirements of all communities impacted by them. In order to meet these requirements, not only do responses need to be based on the "balance-of-equities" decision between the priorities of the different stakeholders whose assets are being attacked, they should also take account of the value of intelligence (both local and collaborative) associated with an attack and consider a more flexible suite of response options. The proposed Dynamic Cyber-Incident Response model enables those responsible for cyber-incident response and their key decision-makers to develop a more dynamic set of response procedures within their legal and organisational constraints. That is not to say that if a high-value or critical asset is being attacked that it should necessarily be allowed to fall in order to gain intelligence; however, if a low value asset is being attacked and the attack or attacker is unknown or novel, the organisation might be better served by learning about the attack rather than defending the asset. With this approach, the gained intelligence could well help to defend a higher-value asset in the future.

6. FURTHER WORK

The next stages of this work will be to evaluate the survey data and refine and develop the proposed model. The intention is evaluate the model in a variety of deployment scenarios utilising a purpose-built Cyber Range at the university. The current evaluation criteria for the model are expected to be

i) Assessment of intelligence gains which may be achieved by allowing a predefined set of cyber incidents to continue under observation.
ii) The contribution of DRVA to the situational awareness of the decision-maker and consequent influence on their ability to make the optimal decisions.

REFERENCES:

Barnum, S. (2012). *Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX)*. Online: Mitre Corporation.

Calder, A., & Watkins, S. (2008). *IT Governance A Manager's Guide to Data Security and ISO27001/ISO 27002*. London and Philidelphia: Kogan Page.

Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). *Special Publication 800-61 Revision 2; Computer Security Incident Handling Guide*. NIST, US Department of Commerce.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors The Journal of the Human Factors and Ergonomics Society, 37*(1), pp. 32-64.

Hallingstad, G., & Dandurand, L. (2011). *CIS Security (including Cyber Defence) Capability Breakdown*. The Hague: NATO Consultation, Command and Control Agency (NC3A).

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis - Pearson New International Edition* (7th ed.). Upper Saddle River, New Jersey, US: Pearson.

Killcrece, G , Kossakowski, K.-P., Ruefle, R., & Zajicek, M. (2003). *State of the Practice of Computer Incident Response Teams (CSIRTs)*. Pittsburgh, PA, USA: SEI Carnegie Mellon University.

Lawson, J. S. (1980). Command control as a process. *19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes* (pp. 1-6). IEEE.

MoD, UK. (2011). *JDP 0-01 British Defence Doctrine*. Shrivenham, UK: MoD Development, Concepts and Doctrine Centre.

MoD, UK. (2011). *Joint Doctrine Publication 2-00 Understanding and Intelligence Support to Joint Operations*. Shrivenham, UK: MoD Development, Concepts and Doctrine Centre.

Northcutt, S. (2003). *Computer Security Incident Handling*. SANS Institute.

Orr, G. E. (1983). *Combat Operations C3I (Command, Control, Communications, and Intelligence) Fundamentals and Interactions*. Maxwell Air Force Base, Alabama : Air University - Center for Aerospace Doctrine, Research and Education.

Rowe, N. C. (2006). Measuring the Effectiveness of Honeypot Counter-Counterdeception. *Proceedings of the 39th Annual Hawaii International Conference on Periodical, System Sciences, 2006*. (pp. 129c-129c). IEEE.

Tzu, S. (2011). *The Art of War (translated by Thomas Cleary)*. Boston and London: Shambala.

Verendel, V. (2009). Quantified Security is a Weak Hypothesis. *Proceedings of the 2009 Workshop on New Security Paradigms* (pp. 37-50). New York, USA: ACM.

Wang, P , Wu, L., Cunningham, R., & Zou, C. C. (2010). Honeypot detection in advanced botnet attacks. *International Journal of Information and Computer Security*, 30-51.

West-Brown, M. J., Stikvoort, D., Kossakowski, K.-P., Killcrece, G., Ruefle, R , & Zaijicek, M. (2003). *Handbook for Computer Security Incident Response Teams (CSIRTs) 2nd Edition*. Carnegie Mellon University.

Yuill, J., Wu, F , Settle, J., Gong, F., Forno, R , Huang, M., & Asbery, J. (2000). Intrusion-detection for incident-response, using a military battlefield-intelligence process. *Computer Networks*, 671-697.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Beyond technical data - a more comprehensive Situational Awareness fed by available Intelligence Information

Andreas Kornmaier
Faculty of Computer Science
Universität der Bundeswehr München
D-85577 Neubiberg, Germany
andreas.kornmaier@unibw.de

Fabrice Jaouën
Deputy Assistant Chief of Staff - CJ35
Multinational Joint Headquarters Ulm
D-89081 Ulm
fabricejaouen@bundeswehr.org

Abstract: Information on cyber incidents and threats are currently collected and processed with a strong technical focus.

Threat and vulnerability information alone are not a solid base for effective, affordable or actionable security advice for decision makers. They need more than a small technical cut of a bigger situational picture to combat and not only to mitigate the cyber threat.

We first give a short overview over the related work that can be found in the literature. We found that the approaches mostly analysed "what" has been done, instead of looking more generically beyond the technical aspects for the tactics, techniques and procedures to identify the "how" it was done, by whom and why.

We examine then, what information categories and data already exist to answer the question for an adversary's capabilities and objectives. As traditional intelligence tries to serve a better understanding of adversaries' capabilities, actions, and intent, the same is feasible in the cyber space with cyber intelligence. Thus, we identify information sources in the military and civil environment, before we propose to link that traditional information with the technical data for a better situational picture. We give examples of information that can be collected from traditional intelligence for correlation with technical data. Thus, the same intelligence operational picture for the cyber sphere could be developed like the one that is traditionally fed from conventional intelligence disciplines. Finally we propose a way of including intelligence processing in cyber analysis.

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any agency of the government Examples within this article are based only on very limited and dated open source information Assumptions made within the analysis are not reflective of the position of any government entity

We finally outline requirements that are key for a successful exchange of information and intelligence between military/civil information providers.

Keywords: *cyber, intelligence, cyber intelligence, information collection fusion*

1. INTRODUCTION

Cyber attacks and incidents take place on a daily basis, but only few become known to a broader community. Nevertheless, the known cyber attacks with their severe results, e.g. the closure of the company HB Gary Federal, motivate IT Security to improve defensive measures to protect their organizational networks and the data and information stored in these.

In order to protect the networks they are monitored with sensors and tools on servers and network nodes to provide lower-level network event-oriented alerts. The use of the tools and the analysis of the lower-level data require in most cases highly technical trained network security experts.

They are also analysing detected attacks to understand how the attacker was able to gain access to the system using vulnerabilities and weaknesses in hard- and software and their configuration [1].

The information collected by the sensors and the evaluated attack data that are currently collected and processed have a strong technical focus that is mainly directed inwards.

Threat and vulnerability information alone are not a solid base for effective, affordable or actionable security advice for decision makers. They need more than a small technical cut of a bigger situational picture not only to mitigate, but to combat the cyber threat. The technical information needs to be transferred from “geek” vocabulary to a format understandable by the decision maker [2]. Nevertheless, one must admit that not even when this process is completed the decision maker has a real and full understanding over the situation, although this should ideally be appropriate for him to develop and coordinate detailed plans, ensuring by the way that he stays interested in cyber defence planning [2].

Thus, cyber specialists are encouraged to go this way as it is true that the principles of war have not changed with the development of the cyber dimension. Clausewitz’ statement “War is the province of uncertainty: three-fourth of those things upon which action in war must be calculated, are hidden more or less in the clouds of great uncertainty.”[3] applies to features of the modern Information Technologies. The tempo set by cyber-attacks, in some cases their hidden or at least discrete infiltration into the systems keep the decision maker in a false sense of security, being completely ignorant of the inherent danger. On the other hand, lacking any understanding in cyber matters could as well drive him to a form of paranoia by fear of full scale cyber-attacks, this feeling being fed by some part of irrationality.

In that sense the cyber specialist plays a critical role in the decision making process, helping the leader to strike a balance in the effective threat level posed by cyber issues. Fully involved in the leader’s support and advisors’ team, this expert is expected to cover one major task: developing the awareness on cyber issues in support of those making a decision.

Therefore it is necessary to transfer the technical information into the language of the decision maker and put it into his/her context by supplementing the technical monitoring data with

further available information or intelligence [4], thus relating the cyber dimension to the overall operational framework. This approach will also provide a more extensive situational awareness, enabling a more comprehensive decision making. The required information is very often already available, even correlated, but not linked. Thus, it is now necessary to take a look at already available conventional data that needs to be collected and fused with the traditional security event data, not only to be reactive to threats, but to be enabled to predict and prevent attacks [5].

Very little research has addressed the use of already available information to put technical data into an operational/ strategic context. In this paper, we first give a short overview over the related work that can be found in the literature. We then evaluate the approaches. Following this we examine what information categories and data already exist, identify information sources in the military and civil environment, before we propose to link that traditional information together with the technical data for a better situational picture. Finally we propose a way of including intelligence processing in cyber analysis.

This paper is not intended to describe specific techniques or potential theoretical frameworks for a better situational awareness through correlation of context information. Legal constraints and regulations like privacy laws that legitimately limit data acquisition are also beyond the scope of the effort of this paper. This is also the case for lack of cross-border treaties for data sharing and data constraints and restraints that might exist in regards to mission, civilians, enemy, time, ROE.

We further identify requirements, where information needs proactively to be looked for by tasking. We approached the field through a literature review, experience, participation in cyber defence exercises and many fruitful discussions with IT Security specialists and intelligence officers.

In the next section we begin by describing the related work identified by performing a literature review on conventional data and information to be used for better situational awareness and more comprehensive decision making in the cyber context complementing technical data.

2. RELATED WORK

For the literature query we were looking at several approaches in the literature for fusing data and structuring information in a format. We also looked at contributions to situational awareness, the common operational picture (COP) and the decision making during the literature review. All papers have a limited focus in regards to our research, so we only touch the most relevant developments with findings for our research.

In [6] we found generic threat matrixes that allow to categorize threats and thus to define a common vocabulary for them. Although a common terminology as a basis for successful understanding of different groups (e.g. technicians and decision makers) is still missing [2], some different categories of players can be discriminated [7]. While in the past and in some current conflicts the organized masses (states, armies, ethnicities) of people were at the core of the analysis, the cyber dimension has led to the emergence of smaller groups or even individuals as possible adversaries of a much larger organization.

Opposing in some way Clausewitz' approach of war to the cyber dimension of conflicts, Kempf

underlines the emerging role of the individual. While in former albeit various forms of conflicts between states, organized bodies were in the leading role, individuals are now able to operate, even in a limited dimension, against stronger, larger structures from remote and safe locations. In addition to those isolated persons, formal or informal groups act in the cyber dimension, either motivated by crime or political activism, finding there a good opportunity to set plans, reach their goals or get some financial or political profit.
However, their large diversity prevents the analysts from any simplification as this could drive them to a misleading understanding of the threat. As a matter of fact, the knowledge of the 'hostile' Tactics, Techniques and Procedures (TTP) has to be permanently checked and balanced with the effective capabilities of the most probable adversary, without excluding the other ones. Yet, this overall framework being in a permanent movement and transformation plays different roles in the decision making of leaders, depending on their objectives and on the vulnerabilities offered in reaching for their own goals to those individuals or groups.

The large amount of potential third players who could influence the own action gives then the analysis of the cyber threat a paramount importance, in order to provide the leader an appropriate level of information before making his decision.
To reach this goal a structured and comprehensive approach is required and provided by different tools developed by the specialists in cyber issues. If not, the result would be giving the potential threat an infinite complexity that would severely hamper any trial for a sound cyber defence.

The Structured Threat Information eXpression (STIX) is a collection that includes various sets of cyber threat information. The available sets in STIX offer a structure to store information on Indicators, Incidents and Adversary TTPs including attack patterns, malware, exploits, tools, infrastructure, targeting, etc. Also information on exploitable targets like their vulnerabilities and weaknesses can be put into STIX, as well as different remedial actions (Courses of Action) to respond to incidents or to vulnerabilities/weaknesses.
In STIX also information can be included on Cyber Threat Actors and their Cyber Attack Campaigns [1].
For the representation of the information STIX uses other, already developed structures. For information like 'cyber observables' (operational cyber events or stateful properties such as registry keys, email, and network flow data) it uses the definitions of the Cyber Observable eXpression (CybOX) language. The Common Vulnerability Enumeration (CVE), Common Platform Enumeration (CPE), Common Weaknesses Enumeration (CWE), and Malware Attribute Enumeration and Characterization (MAEC) are ingredients of STIX to describe standard information about vulnerability (using OVAL, the Open Vulnerability and Assessment Language), platform, weakness and malware. For describing an attack it uses the Common Attack Pattern Enumeration and Classification (CAPEC).
In summary it can be stated that STIX allows to represent cyber threat information in a structured, standardized manner [1].

Data fusion is in [8] described to be extended into the cyber security incident management domain. In [9] the basic data for several fusion levels come from Sys Logs, Web Logs, IDS and IPS alerts. All four data sources are technically aligned in that Data Fusion Approach for Cyber Situation Awareness and Impact Assessment.

Other approaches focus on establishing a methodology or metrics to characterize the threats consistently and add with the measured observables to a situational picture [4], [6].
Usually open-source information is utilized and not necessarily secret intelligence [4], although the latter will never be excluded depending of the threat level against the vital functions of the target.
Hutchins states in [10] that "it is possible to anticipate and mitigate future intrusions based on knowledge of the threat" and proposes an "intelligence-driven, threat-focused approach to study intrusions from the adversaries' perspective."
In the military and security environment the term intelligence stands for understanding and knowledge in the military and security context. But it is also used for reports and summaries that provide information with an assessment and added benefit to decision makers, operational planners and intelligence specialists to round up their situational picture for their further work [11], [12].
Classical questions for the intelligence community are the adversary's intent as well as TTPs.
In the context of countering Cyber Terrorism David proposes in [5] the establishment of a Cyber Intelligence Analysis Centre generically outlining a cooperation of governmental and civil entities focusing on technical means.
[5] postulates that intelligence "should provide the essential elements of enemy information: who, what, when, where, why and how. That is, who will attack what, at what time and place, for what purpose and objective, and with what type of resources and methods."
In [5] it is proposed to achieve this goal by fusing information from multiple sources to learn and analyse the tools, tactics and motives.

As traditional intelligence tries to serve a better understanding of adversaries' capabilities, actions, and intent, [1] argues that the same is feasible in the cyber space. He uses the term cyber intelligence for this cyber focused field. According to [1] cyber intelligence is to give responses to relevant threat actors, their suspected intent, and adversary's possible and taken Course of Action. This includes technical targets like sort of vulnerabilities, misconfigurations, or weaknesses an opponent is likely or used to exploit in attacking their objective [1]. To achieve this, cyber intelligence has to analyse opponent's capabilities in the form of their TTPs. TTPs are derived from the traditional military sphere, where they are used to analyse and predict an adversary's actions and methods. Therefore, TTPs have a central role not only in traditional intelligence, but also in the cyber sphere [1].
Nevertheless, all approaches are missing to take a view on already available information, traditional established information structures and how they could be benefited from.

3. EVALUATION OF EXISTING APPROACHES

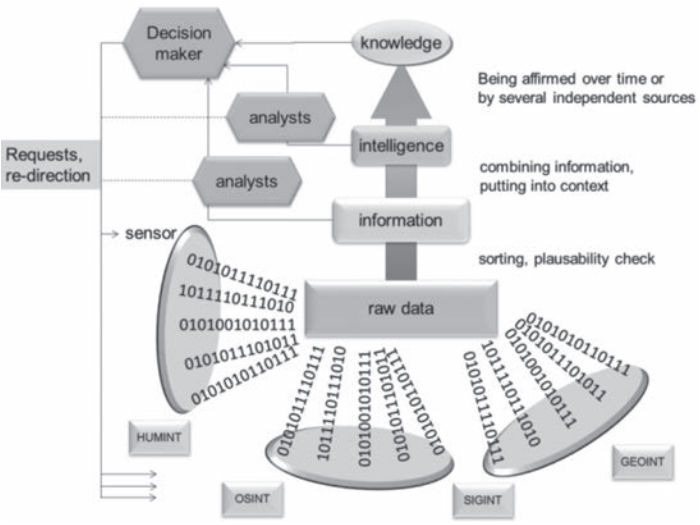
Our centre of interest being set on the efforts to collect, link and fuse information or exchange it [8], we left apart the understanding of the different groups, for which we suggest to refer to already existing typologies [6], [7].

In the related work the focus is mostly set on technical (network- and packet-level) data derived from lower-level security tools and their storing in data structures for further processing as

described in [1] and [9]. That information is of course relevant to describe a network topology or events within a known network infrastructure [13].
For this purpose STIX includes several other well defined and established structures. It can be summarized as overarching framework of several specialized smaller frameworks. Nevertheless, all found efforts concentrate on technical aspects and their assessment. But it must be stated, that threat and vulnerability feeds by themselves do not produce intelligence on cyber threats. Nor are the results effective or actionable in regards to a situational awareness or for a decision making.

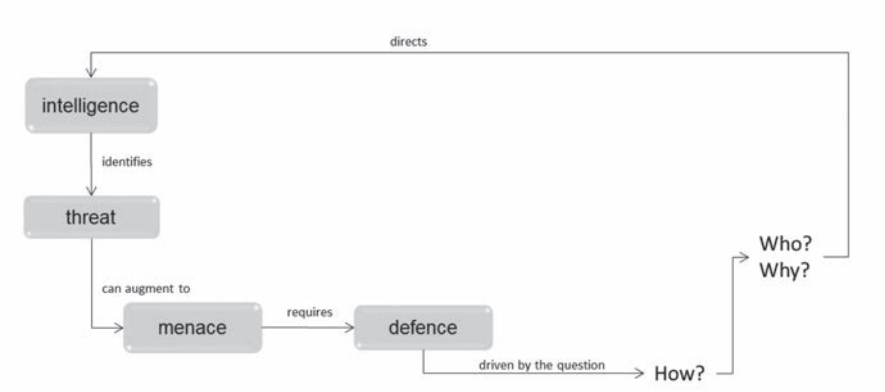
At first technical data comes unstructured and it needs to be decided, what is relevant and/ or representative for further processing and assessment by a skilled analyst [6], [14] (comp. figure 1).

FIGURE 1: PROCESSING OF RAW SENSOR DATA TO INTELLIGENCE AND TO KNOWLEDGE; READJUSTMENT OF SENSORS (OWN ILLUSTRATION)



He can assess the actual and mostly historic data to give an estimate on the current threat or on a preceded attack/incident from a technical perspective. That kind of information has been seen as an important type of knowledge by almost all above described approaches. But for a proper assessment on a more abstract layer, where non-technical information is in the focus, further information that is collected and processed is needed. For example in an assessment on taken informational damages during an incident/ attack that bases solely on technical data, it is mostly analysed “what” has been done, instead of looking more generically for the tactics, techniques and procedures to identify the “how” it was done [10]. The “how” allows the defender to evaluate capabilities and objectives, maybe even limitations and doctrine of the attacker [10] (comp. figure 2).

FIGURE 2: IMPORTANT ROLE OF INTELLIGENCE (OWN ILLUSTRATION)



In some way, critically needed are "intelligence-based earliest assessments of adversaries' intent" [4].

The intelligence analysis gets its real value by prioritizing the potential threats depending on their level of technological danger and their will or intent to effectively disturb the networks or activity of own assets. Dossé pledges for such a discrimination of the threat [15]; e.g. in conventional military assessment, the different levels of threats are to be discriminated: a single man attack with a rifle that is not considered to be at the same level as an offensive with an armoured corps.

Very often the statistics published by administrations do not help figuring out the effective threat they are confronted with, as they release the number of attacks they are confronted with on a certain period of time, without sorting out which were of critical importance and which could be simply disregarded as considered irrelevant.

Although it has never been and will never be an exact science, intelligence analysis provides the appropriate understanding needed to support a sound and efficient decision process.

The combination of capability and intent allows an assessment beyond forensic after-attack assessments in form of predictions and warnings that address events in the (near) future [4].

The approaches that are based on technical data miss mostly the aspect of the adversary's intent, also if expressions like adversary's intent, Tactics, Techniques, and Procedures, courses of action are considered, but they are used always in a technical context. Also if technical experts hypothesize about intent and goal of attacks, the "task of drawing such conclusions is more professionally handled by judiciary, intelligence and diplomatic authorities" [8].

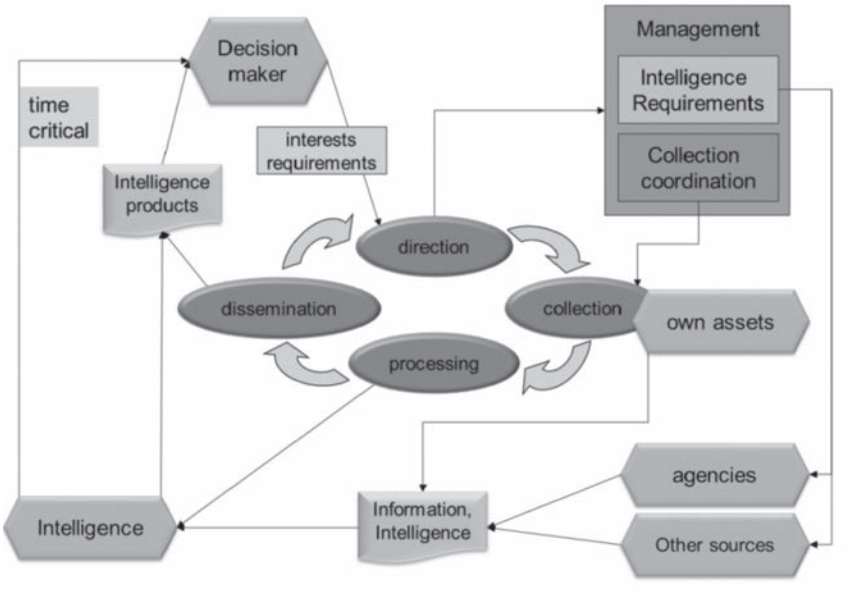
Intelligence is dealing with uncertainties; the more information is cross-checked and subsequently validated or confirmed, the more accurate the assessment will be, in an attempt to decrease as much as possible the number of mistakes, through the intelligence cycle depicted in figure 3.

Information that is needed by and relevant for the decision maker provides through the Intelligence processing an accurate situational awareness [11], [16].

Such a best possible accurate situational awareness is a prerequisite to make appropriate

decisions [9], [11]. The intelligence efforts are driven by the information requirements of the decision maker, who can directly readjust the efforts by giving guidance [11], [16]. As questions always aim to recent developments and changes, information in databases and repositories are never sufficient to respond to the information request. Therefore, a need arises with the decision makers' request to collect more information via the available various collection disciplines [16].

FIGURE 3: THE INTELLIGENCE CYCLE IN THE COLLECTION COORDINATION AND INTELLIGENCE REQUIREMENTS MANAGEMENT (OWN ILLUSTRATION).



It can be summarized that Intelligence is the basis of on which a decision for operational activities is built [12]. Or in other words Intelligence drives the mission. "Thus the intelligence contribution must begin even before operational planning starts."[12]

The cyber domain is reaching into the domains air, land, sea and space as 5th dimension [19]. Reaching implies overlapping areas. This is underlined by the fact that many assets have a position in the physical as well as in the virtual cyber environment [2]. This feeds the assumption that cyber might be a different "view" on or classification of information, data, assets etc. Thus, cyberspace is not really something completely new and we can examine existing traditional information sources and repositories.

It is necessary to take into consideration that although defining cyberspace as an abstract fifth dimension, it is physically based on hardware components [23]. The hardware is used by persons with capabilities with some intent. Therefore, fusing technical data, e.g. derived from raw network packets, with traditional intelligence appears to provide more comprehensive analysis of the cyber threat on a more precise level than before as it includes the human factor, which is per se neglected in any exclusively technical analysis [20], [23].

In short, a technical capability to harm one's systems is irrelevant as long as there is no intent to do so.
If this discrimination process is not implemented, the decision maker will undoubtedly suffer an overdose of possible threats that could paralyze his action. This critical mitigation between risk and opportunity makes the decision making much easier.

Focusing only on technical data that is delivered by physics-based sensors, it must be kept in mind, that sensors can only be put in dominated or at least controlled areas. Otherwise they become vulnerable and can be manipulated [14].
New and/ or actionable knowledge may result from low-level data that became meaningful information by a goal-directed cross-linking of different information products [20], [22].
Finally trustworthy intelligence will be created from this knowledge in a cyclic (intelligence) process [20] that has several iterations and readjustments caused by quality of source and information as well as by cross-checks. In cross-checks often available and potentially conflicting information shall be verified or falsified to confirm a situation. This makes it a time intensive challenge for the human analysts although absolutely necessary in order to avoid misleading conclusions [11], [14].

The traditional security tools that are used in network monitoring are generally only point solutions that provide only a small technical section of a bigger context [20]. Thus, it becomes apparent that the technical data needs to be merged with complementary information [14].
Information elements are generated by different, often heterogeneous sources [22]. They do not only include computer network specific sensors, but also other physical sensors and human sources.
Following [19] in dividing cyberspace into a physical, a logical, and a social layer gives a good first base for the types of information that need to be looked at, further examined and exploited for a more comprehensive situational awareness.

Intelligence, surveillance and reconnaissance in and of cyberspace need to be conducted to "bring light" into uncertain situations and meet the information need.
With the novelty of cyber some introduce in the domain of intelligence the term Cyber Intelligence. If it is used in the sense of 'collecting, analysing and countering of cyber security threat information' it might fall short, especially when the focus lies only on technical information.
The emphasis of the intelligence efforts for Cyber or in short *Cyber Intelligence* is different from those for conventional intelligence operations, also if adversary intent and capability are for both of interest. Cyber Intelligence identifies Cyber Threats on the understanding of the global network and computer architectures and associated threats by analysing and fusing conventional threat data with network information. By merging those with global events the actual technical network border can be penetrated.

At the moment Cyber Intelligence appears to be strictly defined in technological terms by technical experts, what is not in the best interest for the task and needs to be completed by a broader inter-discipline view in order to meet the operational requirements [18].

Missing is the connection between the collected technical data and information that is already available in different traditional established information structures and domains. Thus, we follow [5]'s argumentation that the focus should be on fusing information from multiple sources to learn and analyse the tools, tactics and motives and take a look in the next section to the different disciplines of "traditional intelligence in possible support for cyber aspects [4].

4. INTELLIGENCE SUPPORT FOR CYBER SPHERE

Technical data has often been collected mindlessly and it was tried to make sense of the huge data sets [21]. To find useful information or even intelligence in that enormous amount of data, the strategy of mindless collection and purely technical assessment must be changed. The technical data must be a part of the bigger situational picture that gets information also from the traditional intelligence disciplines for fusion.
We state that in the traditional intelligence fields information is already available or can easily be collected by adjusting the intelligence collection plan.
Therefore it is necessary to take a look at the different disciplines and the conventional data produced and available in them, waiting to be collected and merged with the technical data.
The basic groups of collection disciplines are Human Based Intelligence (HUMINT), Imagery Intelligence (IMINT), Open Sources Intelligence (OSINT) and Signals Intelligence (SIGINT) [12], [16], [24].

In HUMINT data like names, locations, as well as motivations and capabilities are processed. In addition it could be also directed to find WebIDs. As well, HUMINT contributes to the drawing of human networks, thus enabling to understand the possible underground ramifications of an apparently isolated threat.
OSINT can provide host information, IP numbers, information on the used ISP, the location, WebID, homepage(s), blogs etc. It can be done in a technical approach, but also in a more abstract level, e.g. via scanning social media. Associated with HUMINT, OSINT enhances the merging process between the verbally expressed intent and the effective behaviour.
IMINT can provide further information about a location, used infrastructure, types of antenna and possibly about networks, especially in connection with GeoINT, HUMINT and SIGINT.
SIGINT intercepts can not only reveal the transmitted message, but also show the way of data. Thus, further analysis might implicate on top of a physical network a virtual usage network.
GeoINT can bring an invaluable added value to the overall analysis process through their capability to manage large databases originating from various economic fields.
As a summary, any data related to grids is of use, be it servers, data centres, web cafés, that is any facility being assessed to be of interest in the analysis of the cyber threat.
Even though varying from one organization to the other, intelligence reports may be characterized in four categories:

- immediate reports to broadcast brand new information,
- timely reports, which include an assessment and intend to give the heads up,
- ad hoc reports dedicated to one specific issue and
- national intelligence reports.

The latest category is of a peculiar interest in the field of international cooperation, as those documents are the steppingstone for deciding what can be shared or not.
These products usually include analysis of adversaries, their capabilities, objectives, doctrine and limitations [10].
To support the cyber efforts, those products should include information or details relevant for cyber intelligence. That is the case, when they are in relation to the cyber environment, either in the physical or in the virtual cyber sphere. Many relevant intelligence snippets can be found in open sources like chat rooms, postings in forums, blogs and news groups, but also in e-mails, wikis, web sites, social media and messaging communications. By looking for identified buzz-words in a first automated scan and then refining the search taking into consideration further information that is connected to the first results. Those sources are open and thus available and easily accessible. Information from private communication channels in blogs and forums can be obtained, but this by passing control mechanisms, e.g. a registration.
Information from hacker forums is of special interest. There are chances to find commonalities in different attacks by correlating network data. Thus, not only in regards to content-analysis the forums have to be examined deeper, but also network data can be gained by specific collection efforts [21]. Those forums provide rich conventional intelligence, but also cyber specific information as it is distributed via or hosted in cyberspace.

A cyber skilled analyst could merge conventional information/ knowledge and political events with the cyber specific data [21]. He can develop the same intelligence operational picture for the cyber sphere like the one that is traditionally fed from conventional intelligence disciplines. Thus, he can create a more comprehensive understanding of a potential threat or attack by including context and his experience.
As intelligence is looking over longer periods for reoccurrences, the aggregating of data from multiple sources will reveal patterns that are not evident from a single source [21]. Intelligence can be distinguished between tactical, operational and strategic level. Technical data of a machine or in a network segment corresponds to tactical intelligence level [17]. Operational or even strategic intelligence needs to look beyond the bits and bytes correlating and linking activities of maybe years as for Intelligence on that level not so much the single event or a phase of the event is of interest. It's more the cyclic reoccurrence and the pattern that enables to possibly predict further adversary measures.

After an attack a forensic examination of the intrusion artefacts will provide at least a section of the timeline of the attack/ incident, but also technical data for further investigation and intelligence tasking. For example the examination of the STUXNET source code included snippets that were giving hints to where the originators come from.
Starting from an IP that is associated with an attack a lot of information can be collected on a technical level. The IP allows to get e.g. host name, geo-location to identify the physical origin (or last used echelon) of an attack, the ISP that has registered that IP etc. However, as IP addresses can easily be masked or spoofed, the reliability of that information is poor. Therefore, it is up to the intelligence disciplines to provide further information. Who was using the host with the given IP? What WebID was he using for his actions? Exist further occurrences of this WebID, maybe in similar context? Is only one person using the WebID or several persons? Are there other services he is using with this WebID in the internet? Is he using other WebIDs (e.g.

different e-mail accounts, different login names)? Is he using blogs and social media, what information are contained and published there? What motivations and capabilities does the person have to initiate an action that he is now being under examination? What pictures are published, do they contain geo-tags so that through IMINT and HUMINT further information can be gained by a redirection of intelligence/ reconnaissance efforts. If there is enough information the person can be profiled by his customs, locations, used internet services and also his interests, his intent and capabilities.

Also other starting points are possible like data from an investigation or from a signal intelligence measure. An examination and fusion of social media profile data is also thinkable, if indications exist that justify this proceeding.

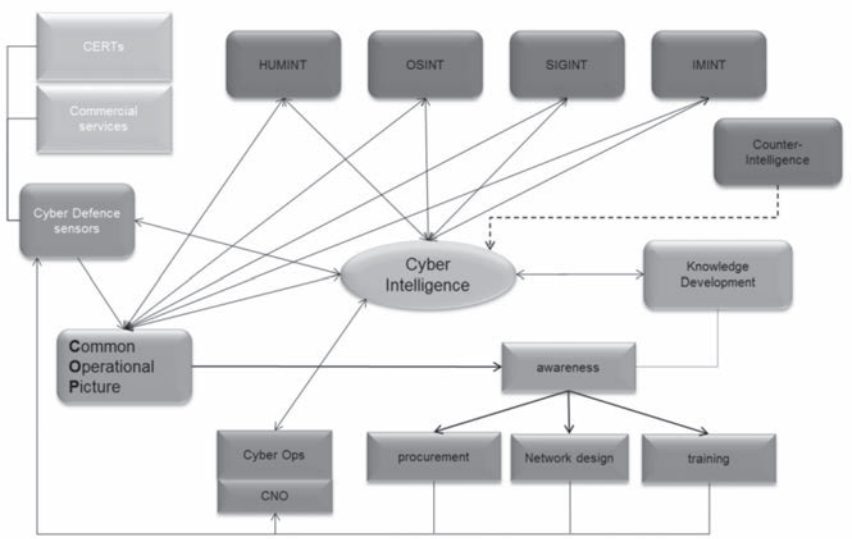
With this approach a professional assessment beyond pure technical evaluation and hypothetical assessments can be made. Fusing the different information from the various intelligence disciplines creates knowledge about the adversaries.

Presenting the relevant information and intelligence in an appropriate way for the receiving audience in order that they can understand the given information and possible effects and results from it, this increases the awareness and causes a better common operational picture serving decision makers as basis for proper and

comprehensive decisions [2], [17]. Better fundamental information can be fed back into the intelligence loop and a more focused readjustment of the efforts by the decision maker is possible. (see figure 3)

For example, an assessment that attacks are not likely at the moment allows the decision makers to turn their attention to more pressing matters [4].

FIGURE 4: VARIOUS SOURCES FOR CYBER INTELLIGENCE, INTERDEPENDENCIES AND POSSIBLE PURPOSES OF USE (OWN ILLUSTRATION)



The gained intelligence must iteratively be exploited and pursued for own objectives. It allows advancing development of own procedures, standards, doctrine and policies. Thus, the findings from analysing the adversaries TTPs and capabilities can be used to adapt own defensive cyber training. In addition it also allows to change passive devices' settings as well as the consideration of the findings in a re-design of the own network or at least in the design of own future networks. They finally should also be taken into consideration in decisions for procurement of hard- and software. (see figure 4)
When technical data and gained intelligence are supplemented with information from the private sector, a very comprehensive picture is created, because commercial/ private/ civil companies/ organisations have other resources and legal constraints. Finally, they are complementary.

5. REQUIREMENTS

For cooperation a common terminology is essential. Only then, there will be clarity among different, probably far away located and maybe even multi-lingual actors. Such a basic understanding is prerequisite for common data analysis in conjunction with all possible intelligence sources and for any following further dissemination of information/ intelligence.
The sharing and exchange of information must be driven by the aim to be better than the status quo by providing effective, timely and actionable intelligence. This allows a comprehensive situational awareness and supports the decision maker in continuous planning and executing Cyber Defence actions. This is achieved by observing and analysing menacing cyber activities and trends [17].
All efforts need to be designed for sustainment. This is underlined by the fact that neither the government, nor the private sector alone can defend against the cyber threat effectively and efficiently. In addition there exist too many approaches to defend everything [21]. Therefore, the efforts must be focused appropriately, which is the main role of intelligence. As developed by Lieutenant-Colonel Foch in his conferences at the French War College, 'economy of forces' consists in selecting where and when forces are to be used the best, instead of trying to face all the possible situations [25]. The cyber threat genuinely and from a purely technical point of view being possibly originating from various locations and using different vehicles, this fine selection of the directions and locations where the cyber defence should focus is of primary importance.
Information and data of penetrations or attacks that are directed against the entire critical infrastructure (CI) are of interest for fusion. By the mainly private nature of the CI and the high interest due to the dependency for governmental functioning, an information exchange between companies and organisations of the CI sectors and governmental institutions will be essential to counter the menace [17]. Neither an intelligence organisation (most are specialised in one intelligence discipline) nor a governmental institution nor a private company can collect, produce or even access adequate intelligence on their own. To keep that status quo will not improve the chance to have reliable data in an environment that has to deal with many uncertainties [1]. Only sharing of relevant cyber threat information will overcome this limitation and enable an informed decision making. For success all sharing partner must contribute. It is not only a "give", but also a "take" liaison. Benefiting from partner's information and intelligence a potentially more complete understanding of the threat landscape can be achieved [1].

A first step will be to cross train cyber and intel personnel in organisations, so that they are able to understand and transfer requirements and limitations of the other work domain.
In consequence, this approach will take the technical based abstract level to a more concrete level, specifying more precisely the attacker. Maybe in the future even an identification and attribution could be possible, when the limits of governmental institutions, international organisations, and civil companies have burst.
Thus, establishing regulations in strategies and policies for the exchange of information and intelligence in the above outlined cyber context is the essential first step in the described process. It must be defined who shares what, with who, under what circumstances, how the information is handled, classified, processed and stored [1]. These regulations are necessary, because on the one hand there exists no broadly accepted standard for sharing information or even intelligence across agencies or private companies. On the other hand – mentioned for completeness – trust is the key for increasing the sharing behaviour. Trust for the exchange occurs at the individual and organizational level [26], [27]. It is the degree of confidence to handle the information/ intelligence with the same sensitivity. Only then the exchange will take place. As well, cooperation between sovereign states is to be fostered for a better efficiency in cyber defence [18].
Agreements between organizations, agencies and private companies and the consequent, augmenting exchange of information are a way to build this trust. On the individual level it is the personal relation, or better interpersonal confidence between the subject matter experts that builds up trust over time and generates consistent and positive effects.
A combination of both is established when institutions are created that host several representatives of different institutions and they meet in order to exchange and merge information [23]. On top of the organisational trust this promulgates the individual one.
The agreements for an exchange of information are not only basis to build this trust, but also necessary to formulate the regulations and control mechanisms as well as the interoperability needs.

6. CONCLUSION

The purely technical data feed is always there and therefore certain. However, now uncertainties have to be accepted and dealt with, when information is merged in cooperation with the intelligence community and other information providers like CERTs. We have shown that persistently consolidating data from disparate sources into meaningful and complementary information allows better and more precise assessments about an adversary's capabilities, his intent and his location. This enhances the cyber situation-awareness and allows a more extensive situational cyber picture.
Those resulting details are actionable intelligence (with all the uncertainties being inherent in such assessments) that allow not only after action or tactical situation updates, but predictive, strategic warning in regards to cyber threat activities. Thus we get away from a purely reactive and defensive position to a foreseeing, flexible, proactive one.
The analysis and fusion of the technical and (geo)political events remains rooted in each analyst's experience, background, and expert opinion. Therefore, providing clear intelligence to the analysts is essential to prevent erroneous conclusions finding their way into the situational awareness.

Of new importance are intelligence, surveillance and reconnaissance operations across multiple intelligence disciplines in the context of cyber. Those operations and methods have not changed, they must only be adapted to the cyber sphere. This requires that the cyberspace is better understood and processes take special properties of cyberspace into consideration.
Same is true for sharing cyber intelligence and cyber threat information: Though we have established agreements and mechanisms to exchange conventional information and intelligence, the supposed novelty of cyberspace and specific properties of the cyber sphere hinder the sharing and exchange of that information.
The national and international organisational structures need to adapt to the new challenges and needs.
Our approach helps to create awareness for the correlation requirements of information of the cyber sphere and the traditional intelligence disciplines.
Further research will have to address who shares what, with who, under what circumstances, how the information is handled, classified, processed and stored; also if the trust question is still not solved.
In the exchange of that valuable fused information rests a high potential to shift the balance between attacker and the defender [1].
Fusing the complementary information to actionable intelligence allows decision makers better to prevent surprise attacks in Cyberspace and the way we respond. The information "nuggets" are out there waiting to be collected.

REFERENCES:

[1] S. Barnum, Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX™) [Online]. MITRE Corporation, 2012. Available: https://msm.mitre.org/docs/STIX-Whitepaper.pdf
[2] M. Lanham (2012), Operating on unconventional Terrain - Cyber Defense Planning [Online]. Army Communicator. Available: http://www.dtic.mil/dtic/tr/fulltext/u2/a571985.pdf
[3] C. v. Clausewitz, On War [Online]. Available: http://www.gutenberg.org/files/1946/1946-h/1946-h htm
[4] J. Healy and L. van Bochoven (2012), Strategic Cyber Early Warning: A Phased Adaptive Approach for NATO [Online]. Atlantic Council. Available: http://mercury.ethz.ch/serviceengine/Files/ISN/155419/ipublicationdocument_singledocument/22f57269-9d44-4cac-ac21-21423273e1d1/en/NATO+Cyber+Warning+2012.pdf
[5] M. David and K. Sakurai, "Combating Cyber Terrorism: Countering Cyber Terrorist Advantages of Surprise and Anonymity", Proc. IEEE AINA, pp. 716 – 721, 2003
[6] M. Mateski et al., "Cyber Threat Metrics", Sandia National Laboratories, SANDIA REPORT SAND2012-2427, Unlimited Release, 2012
[7] O. Kempf, in: Introduction à la Cyberstratégie. Chapt 6, pp. 78-99. Economica. 2012.
[8] M. Osorno et al. (2011), Coordinated Cybersecurity Incident Handling - Roles, Processes, and Coordination Networks for Crosscutting Incidents [Online]. Available: http://dodccrp.org/events/16th_iccrts_2011/papers/189.pdf
[9] W. Koch et al., "The JDL Model of Data Fusion Applied to Cyber-Defence – a Review Paper", Workshop on Sensor Data Fusion - Trends, solutions, Applications. IEEE. 2012
[10] E. M. Hutchins et al., "Intelligence-Driven Computer Network Defense - Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains", Proc. of International Conference on Information Warfare & Security, pp. 113, March 2011
[11] G. Thibault, "Intelligence Collation in Asymmetric Conflict: A Canadian Armed Forces Perspective", Proc. IEEE FUSION, 2007
[12] The Joint Staff, United States Army, "Joint Publication 2-0: Doctrine For Intelligence Support To Joint Operations" [Online]. Washington D.C., 30 june 1991. Available: http://www.dod.mil/pubs/foi/joint_staff/jointStaff_jointOperations/

[13] F. Cheng et al.: RemodelingVulnerability Information; in F.Bao et al (Eds.) : Inscrypt2009, LNCS 6151, pp. 324-336, 2010 ; Springer Verlag Berlin Heidelberg 2010
[14] M A. Pravia et al.: “Generation of a Fundamental Data Set for Hard/Soft Information Fusion”, Proc. IEEE FUSION, 2008
[15] S. Dossé: Le Cyberespace, Nouveau Domaine de la Pensée Stratégique. p.119, Economica, 2013.
[16] A. Koltuksuz and S. Tekir, “Intelligence Analysis Modeling“, Proc. IEEE ICHIT, Vol. 1 , pp. 146 – 151, 2006
[17] B. Norquist, “Governmental Effects upon the Cyber Security Decision Making Cycle” , SANS Institute, White Paper, 2005
[18] “French White Paper on Defence and Security” [Online], 2013. Available: http://www.gouvernement fr/sites/default/files/fichiers_joints/livre-blanc-sur-la-defense-et-la-securite-nationale_2013.pdf
[19] United States Army Training and Doctrine Command, “ The United States Army’s Cyberspace Operatins Concept Capability Plan 2016 - 2028”, TRADOC Pamphlet 525-7-8; 2010.
[20] S. Jajodia et al. (Eds.), “Cyber Situational Awareness”, Advances in Information Security, Vol. 46, Springer Verlag, 2010
[21] S. Goel, “Cyberwarfare connecting the dots in cyber intelligence”, Communications of the ACM, Vol. 54, pp. 132-1408, august 2011
[22] J. Sander et al., “ISR Analytics: Architectual and Methodic Concepts”, Workshop on Sensor Data Fusion: Trends, solutions, Applications. Pp 99 – 104, IEEE, 2012
[23] R. Clarke and R. Knake, Cyber War, Harper Collins, New York, 2010
[24] E. Rosenbach and A. J. Peritz, “Confrontation or Collaboration? Congress and the Intelligence Community”; The Intelligence and Policy Project; Belfer Center for Science and International Affairs; John F. Kennedy School of Government, Harvard University; 2009
[25] F. Foch, Conférences faites à l’Ecole Supérieure de Guerre [Online], pp. 46, 1903. Available : http://gallica.bnf.fr/ark:/12148/bpt6k86515g
[26] J. V. Treglia, “Towards Trusted Intelligence Information Sharing”, Proc. ACM, SIG KDD, Workshop on CyberSecurity and Intelligence Informatics pp. 45-52, 2009
[27] S. Ritter, „Computernotfallteams: CERTs als zentrales Element nationaler Cyber-Sicherheit“, BSI Forum, Nr. 6, 20. Jahrgang, 2012

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Situational awareness and information collection from critical infrastructure

Jussi Timonen
Department of Military Technology
The Finnish Defence Forces
Helsinki, Finland
jussi.timonen@mil.fi

Samir Puuska
Department of Computer Science
University of Helsinki
Helsinki, Finland
puuska@cs helsinki.fi

Lauri Lääperi
Department of Military Technology
The Finnish Defence Forces
Helsinki, Finland
lauri.laaperi@mil.fi

Jouko Vankka
Department of Military Technology
The Finnish Defence Forces
Helsinki, Finland
Jouko.vankka@mil.fi

Lauri Rummukainen
Department of Military Technology
The Finnish Defence Forces
Helsinki, Finland
lauri.rummukainen@mil.fi

Abstract: Critical infrastructure (CI) is a complex part of society consisting of multiple sectors. Although these sectors are usually administered independently, they are functionally interconnected and interdependent. This paper presents a concept and a system that is able to provide the common operating picture (COP) of critical infrastructure (CI). The goal is to provide support for decision making on different management layers. The developed Situational Awareness of Critical Infrastructure and Networks (SACIN) framework implements key features of the system and is used to evaluate the concept.

The architecture for the SACIN framework combines an agent-based brokered architecture and Joint Directors of Laboratories (JDL) data fusion model. In the SACIN context, agent software produces events from the source systems and is maintained by the source system expert. The expert plays an important role, as he or she is the specialist in understanding the source system. He or she determines the meaningful events from the system with provided guidelines. The brokered architecture provides scalable platform to allow a large number of software agents and multiple analysis components to collaborate, in accordance with the JDL model. A modular and scalable user interface is provided through a web application and is usable for all SACIN participants. One of the main incentives for actors to provide data to the SACIN is the resultant access to the created COP.

The proposed concept provides improved situational awareness by modeling the complex dependency network within CI. The current state of the infrastructure can be determined by combining and analyzing event streams. Future states can be proactively determined by modeling dependencies between actors. Additionally, it is possible to evaluate the impact of an event by simulating different scenarios according to real-world and hypothetical use cases. As a result, understanding of CI and the ability to react to anomalies is improved amongst the decision makers.

Keywords: *Common Operating Picture, Critical Infrastructure, Situational Awareness, JDL data fusion model*

1. INTRODUCTION

This research presents the Situational Awareness of Critical infrastructure and Networks (SACIN) framework for gathering information from the different entities of critical infrastructure (CI). The main contributions of this paper are the created concept framework and the designed SACIN framework, including the implemented demonstration system. The framework provides tools for gathering information from CI, architecture for information fusion, and a user interface. Based on the derived information, it is possible to support decision making and expand the scope from situational awareness to a decision-making platform.

CI consists of a large number of different and constantly evolving source systems, which are impossible to integrate directly together. A big data system, where raw data from the source systems is gathered and analyzed, is not feasible in this context, because no single entity can understand the operation of all CI sectors. Additionally, most CI systems are privately administered and use equipment to which vendors are not usually allowing access. The solution for the system in this kind of environment is agent-based architecture, where some responsibility of the data integration is placed on the source system experts. The agent is a tool that is able to produce events from the system being monitored and to deliver them onwards. The autonomous agent enables information to be gathered from the source system without affecting the system being monitored.

Our approach is technical; first, we define the problem to be solved in chapter 1 and explore the prior research in chapter 2. In chapter 3, the concept framework is presented, and the architecture supporting the framework is studied in chapter 4. The designed agent component is presented in chapter 5 and the user interface in chapter 6. The empirical part of the study is the implementation discussed throughout chapters 3–6. Finally, in chapter 7, the results and future research are discussed.

2. RELATED WORK

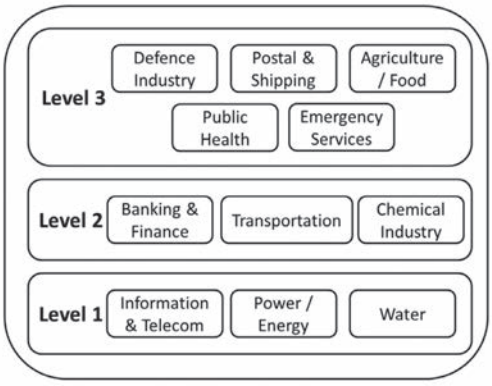
A basic information source on CI protection is the book by Lewis [1]. It presents a model of the sectors in CI and evaluates the threats faced. Modeling CI for use in different simulations is evaluated by Tolone et al. [2]. A project worth mentioning is the Executive order 13636 – Preliminary Cybersecurity Framework [3]. This order presents the basics for a risk based framework; its purpose is to unify and provide an improved understanding of the situation inside organizations. Wide-area situational awareness methodological framework is presented by Alcaraz and Lopez [4]. This research focuses on improving the situational awareness of CIs. An agent-based solution for modeling and simulation of interdependencies in the CI is presented by Casalicchio [5]. This study presents Agent-based Modelling and Simulation Framework, which is also implemented and tested. Attwood et al. present the Smart Cities Critical Infrastructure Response Framework [6]. This framework aims to provide an understanding of linked infrastructure and enable more efficient reactions on failing entities. The dependencies in CI are analyzed [7-10].

According to the literature review, there seems to be a lack of applying the Joint Directories of Laboratories (JDL) data fusion model with an agent-based solution to CI protection. Therefore, this paper combines these two approaches for the use of the common operating picture (COP) of CI. Additionally, the paper presents a concept framework that includes an implementation of the designed system.

3. CONCEPT FRAMEWORK

An important basis for the study is the taxonomy of CI defined by Lewis [1]. This taxonomy, presented in Figure 1, operates as a guide for dividing the entities in CI. Furthermore, the taxonomy provides a means to understand the interdependencies of objects in CI. The taxonomy is applied throughout the framework from low-level components to the COP. The taxonomy is complemented with event ratings [11] and event categories [12].

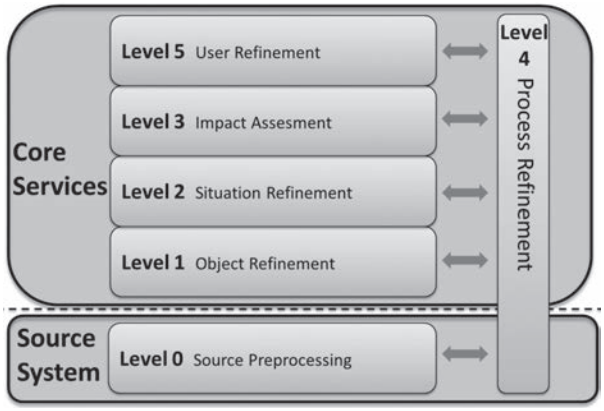
FIGURE 1 – SECTORS OF CI [1]



The vast amount of dependencies has an important role in the concept framework. The strength is in understanding the dependencies amongst the systems. For this purpose, the concept includes means to define and analyze the dependencies. The goal is to offer a source system in the CI means to share information and update the relations to the other entities.

The data fusion model used for SACIN is the JDL model, which presents a process supporting data collection and integration for the COP. The implementation of JDL model to cyberspace has been studied in [13, 14]. Challenges of information and data fusion in the context of urban operations are examined in [15-17]. Although the applied environment differs [15-17], the challenges in fusion are remarkably similar. The JDL model applied to SACIN is presented in Figure 2.

FIGURE 2 – JDL ADOPTED FROM [18]



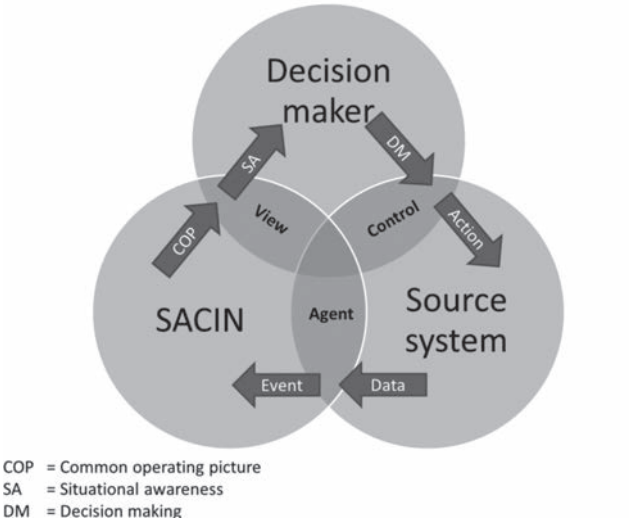
CI source systems with their monitoring components act as a sensor in the fusion process point of view. These systems are integrated with the SACIN through the agent software belonging to the JDL level 0. The purpose of the first, second, and third fusion levels is to analyze and form a model of CI in its current and future state. Analysis is initiated at level 1 by creating objects from the event stream. Objects can be created from just one significant event or from information gained through multiple events. For example, recognizing systematic port scans from multiple agents could create more serious reconnaissance objects. The aim for level 2 is to combine the information from the objects delivered from level 1 and construct the current state of the whole system. The acquired system state is then supplemented with the information at level 3. The focus on level 3 is the prediction the futures risks, possible vulnerabilities, and an estimation of their effects. Level 4 provides the ability for the system to control its operation through automated and user defined mechanisms. Finally, the COP is presented to the user with the level 5 user interface.

In Figure 3, the action flow in the concept framework is presented. The process starts from the source system, which provides data to the SACIN. Data collection is made possible by creating

an agent component, which can be deployed directly to the source system. The agent is able to connect to the SACIN and also extract important data from the source system. From these data, the agent produces events that are directed to the SACIN framework. From these events, the SACIN creates the COP according to the JDL model (Figure 2). The user interface supports the situational awareness of the decision makers at different levels.

In Figure 3, the decision maker includes authorities and source system operators. Authorities focus on maintaining society, whereas source system operators provide data to the SACIN and aim to improve their own processes. In the context of this study, a straight gateway for the means of effect (MoE) is not offered to the authorities' level, since the source systems are usually not owned or controlled by the high level decision makers. Control and action represent the communication between source system operators and authorities via every possible gateway (automated, email, phone, etc.).

FIGURE 3 – RELATIONS OF THE ENTITIES



An important part of the source system is the domain expert, who is responsible for understanding the state of the particular source system. The agents deployed to the source systems create and deliver the data forward. These data are aggregated using operators (human) and analyzers (automatic) to detect the relationships between the events based on dependencies, adding information from external systems, developing conclusions, and combining information. The COP is being created by SACIN, and the resolution is maintained all the way to the individual source systems and complemented with the aggregated information and dependencies. Source system-specific views are available amongst all actors. Furthermore, the COP is available in its entirety to the authorities. From this information, it is possible for a decision maker at the authority level to supplement one's situational awareness and use the desired MoEs. In Table 1, the different roles in decision making are presented.

FIGURE 4 – CREATION OF THE COP OF CI

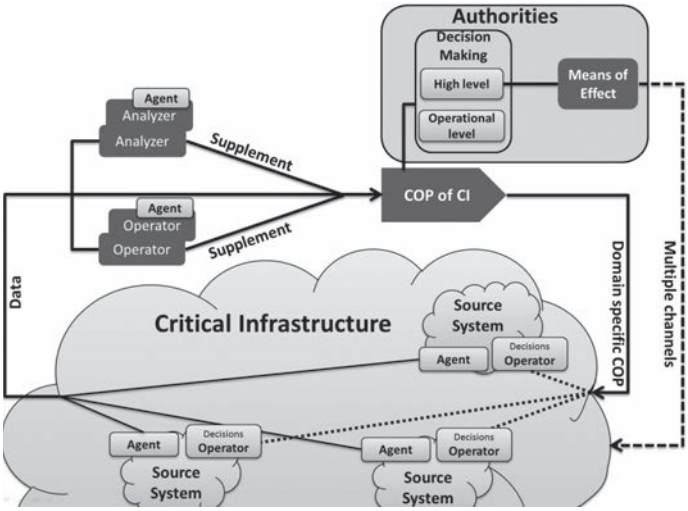


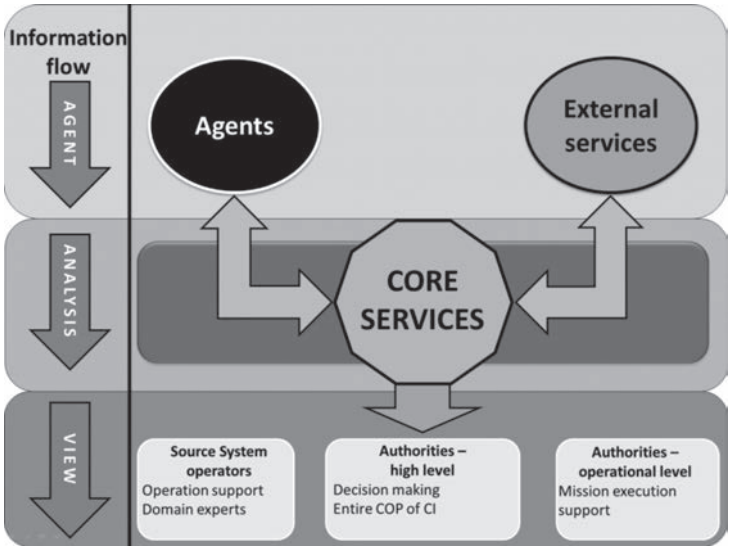
TABLE 1 – DECISION MAKING

Actor	Decisions	Means of Effect	Purpose for using SACIN	Examples of the entities
Source System operator	-Decisions effecting one's own system -Business-related decision.	-Internal means of the source system -Control over the own system	-Improved SA of the surroundings and connecting entities -Prediction and improving resilience of one's own business	Power grid operator, water supply company
Authorities high level	-Decisions effecting society as a whole -How to deal with and recover from a situation of crisis -Prediction and simulation of complex event chains	-Political -Military -Information sharing -Guidance -Preparation (emergency supply plans)	-To protect society from crisis situations -Improved recovery -To test the scenarios	Ministries, council of the state
Authorities operational level	-Decisions concerning one's own operations	-One's own operations	-To improve the efficiency and predictability of one's own operations	Police, fire department, rescue department

A SACIN core service (Figure 5) is the component where information is analyzed, stored, and organized. The agents are connected to the core services using a two-way communication

channel. The final layer from the perspective of information flow is the view, where the analyzed information is delivered from the core services to the user interface. The operators of the user interface are fundamentally the same as presented in Table 1. Information providers (source systems) are interested in the state of CI on which they are dependent.

FIGURE 5 - CONCEPT FRAMEWORK



4. SYSTEM ARCHITECHTURE

The main goal of the SACIN system architecture is to provide a platform supporting data integration and analysis of CI sectors. Different JDL data fusion processes should be supported to allow the integration of different CI systems. Scalability, closed systems, and data privacy are only a few requirements that lead to an agent-based integration approach. From the architecture point of view, the JDL model (Figure 2) and agent-based approach are the main influences regarding critical design choices.

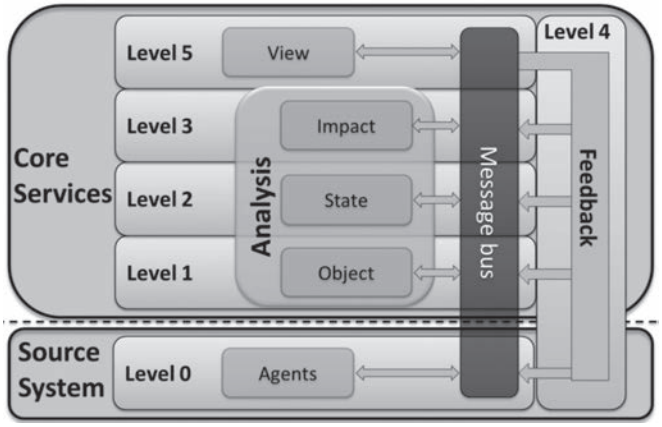
The JDL model itself does not take a stand on architectural decisions; it defines required steps the system must be able to offer. The architecture must accommodate all six data fusion sub processes and allow them to work together in a flexible and scalable way. The inter component communication channel is the key feature allowing operation in distributed environments and implementation on a national scale. Sufficient communication channels can be achieved with a common message bus.

Requirements for the common message bus are first, to have the capacity to handle large numbers of messages from multiple sources and second, to allow the routing of message to

one or multiple destinations. The first requirement is to allow a large number of agents to send information from their respective systems to analyzer components. The second one allows a flexible and scalable way for a component to communicate with any other one within the fusion chain. The role of the message bus is central to the functioning of the system. Therefore, it is important to be able to scale the capacity by distributing the load to multiple servers as well as to ensure service availability by duplicating the access points.

Figure 6 depicts a logical architecture diagram for the SACIN following the JDL model. All fusion sub processes are handled with respective components that communicate through the common message bus. Separation between domain and SACIN entities presents the administrative boundary between systems. The agent acts as a middle component between the separately administered source system and the SACIN. Event analysis is separated into three different components, which together, provide current and future states of CI. The analysis result is presented to the users through the view component in the form of the COP.

FIGURE 6 – JDL AND ARCHITECTURE



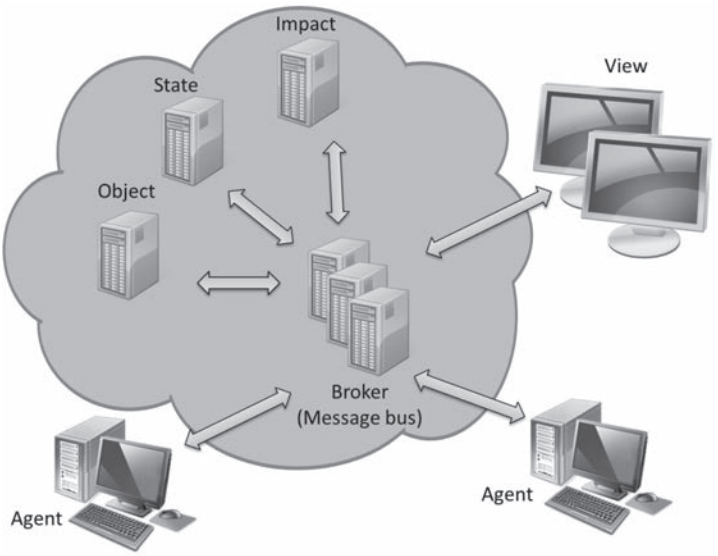
The message bus functionality can be achieved by various technologies, such as an enterprise service bus, a p2p network, or as a cloud service. The most important function of the message bus is to allow a large number of agents to send their events to the analyzers. Additionally, there may be separate analysis components that require the same streams through broadcasting. It is necessary to keep the system simple to manage, and the agent in particular should be able to run on low-end equipment.

A suitable technology implementing the message bus is brokered architecture from a cloud service point of view. The broker can be seen as a cloud service where a group of servers together offer message transfer services. Various services such as broadcasting and bi-directional messaging can be offered with little overhead. The same events can be directed to multiple destinations almost simultaneously. Additionally, as most of the communication between components is the “fire and forget” type, brokers can easily allow all inter component communication.

Although data fusion is the system's primary task, there are a few other topics that need to be addressed before the system is functional. The first and most important one is agent identification, which is required for separating and linking events to source systems. Because a large number of agents may be present, the ID pace should be large. Additionally, ids should be allocated randomly to make it more challenging to enlist brute force or guess used IDs. The second is the handling of user accounts that are used to operate within the system. User accounts are necessary for assigning ownership status to the agents. There needs to be a registrar component that is responsible for allocating and registering the agent ids as well as users to the system.

Figure 7 presents the interactions between different components. The broker acts as an intermediate service for routing messages between components. It does not orchestrate the operation in any way, but only allows inter component communication. All the operation logic and actions originate from the components and users. The broker, i.e., message bus, and other presented components together form a SACIN framework, which allows the integration of data from a separate CI sector. SACIN system components should be as separate and independent as possible. Each component should define an interface that other components are able to use through the message bus. Interfaces allow the addition of third-party services in the analysis chain. More sophisticated components complement the basic functionalities provided by SACIN. The primary messaging format between different components is an event. The agent component is presented in more detail in chapter 5 and user interface in chapter 6.

FIGURE 7 – BROKER



Scalability is a major requirement for the common operating picture system as the goal is to allow implementation on national scale. Therefore introducing new information sources, i.e. agents, to the system should increase resource requirements as little as possible. Networking

between agents and analysis components should be flexible enough to allow traffic load sharing between multiple servers.

In accordance to the JDL model, the agent is the interface between the source system and the SACIN. The level 0 source pre-processing allows the addition of new source systems that differ considerably from the other ones. Additionally, the agent acts as a low pass filter when it analyses and categorizes the source systems raw data. By reporting only relevant events to the SACIN the amount of transmitted data can be reduced greatly and not to overwhelm the broker servers. On average the expected amount of traffic from agents shouldn't be more than a few events per minute and a few events per second when certain incident occur.

Although core analysis components of the system are affected by the number of agents that produce events to the system they should not be the bottleneck of the system. As the JDL model levels 1 to 3 are all able to continue the filtering of the input data they can limit the traffic volume on such levels that the core services are not congested. Especially level 1 object refinement has an important role as it is the first analysis component handling the events. Although the filtering can reduce the load to other levels the level 1 must support load balancing to multiple servers. As the level 1 analysis focuses more on individual agents than dependencies between agents, it is possible to separate agent to groups that are handled by dedicated servers.

Analysis

As mentioned above, the analysis components produce events at object, state, and impact levels (see Figure 6). These follow the JDL data fusion model and handle the tasks defined in chapter 3. All analysis components are connected through the message bus and therefore can be distributed to separate servers. However, the state analyzers require access to the common database to achieve state for the whole system, and the impact analyzer requires access to the dependency information between different source systems.

Object

The object analyzer is responsible for handling the events that originate from agents. It analyzes the event streams and filters out the desired events. Additionally, object analyzer can detect and generate new events by combining information from different sources. For example, if level one analyser detects multiple port scanning operations in a given time frame which are directed to multiple agents in one sector or geographical location, a new event with greater severity can be generated to represent a possible network reconnaissance. Complex event processing techniques should be utilized in this analysis because the input is event stream [19].

State

The state analyzer forms states of all source systems based on object analysis events. Here the state is linked with agents and stored in the database. State information is constantly updated and new events are generated as the state changes. Additionally, current states of the agents can be queried through the message bus by other components. Severity of the events is largely assessed on the source system experts when they are defining how severely the detected event affects their own system operation.

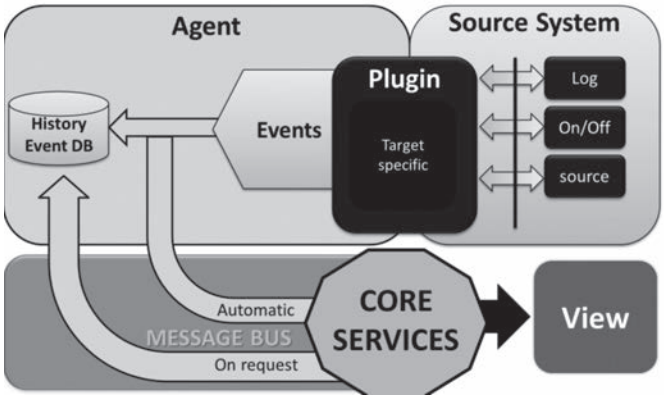
Impact
The impact analyzer focuses on determining the future state of the CI. Dependency information between different source systems, i.e., agents, is required for the analysis to allow various network analysis methods to be utilized. For example, vulnerability analysis can be performed to detect critical nodes or failure propagation throughout CI. Additionally, the alarms can be quickly propagated to specific systems to inform incidents such as telecommunication power outages.

5. AGENT

A SACIN agent (see Figures 4 & 7) is a middleware component designed to facilitate centralized event logging and analysis. All agents are assigned unique identifiers, which are used to separate them from each other within the SACIN framework. The purpose is to collect and log events from diverse sources and unify the event format for further analysis. A SACIN agent is designed to collect important status information from systems or processes that are part of CI. These systems can vary from industrial automation to custom intrusion detection systems. Because these systems have vastly different logging and error reporting capabilities, the middleware approach provides the needed flexibility between ease-of-use and wide compatibility.

Figure 8 describes the agent attachment to the source system. The actual event generation is done by the SACIN agent through a domain-specific software component called a plugin. This component will be built by a source system expert and it will take care of gaining and interpreting system incidents and providing events to the SACIN. The agent stores the events into a database, if allowed by the used platform, from which it is possible to collect the events for a more detailed analysis of the core services.

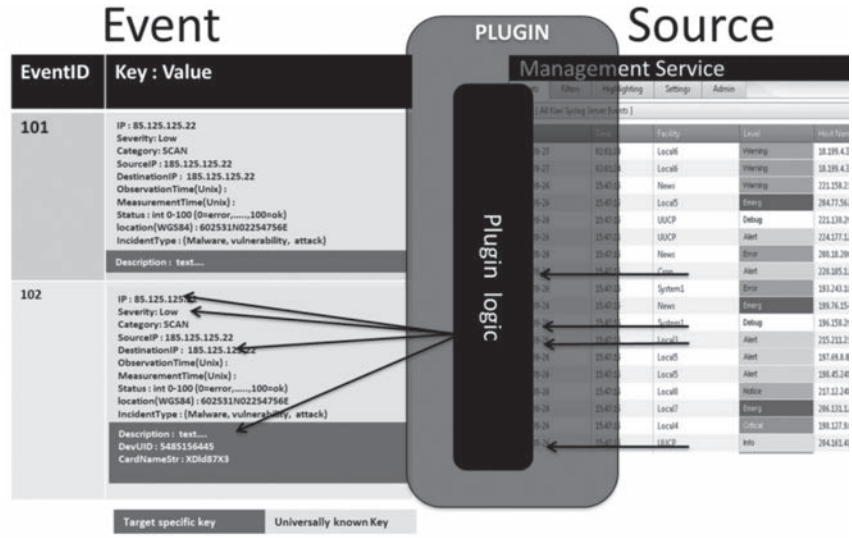
FIGURE 8 – AGENT



The most common place for the agent to be installed is a centralized component controlling a large entity from the same branch. The output of the system is used to understand the situation

of the source system. Figure 9 presents an example of the information flow when the agent generates events from a log file. From the perspective of SACIN, there is no direct visibility to the actual source system or the cause of event. The source system expert is offered tools and guidelines to be able to build the plugin. This means that the source system expert is, in fact, responsible for making system observations.

FIGURE 9 – PLUGIN



6. USER INTERFACE

The user interface of SACIN presents a way to visualize the COP. It serves as level 5 of the JDL model and tries to address user refinement issues, such as workload, visual attention, and particularly situational awareness, as presented by Blasch and Plano [20]. The user interface receives events of the infrastructure from the SACIN system back end. When new events are received, the user interface visualizes them into four different views that attempt to increase the situational awareness of operators of the SACIN system. As interpreted at JDL level 5 [20] and defined by Endsley [21], situational awareness is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." The views of the SACIN user interface attempt to cover all three aspects of this definition: the knowledge about all actors in the CI environment and their current and future states.

A general overview, presented in Figure 10, of the monitored infrastructure is provided to the operator as a quick way to check whether all parts of the infrastructure are working correctly. The operator has the ability to select the actors he wishes to monitor. The actors of the CI are divided into 11 different sectors according to the industry to which they belong. This

categorization follows the taxonomy presented by Lewis [1] (see Figure 1) plus one extra sector for actors that do not necessarily belong to any other sector. The current statuses of each sector are then visualized as six-segmented circles, as shown in Figures 10 and 11. These six segments represent the Federal Agency Incident Categories [22].

FIGURE 10 – OVERVIEW

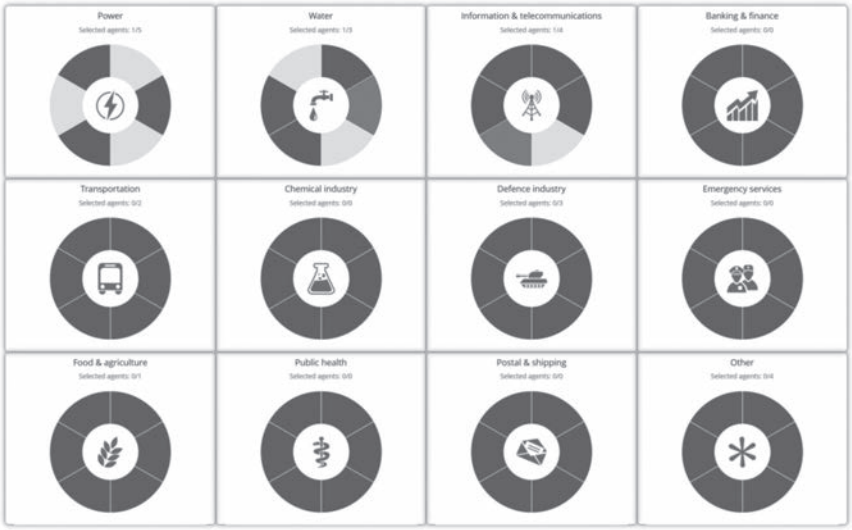
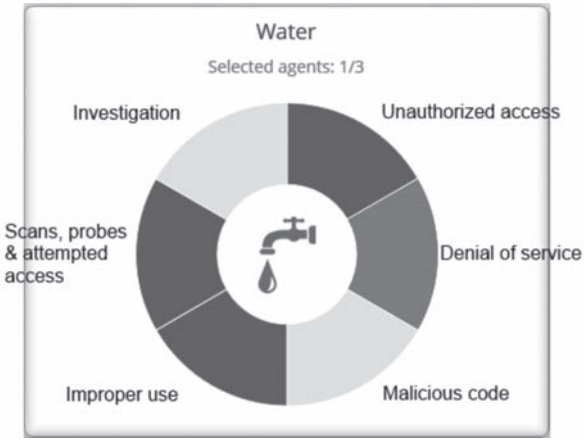


FIGURE 11 – STATUS CIRCLE IN FIGURE 10



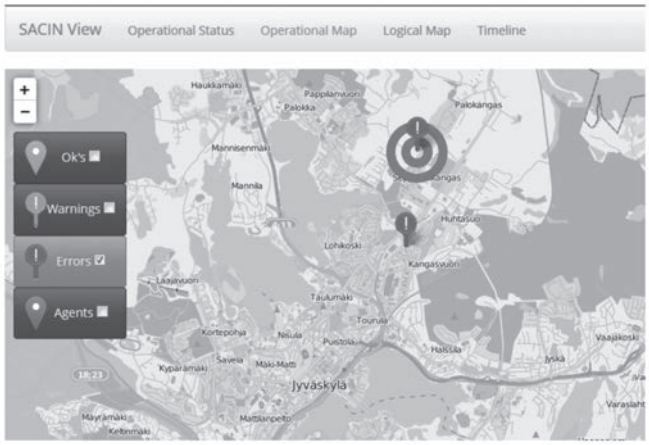
A timeline and a common event log, as shown in Figure 12, offer a temporal view for the operator to see when events have actually happened. This way the operator may, for example,

analyze consecutive events and link them together even if there are no indications of a relationship between the two in other views or external sources. This offers an advantage when doing risk analysis. Operators also use this view to receipt new events. This ensures that they have consciously seen all the events.

FIGURE 12 – TIMELINE

A map view, as shown in Figure 13, is offered to the operator so that the geographical distribution of agents and faults becomes clear. Regional events, such as floods, storms, or alike, may also be spotted on the map view. The implementation itself works as most contemporary map interfaces such as Google Maps. Operators also have the option to filter out types of events in which they are not interested.

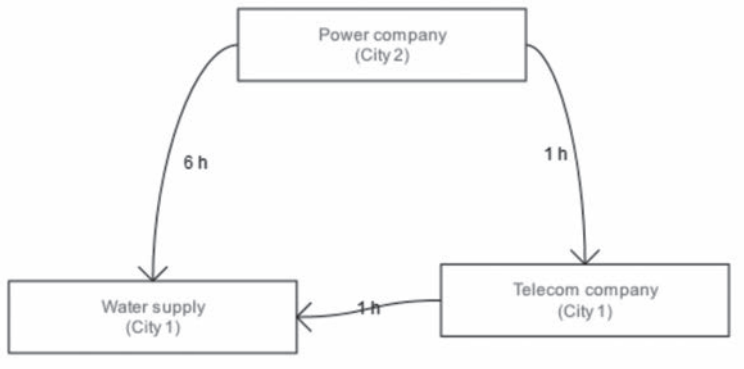
FIGURE 13 – MAP AND GEOGRAPHICAL DISTRIBUTION



Operators also need knowledge about potential escalating events that may occur. The logical view of the user interface incorporates logical dependencies between different actors in the CI, as shown in Figure 14. For example, a water supply company that is highly dependent on a power station may suffer from system failures due to power outages in the power station.

Because of this, an operator at the water supply company needs to know to which actors his company is dependent on and how fast faults will propagate. The dependencies are visualized in a simple directed graph that is drawn based on the selected actors. Each edge is accompanied with a time estimate that tells the operator how long the dependent can function normally without the other actor.

FIGURE 14 – LOGICAL MAP



As stated, these views try to support situational awareness, as operators are shown the current statuses of each industry, offered varied ways to see the events of the CI, and the dependencies between the actors are displayed so a projection of the future is possible. At an operator's workstation, these four views are positioned as shown in Figure 15. The layout is based on the idea that the most interesting view, the timeline, is placed in the center. The overview is placed on the left and the map on the right, so the general workflow supports the left-to-right type of reading. Ideally, an operator first checks the overview to see if everything is working correctly, continues to the timeline view to receipt new events, and finally, examines the map to look for regional events. The logical dependencies view is placed on top of the center view, as it is assumed to be used infrequently.

FIGURE 15 – DISPLAY LAYOUT

The usefulness and performance of the user interface was tested on several test sessions. During these sessions, test participants evaluated the usability of the system using the System Usability Scale (SUS) [23]. It was also tested on how well the system works in a real-life-like simulation using the Situation Awareness Global Assessment Technique (SAGAT) [24]. As a result, the overall SUS score was 71 on average, and all the error events were remembered and placed on a map with an average of 60% hit ratio. Test participants considered the timeline view as the most interesting of all four views. This was backed up by the fact that on average approximately 42 percent of total participant gaze time was focused on the timeline view. The user tests also raised a few issues about the necessary functions in the user interface such as the receipt functionality and the linkage of events between different views.

7. RESULTS AND FUTURE RESEARCH

In this paper, the authors presented a concept framework for creating a COP from CI. The implemented SACIN framework demonstrates the key features of the concept. The main contributions of this paper are the combination of the JDL model and the agent-based architecture, backed up by the implementation. In this paper we also present the results of the user tests carried out to the system operators.

Currently, the functionality corresponding the JDL model levels 0, 1, and 5 is being implemented, while other fusion levels are still in the early stages of development. In other words, events from source systems are created, categorized, rated based on their severity, and transmitted to the user interface. Analysis of the current and future states of the source system has still only been partially implemented.

Future research will focus on analyzing the dependencies and information flow to the system. At this time, SACIN does not implement the module for analysis, but there is an interface to attach the module. Similarly, the user interface will be a subject of further development. The usability tests for this paper were performed at the operator level. In future tests, the decision makers will be included in the testing to a greater extent. This will enable real-world scenario-based operations, as at this point the SACIN has the capability to reflect events from real-world data, in real time or simulated.

REFERENCES:

[1] T. Lewis, *Critical Infrastructure Protection in Homeland Security - Defending a Networked Nation*. Monterey, California: John Wiley & Sons Inc, 2006.

[2] W. Tolone et al., “Critical infrastructure integration modeling and simulation,” in *Intelligence and Security Informatics*, Berlin, 2004, pp. 214-225.

[3] The White House, “Executive Order - Improving Critical Infrastructure Cybersecurity,” Washington DC, 2013.

[4] C. Alcaraz. and J. Lopez, “Wide-area situational awareness for critical infrastructure protection,” in *Computer*, vol. 46, April, 2013, pp. 30-37.

[5] E. Casalicchio et al., “Federated agent-based modeling and simulation approach to study interdependencies in IT critical infrastructures,” in *11th IEEE International Symposium Distributed Simulation and Real-Time Applications (DS-RT 2007)*, Greece, 2007, pp. 182-189.

[6] A. Attwood et al., “SCCIR: Smart Cities Critical Infrastructure Response Framework,” *Developments in E-systems Engineering (DeSE)*,United Arab Emirates, 2011, pp. 460-464.
[7] Z. Liu and B. Xi, “COPULA model design and analysis on critical infrastructure interdependency,” in *International Conference on Management Science and Engineering (ICMSE)*, Melbourne, 2012, pp. 1890-1898.
[8] C. Wang et al., “National critical infrastructure modeling and analysis based on complex system theory,” in *First International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC)*, Beijing, 2011, pp. 832-836.
[9] R. Zimmerman, “Decision-making and the vulnerability of interdependent critical infrastructure,” in *International Conference on Systems, Man and Cybernetics*, Hague, 2004, pp. 4059-4063.
[10] R. Zimmerman and C. E. Restrepo, “Analyzing cascading effects within infrastructure sectors for consequence reduction,” in *IEEE Conference on Technologies for Homeland Security (HST 09)*, Washington DC, 2009, pp. 165-170.
[11] P. R. Garvey et al., “A macro method for measuring economic-benefit returns on cybersecurity investments: The table top approach,” in *The Journal of International Council on Systems Engineering*, vol. 16, no. 3, December, 2012, pp. 313-328.
[12] C. Stock and P. Curry, “MNE7 Collaborative Cyber Situational Awareness (CCSA) Information Sharing Framework,” 2013.
[13] S. Schreiber-Ehle and W. Koch, “The JDL model of data fusion applied to cyber-defence,” in *Workshop on Sensor Data Fusion Trends, Solutions, Applications (SDF)*, Bonn, 2012, pp. 116-119.
[14] G. P. Tadda, “Measuring performance of Cyber situation awareness systems,” in *11th International Conference on Information Fusion*, Köln, 2008, pp. 1-8.
[15] M. Bjorkbom, et al., “Localization services for online common operational picture and situation awareness,” *IEEE Access*, vol. 1, November, 2013, pp. 742-757.
[16] J. Timonen and J. Vankka, “Enhancing situational awareness by means of visualization and information integration of sensor networks,” in *Proc. SPIE 8756, Multisensor, Multisource Information Fusion Architectures, Algorithms, and Applications*, Baltimore, 2013.
[17] R. Virrankoski, “Wireless sensor systems in indoor situation modeling II (WISM II),” Proceedings of the University of Vaasa, Vaasa, Tech. Rep. 2013.
[18] N. A. Giacobe, “Application of the JDL data fusion process model for cyber security,” in *Proc. SPIE 7710 Multisensor, Multisource Information Fusion Architectures, Algorithms, and Applications*, Orlando, 2010.
[19] S. Vranes, et al., “Application of Complex Event Processing Paradigm in Situation Awareness and Management,” *22nd International Workshop on Database and Expert Systems Applications*, Tolouse, 2011, pp. 289-293.
[20] E.P. Blasch and S. Plano, “JDL level 5 fusion model: User refinement issues and applications in group tracking,” in *Proc. SPIE 4729, Aerosense*, 2002, pp. 270-279.
[21] M R. Endsley, “Toward a theory of situation awareness in dynamic systems,” *Human Factors*, vol. 37, no. 1, pp. 32-64, March, 1995.
[22] United States Computer Emergency Readiness Team (n.d.). Federal Incident Reporting Guidelines [Online]. Available: https://www.us-cert.gov/government-users/reporting-requirements
[23] J. Brooke, “SUS-A: A quick and dirty usability scale,” in *Usability Evaluation in Industry*, London, United Kingdom: Taylor & Francis, 1996, pp. 189-194.
[24] M R. Endsley, “Situation awareness global assessment technique (SAGAT),” *Aerospace and Electronic Conference (NAECON)*, vol. 3, pp. 789-795, 1988.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Operational Data Classes for Establishing Situational Awareness in Cyberspace

Judson Dressler
Department of Computer Science
Rice University
Houston, Texas, USA

William Moody
School of Computing
Clemson University
Clemson, South Carolina, USA

Calvert L. Bowen, III
Johns Hopkins University
Applied Physics Lab
Laurel, Maryland, USA

Jason Koepke
Towson University
Baltimore, Maryland, USA

Abstract: The United States, including the Department of Defense, relies heavily on information systems and networking technologies to efficiently conduct a wide variety of missions across the globe. With the ever-increasing rate of cyber attacks, this dependency places the nation at risk of a loss of confidentiality, integrity, and availability of its critical information resources; degrading its ability to complete the mission. In this paper, we introduce the operational data classes for establishing situational awareness in cyberspace. A system effectively using our key information components will be able to provide the nation's leadership timely and accurate information to gain an understanding of the operational cyber environment to enable strategic, operational, and tactical decision-making. In doing so, we present, define and provide examples of our key classes of operational data for cyber situational awareness and present a hypothetical case study demonstrating how they must be consolidated to provide a clear and relevant picture to a commander. In addition, current organizational and technical challenges are discussed, and areas for future research are addressed.

Keywords: *cyber situational awareness, cyberspace operations, operational needs*

1. INTRODUCTION

The critical computer networks of the United States play a key role in our everyday lives, controlling the nation's energy, transportation, and financial systems. As such, the Department of Defense (DoD) has built operational dependency on its information systems and their associated networks. Disruption of these networks would have significantly damaging effects on the United States' ability to operate and defend itself. With the constantly increasing rate

of cyber-attacks against our nation's network infrastructure and the ever-changing nature of computing, it is vitally important for the DoD to have an understanding of the cyber operating environment in order to properly secure and defend the nation.

More than a decade ago, Bass [1] observed that current intrusion detection technologies were not maturing at the rate of new attacks. Former Director of the National Security Agency (NSA), Mike McConnell, echoed this sentiment in February 2010 when he stated: "The United States is fighting a cyber-war today, and we are losing. It's that simple. As the most wired nation on Earth, we offer the most targets of significance, yet our cyber-defenses are woefully lacking" [2]. Commander, United States Cyber Command (USCYBERCOM) and Director of the NSA General Keith Alexander continued: "... to defend those networks and make good decision in exercising operational control over them ... will require much greater situational awareness and real-time visibility of intrusions into our networks" [3]. These concerns clearly identify the need for a comprehensive strategy to gain situational awareness over the cyber domain, which enables commanders at all levels to consider cyber as they make operational decisions and direct actions for their forces.

To successfully operate in the cyberspace domain, Cyber Situational Awareness (CSA) must be effectively enabled to empower commanders and government leaders to drive action and support rapid decision-making.

In this paper we propose six classes of data for establishing situational awareness in cyberspace. Section 2 provides background information and motivations for situational awareness. Section 3 describes related works in cyberspace research. We describe our data classes in Section 4 and present a case study in Section 5. Challenges to establishing cyberspace situational awareness are discussed in Section 6. Sections 7 and 8 present conclusions and areas for future research, respectively.

2. BACKGROUND AND MOTIVATION

Defining the term "situational awareness" is almost as hard as actually building situational awareness. United States Department of Defense joint doctrine does not define situational awareness in its Dictionary of Military and Associated Terms, JP 1-02, though situational awareness is used in the definition of four other terms: blue force tracking, common operational picture, United States Strategic Command's Global Network Operations Center, and national operations center. The closest definition in JP 1-02 was of "battlespace awareness", but it has been removed from the latest version.

> Battlespace Awareness - Knowledge and understanding of the operational area's environment, factors, and conditions, to include the status of friendly and adversary forces, neutrals and noncombatants, weather and terrain, that enables timely, relevant, comprehensive, and accurate assessments, in order to successfully apply combat power, protect the force, and/or complete the mission [4].

Since the DoD has established cyberspace as a warfighting domain, many aspects of that definition hold true in cyberspace. With the key being to enable commanders to issue orders to forces based on timely and accurate information. The ultimate goal of situational awareness in cyberspace is to maintain strategic and tactical understanding while continuously taking action or making operational risk decisions.

Achieving CSA has proven difficult to date. However, there are a series of issues to be addressed that will allow incremental progress towards CSA capabilities enabling any organization to harness the power of near real-time information supporting decision-making and proactive actions. Those issues include:

- Identification of what decisions and actions the organization may need to take with respect to cyber to assure operations can be sustained
- Identification of and access to the appropriate data that supports those decisions and actions
- Analytic tools to make sense of the presented data as it relates to operations
- Technology to consolidate and visualize data for decision makers at multiple levels within the organization

3. RELATED WORKS

Network defense, and in the military realm, information dominance have been hot topics over the last decade [5, 6, 7]. Computer systems have become fully integrated into our very existence, impacting how we live our lives. Research has been focused on defining cyberspace and developing innovative ways to defend it in the ever-changing cyber environment [8, 9, 10], including discussions focused on the unique challenge that most of the network infrastructure is a commercial product outside the control and protection of any one entity [9, 11, 12].

There has also been considerable investment into new hardware and software technologies for intrusion detection systems (IDS), host-based security systems, and anti-virus discovery mechanisms. IDS research has moved closer to the individual user and toward a behavioral based approach, as exemplified in [13, 14]. Automated responses have now been included in these detection tools to effectively shut down an attack once recognized by severing the connection or changing a rule. While progressing, these tools still suffer from a false positive problem which usually causes users to scale back the detection threshold.

Commercial visual analytic tools have been developed in an attempt to provide a CSA picture: IBM's Analyst's Notebook discovers patterns and trends across volumes of data to identify and predict malicious behavior; Palantir's toolset focuses on the fusion of disparate data sources into a unified picture for security analysis; and HP's Arcsite is a security information and event management system for enterprise-level IT architecture [15, 16, 17, 18]. Academic research has also developed visualization techniques in an attempt to provide an insight into the network, most using Ben Shneiderman of the University of Maryland's mantra of "overview first, zoom and filter, and then details-on-demand" [19, 20]. VisFlowConnect uses a parallel axes view

to the volume of network traffic in sender/receiver pairings over time; CNSSA incorporates information from multiple sources including current vulnerabilities to assign a vulnerability score based on the Common Vulnerability Scoring System; and SiLK provides analysts with the ability to understand, query, and summarize recent and historical network traffic data [20, 19].

Many publications in the last few years discuss security frameworks to gain insight into the situational environment [9, 21] and even more recently, the notion of tying network security to mission assurance [9, 22, 23]. In [15], the authors present a major task list that a cyber common operating picture must be able to complete as well as technological concerns in the developing of such a system; the Cyber Attack Modeling and Impact Assessment Framework [24] automates the development of attack graphs for computational analysis and impact assessment; and [25] argues effective policies for near real-time information sharing between multiple parties.

All of these ongoing studies and current analytical tools are inherently important to CSA and the discussion of the optimal way to achieve awareness of the cyber domain; however they do not address the fundamental building block of any situational awareness tool: the data. Our work's novelty springs out of this gap, discussing what classes of information are necessary and how each one builds upon the others to develop a holistic operational picture for establishing situational awareness in cyberspace.

4. CYBER OPERATIONAL DATA CLASSES

To achieve operationally relevant situational awareness of the cyberspace warfighting domain, a system must utilize six classes of information by fusing, correlating, analyzing, and visualizing in near real time. The six classes are as follows: 1) Current and near-future threat environment; 2) Global threats and significant anomalous activity; 3) Vulnerabilities of own computer systems and underlying infrastructure; 4) Prioritized cyber key terrain that allows understanding of operational and technical risks; 5) Current operational readiness and capability of its cyber forces and sensors; and 6) In-depth knowledge of ongoing operations and critical mission dependencies on its cyber assets.

As shown in Figure 1, the intersection of any combination of these classes provides more information and moves towards the sweet spot of SA. The factors from all six classes must be continuously assessed in order to provide a true, accurate and holistic representation of the domain which supports the ability to take critical actions and make decisions.

FIGURE 1. NOTIONAL INTERSECTION OF CLASSES OF INFORMATION REQUIRES CONTINUOUS ASSESSMENT TO PROVIDE CYBER SA AND ENABLE CRITICAL ACTIONS AND DECISIONS

A. Threat Environment

To successfully defend the network, an in-depth analysis of potential threats is crucial. This includes an understanding of who would want to attack the network, what goals are they looking to achieve, and how do they normally operate. A thorough knowledge of a threat's personality and normal behaviors will assist in identifying the threat's tactics, techniques, and procedures (TTP) and developing TTPs for network defense and incident response. Assessing an attack's vector in its early stages may reveal the attacker's capability and behavioral trends, leading to projections of future intrusion activities. This awareness can reap huge rewards in the protection from and reaction to a cyber attack. It also can be used to proactively align resources to counter future attacks using similar TTPs. Development of these adversary profiles could also lead to attribution in the event of an attack.

B. Anomalous Activity

Most networks have firewalls, anti-virus, and intrusion detection systems, which operate under pre-established rules or signatures, to detect or block when an anomalous activity occurs. These tools cannot respond to a zero-day exploit or a polymorphic virus because these events do not trigger the pre-established rules. Network and host-based IDS are essential to successfully defending the network. However, "IDS sensors can only capture systematic phenomena caused by attacks but cannot positively ascertain whether an attack has happened or succeeded" [5]. Baseline historical and current consolidated and normalized data must be incorporated into an automated system in order to understand what is "normal" and what is "anomalous" then take actions to effectively defend against cyber threats represented by this activity.

C. Vulnerabilities

From 2006 to 2011, over 75 thousand new security vulnerabilities were discovered [26]. Vulnerabilities are present in every system no matter how secure the system claims to be. Technology advances so rapidly that it can be virtually impossible to eradicate vulnerabilities altogether. The best one can hope for, in many cases, is simply to minimize them. In order to

assess and minimize the risk to the network, vulnerabilities of the systems and the underlying infrastructure must be known. System administrators and security specialists must have the knowledge and tools to understand the vulnerabilities of their networks and to properly test any new system or application before applying it to the network. Most importantly, these vulnerabilities must be known and continuously assessed. Leadership must be willing to allocate funds for vulnerabilities to be found and fixed.

D. Key Terrain

Though a single organization may have tens of thousands of systems ranging from desktops and mobile devices to routers and switches spread geographically across the world, not all systems have equal criticality to mission success. Defending and garnering full knowledge of all systems, accounts, and processes on the network in real time is impractical. Therefore, it is necessary to identify and prioritize key cyber assets to allow the understanding of critical risks both operationally and technically. Identification of cyber key terrain includes all critical information, systems, and infrastructure; whether owned by the organization or used in transit by its information [27]. That said, even these systems must be prioritized and may be less vital than a specific network link supporting a real-time airborne mission. The identification allows for prioritized defense of assets but cannot fail to consider all systems and assets in the network.

E. Operational Readiness

Organizations must know the operational readiness and capability of their cyber forces and assets. This includes the status of its tools and capabilities along with the ability of its cyber forces to protect its networks. Understanding the training status of all personnel to operate in the current threat environment and the readiness and integrity of network sensors, paths, and systems is critical. A real-time status of the network and personnel resources provides data necessary to recognize an attack and align resources which are available to appropriately respond. Mission impact is another aspect of operational readiness which is often hard to define and keep up to date. For a situational awareness picture to truly be useful, it must be operationally relevant and actionable. For this to occur, an organization must have a thorough understanding of mission dependencies based on cyber assets. With the knowledge and prioritization of intermission and mission-system dependencies, the organization can now depict to leadership the impact of a cyber event, whether an outage or attack, and the significance of securing certain assets [9, 22].

F. Ongoing Operations

Lastly, information about the status of all ongoing operations (cyber, kinetic, and even diplomatic) must be fully understood by commanders at all levels. This knowledge could be used to deconflict controlled outages or upgrades to systems that are currently engaged in support of an operation. It could also be used to dynamically identify key terrain and adjust defensive TTPs during the operational window of time. Understanding which operations are being executed or soon to begin execution, allows commanders to reallocate assets as necessary to support those operations. In addition, this allows leaders to understand the operational impact of systems and their critical operational dependencies.

5. AN OPERATIONAL CASE STUDY

A hypothetical operational case study is presented in order to emphasize the value of holistic fusion of data from all six classes. In this case study, we introduce a commander and staff whom are initially presented data from the ongoing operations, key terrain, and operational readiness classes. We will show the improved situational awareness opportunities to impact the commander's decision-making process as additional information classes are considered.

A US Joint Task Force (JTF) is currently conducting combat operations in an area of operations that requires the continuous flow of logistical and personnel resupply. In the operational planning process, the commander has designated his logistical support information systems as cyber key terrain. These systems operate on an unclassified military network so they can receive updates from commercial shipping and airflow systems on the Internet. The JTF commander also is aware that the network sensors deployed to protect these logistical systems are degraded due to required maintenance upgrades. The upgrades are currently scheduled for implementation by a computer network defense service provider (CND-SP) stationed in the continental United States during the next month. Lastly, the commander has an extremely proficient cyber investigative and forensics unit attending commercial certification refresher training. With this partial set of information, the commander has a good baseline of situational awareness of cyber assets and how they may impact his operations across all warfighting domains.

During the course of operations, a critical vulnerability in the outdated operating system of the logistical support system is discovered. As a DoD program of record, the potential patch for this vulnerability remains in pre- deployment testing and is not scheduled for release for another 30 days. USCYBERCOM has assessed the vulnerability and issued a high priority message across the DoD cyber enterprise announcing the details of the vulnerability. This vulnerability allows root-level access to be gained on the systems potentially enabling the deployment of malicious software on all unpatched systems. The commander is advised of the potential impact to his key logistics systems, but decides to take no action based on requirements for the continued flow of supplies and personnel supporting his operational mission set.

When the intelligence officer advises the commander on a new cyber threat report, an additional class of data (Threat Environment) is fused with the current understanding of the battlespace. In this report, it is assessed that the adversary has ever-increasing interest in disrupting and influencing the logistical flow of forces and supplies into theater. Additionally, supporting cyber assets are known to deploy Trojan-horse software on susceptible systems. This additional information of the threat environment improves the commander's understanding of the cyber environment and drives him to take decisive action to ensure his combat power will be available at the critical point in his operations. He directs his cyber force to cease with their commercial training and refocus their efforts on monitoring the behaviors of his logistical support platforms.

While reviewing the network flow and log data from the logistical system, the team discovers information included in our last class, Anomalous Activity. More than half of the logistical support systems supporting the JTF have been sending irregular sized traffic over TCP port 443 to a subnet outside of the United States. Further forensics work determines documents have

been slowly exfiltrated via covert encrypted and unencrypted channels. The commander is now alarmed and initiates crisis action planning. He directs the stateside CND-SP to immediately upgrade the defensive sensors and remove the logistics systems from the network until appropriate countermeasures can be deployed to protect the systems until the patch becomes available. Further, he requests intelligence and cyber forensics support to determine which files were stolen and the potential operational impact of their loss. Now that he does not fully trust his logistics systems' information, considering future shipping schedules were the exfiltrated files, he reallocates air and naval assets to protect inbound shipping containers to protect his logistical lines of communications. Lastly, he directs his cyber forces to begin detailed log review with daily update briefings.

This case study portrays an environment where all SA information classes have an abundance of data available for consumption by an integrated system or motivated person able to fuse them together to provide the opportunity for total situational awareness. This is not today's reality. Cyber forces rarely track or concern themselves with the status of ongoing operations across all warfighting domains. Strategic and operational commanders do not know or fully understand how to determine their cyber key terrain. If they do, typically, they have not taken the required actions or time to determine and designate cyber key terrain. Additionally, the operational readiness of cyber forces is not well defined or tracked at the level needed to fully understand capabilities and how it could impact operations. In contrast, vulnerability, threat and anomalous activity data is plentiful within the intelligence and cyber communities. That said, the data is often presented to the commander in a way that information overload or technical jargon routinely make it difficult for the commander to assess the value of the information and therefore the information is discounted or ignored. Other challenges that inhibit today's ability to gain, maintain, and adjust the fusion of information that can provide SA to the commander are described in the next section.

6. CURRENT CHALLENGES

Effective Cyber Situational Awareness requires that data and information be collected, analyzed, and displayed to the end customer in a timely and relevant manner. Although numerous challenges exist, the key barrier to successful implementation and execution of enterprise-wide CSA is solving the following organizational and technical challenges.

A. Organizational Fear

Gaining access to all of the necessary network data within different aspects of an organization can lead to a turf war. No entity wants to give up access to their data due to fear. Fear of humiliation in publicizing security flaws, fear of losing a competitive edge or public confidence, or fear of the proverbial 1,000 mile hammer. Regardless of the reason, this fear prevents complete situational awareness. To combat this fear, the United States Department of Defense must define and enforce a single information owner who can aggregate this data for analysis.

B. Data Consolidation & Normalization

Data comes in the form of technical and human collections, including IDS, network sniffers, and computer system log files. Ingesting all of the data is currently impractical but may soon become reality due to the advancement of cloud computing and the ever increasing data transfer rates. Determining the proper metrics and alert thresholds for the organization are essential for real time analysis. The data from these sources needs to be consolidated and put into a normalized format in order to be properly ingested into a CSA tool. Data refinement is simplified when a common format exists and requires a temporal calibration of the different data streams [1].

C. Data Synthesis

Currently, stove-piped data synthesis solutions exist across different parts of organizations that were developed separately over time without a clear coordinated cyber strategy. The challenge arises with how to fuse the data together. The fusion process requires the utilization of processing algorithms, such as Sudit's and Stotz's INFERD system, and comparison with known statistics (from USCERT, MacAfee, Norton, etc) to assess evolving situations and threats in cyberspace [28]. This data synthesis is needed for a full understanding of the normal state of the network, allowing security to move away from signature-based toward true anomaly-based detection. Intruders executing stealth TCP-based attacks on multiple geographically-separated parts of a corporate network may fall below the pre-established security thresholds. A common situational awareness tool which ideally includes all six classes of information may be able to synthesize the data and combine disparate attacks which may paint the picture of a coordinated and sophisticated enemy [28, 29].

D. Result Visualization and Dissemination

Until intrusion detection becomes truly machine to machine automation that responds immediately to anomalous activity, human intervention will require rapid understanding by presenting data in a visual manner. In the traditional warfare domains, situational awareness was represented geospatially on a map. Military leadership is used to this representation of disposition of forces, but this depiction does not always fit well within the cyber realm. Visualization systems need to be much more than PowerPoint presentations and bar charts; however, 2D systems such as parallel axes, logical maps, and temporal visualization of packet flows are limited in their ability to represent all the data attributes in one view. In addition, situational awareness visualizations must be able to illustrate mission impact to truly have meaning to leadership. A dissemination plan must also be established for the actionable results as not all information is appropriate for all personnel. Attributes that clearly identify the mission authorities and identity of the user can be used to present the appropriate data to each user.

E. Timeliness

As the amount of data, rules and signatures increase, analysis accuracy decreases and false positives increase, hampering timely detection and response. Cyber attacks occur frequently and can cause debilitating effects within milliseconds. To combat this, a finely tuned advanced threat detection engine must be used in conjunction with the known normal state to ensure the broadest possible spectrum of threats are identified and to eliminate false positives as much as possible. The challenge pivots on the ability to summarize vast amounts of information at the appropriate level and then provide it to operators at the appropriate levels in a timely fashion.

7. CONCLUSION

The United States' reliance on computer networks is undeniable, and there will never be an impervious defense to all network attacks. Thus, robust situational awareness of the cyber environment, detailing what is happening, where, and what are the best available response options is absolutely critical to operations. In this paper, we developed a new approach for decision makers to assist in rapid decision making. We introduced six classes of information necessary (threat environment, anomalous activity, vulnerabilities, key terrain, operational readiness and ongoing operations) to effectively enable and empower commanders and government leaders to incorporate cyberspace into the decision making process. This data must be continuously analyzed to provide a true and accurate representation of the domain.

However, there still remain many challenges that must be addressed before situational awareness in cyberspace may be obtained. This paper has identified the decisions and actions the United States must take with respect to cyber, whether it be analytic tools to correlate the presented data to an operation or the technology to consolidate and visualize data for decision makers. Once addressed, the operational view of cyberspace can move from one of network assurance to a true mission assurance focused situational awareness picture.

No effective and exhaustive solution exists for recognizing the majority of cyber attacks before they occur and cause damage. With the speed of attack achievable in cyberspace, a fully developed cyber situational awareness picture is as close to an early warning system as one can achieve. Therefore, the challenges must be overcome, and situational awareness in cyberspace must be realized to enable proactive, agile, and successful network defense for the United States.

8. FUTURE WORK

The classes of data introduced in this paper are based on the authors' intensive operational experience working at the highest levels of command in the area of cyber situational awareness for the U.S. Department of Defense. Though the authors have traveled the world talking about Cyber SA to senior leaders in multiple organizations across the Department, experimentation and prototyping of systems uses these classes is necessary to fully validate the claims.

Several key aspects of attaining situational awareness are still not well defined. Every organization depends on cyber assets to accomplish their mission. These assets can encompass thousands of computer systems, network sensors, and personnel spread across the globe. An efficient method for determining cyber key terrain to assure mission accomplishment has yet to be found.

As networks expand and data rates continue to soar, working with massive datasets in real time is becoming more common. More research is necessary in taking sensor event data, efficiently storing and correlating it to mission impact, and then disseminating it in a timely manner to

enable leadership to make better decisions. The advent of cloud computing may make this more achievable.

Many advances are being made in general data visualization techniques. The conventional SA tool displays network events on a geo-referenced map of the network. This method works well for battlefield awareness in ground, naval, and aerial assets, but may not be the best way to view cyberspace based on interconnections that defy geographic boundaries. Other visualization techniques need to be developed which allow SA at various levels to inform the commanders for leadership decisions and the net defenders or system administrators for decisive actions at the operator or analyst level.

REFERENCES:

[1] T. Bass, "Intrusion Detection Systems & Multisensor Data Fusion: Creating Cyberspace Situational Awareness," *Communications of the ACM*, 2000.

[2] M. McConnell, "Mike McConnell on How to Win the Cyber-War We're Losing," *Washington Post*, 28 February 2010.

[3] K. Alexander, "Advance Questions for Lieutenant General Keith Alexander, USA Nominee for Commander, United States Cyber Command". *Washington Post*.

[4] Department of Defense, "Joint Publication 1-02 Dictionary of Military and Associated Terms," 2010.

[5] J. Li, Z. Ou and R. Rajagopaian, "Uncertainty and Risk Management in Cyber Situational Awareness," *Cyber Situational Awareness*, 2010.

[6] C. Croom, "The Defenders 'Kill Chain'," *Military Information Technology*, vol. 14, no. 10, 2010.

[7] K. Deutsch, "Importance of Information Dominance," *Military Information Technology*, vol. 14, no. 10, 2010.

[8] L. Stovall, "People, Processes and Technology," *Military Information Technology*, vol. 14, no. 10, 2010.

[9] L. Cumiford, "Situational Awareness for Cyber Defense," in *2006 CCRTS The State of the Art and the State of the Practice*, 2006.

[10] S. Jajodia and S. Noel, "Topological Vulnerability Analysis," in *Proceedings of the Army Research Office Cyber Situational Awareness Workshop*, 2009.

[11] P. Cuviello and B. Kobel, "Cyber-Awareness is a Team Sport," *Military Information Technology*, vol. 14, no. 10, 2010.

[12] K. Condello, "Working Together for Real-Time Awareness," *Military Information Technology*, vol. 14, no. 10, 2010.

[13] R. Koch and M. Golling, "Architecture for Evaluating and Correlating NIDS in Real-World Networks," in *5th International Conference on Cyber Conflict*, Tallinn, 2013.

[14] O. McCusker, S. Brunza and D. Dasgupta, "Deriving Behavior Primitives from Aggregate Network Features Using Support Vector Machines," in *5th International Conference on Cyber Conflict*, Tallinn, 2013.

[15] G. Conti, J. Nelson and D. Raymond, "Towards a Cyber Common Operating Picture," in *5th International Conference on Cyber Conflict*, Tallinn, 2013.

[16] IBM, "Analyst's Notebook," [Online]. Available: http://www-03.ibm.com/software/products/en/analysts-notebook-family/. [Accessed 6 February 2014].

[17] Palantir, "Palantir," [Online]. Available: https://www.palantir.com. [Accessed 6 February 2014].

[18] HP, "Security Information and Event Management," [Online]. Available: http://www8.hp.com/us/en/software-solutions/siem-arcsight/. [Accessed 6 February 2014].

[19] R. Xi, S. Jin and X. Yun, "CNSSA: A Comprehensive Network Security Situational Awareness System," in *IEEE 10th International Conference on Trust, Security, and Privacy in Computing and Communications (TrustCom)*, 2011.

[20] X. Yin, W. Yurcik, L. Yifan, K. Lakkaraju and C. Abad, "VisFlowConnect: Providing Security Situational Awareness by Visualizing Network Traffic Flows," in *IEEE International Conference on Performance, Computing, and Communications*, 2004.

[21] S. Batsell, N. Rao and M. Shankar, "Distributed Intrusion Detection and Attack Containment for Organizational Cyber Security," 2005. [Online]. Available: http://www.ioc.oml.gov.

[22] W. Heinke, "What Commanders Need to Know," *Military Information Technology*, vol. 10, no. 14, 2010.
[23] M. Gregoire and L. Beaudoin, "The Science of Mission Assurance," *Visualization and the Common Operational Picture*, 2005.
[24] I. Kotenko and A. Chechulin, "A Cyber Attack Modeling and Impact Assessment Framework," in *5th International Conference on Cyber Conflict*, Tallinn, 2013.
[25] D. Ferandez Vazquez, O. Pastor Acosta, S. Brown, E. Reid and C. Spirito, "Conceptual Framework for Cyber Defense Information Sharing Within Trust Relationships," in *4th International Conference on Cyber Conflict*, Tallinn, 2012.
[26] B. Casey, "The IBM Institute for Advanced Security Expert Blog," IBM, 31 March 2011. [Online]. Available: http://www.instituteforadvancedsecurity.com.
[27] T. Pingel, "Key Defensive Terrain in Cyberspace: A Geographic Perspective," in *Proceedings of the International Conference on Politics and Information Systems (PISTA)*, Orlando, 2003.
[28] M. Sudit and A. Stoltz, "Information Fusion Engine for Real-time Decision-making (INFERD): A Perpetual System for Cyber Attack Tracking," in *10th International Conference on Information Fusion*, 2007.
[29] S. Yang, S. Byers and J. Holsopple, "Intrusion Activity Projection for Cyber Situational Awareness," 2008. [Online]. Available: http://www.ieeexplore.ieee.org.

Chapter 4
Detection and Deception

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Towards Multi-layered Intrusion Detection in High-Speed Networks

Mario Golling
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
mario.golling@unibw.de

Robert Koch
Faculty of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
robert koch@unibw.de

Rick Hofstede
Design and Analysis of
Communication Systems (DACS)
University of Twente
Enschede, The Netherlands
r.j hofstede@utwente nl

Abstract: Traditional Intrusion Detection approaches rely on the inspection of individual packets, often referred to as Deep Packet Inspection (DPI), where individual packets are scanned for suspicious patterns. However, the rapid increase of link speeds and throughputs – especially in larger networks such as backbone networks – seriously constrains this approach. First, devices capable of detecting intrusions on high-speed links of 10 Gbps and higher are rather expensive, or must be built based on complex arrays. Second, legislation commonly restricts the way in which backbone network operators can analyse the data in their networks. To overcome these constraints, flow-based intrusion detection can be applied, which traditionally focuses only on packet header fields and packet characteristics. Flow export technologies are nowadays embedded in most high-end packet forwarding devices and are widely used for network management, which makes this approach economically attractive.

In the context of large, high-speed networks, such as backbone networks, we make two observations with respect to flow-based and packet-based intrusion detection. First, although flow-based intrusion detection offers several advantages in terms of processing requirements, the aggregation of packets into flows obviously entails a loss of information. Second, the quantity of information is not constrained when packet-based intrusion detection is performed, but its application is often unfeasible, due to stringent processing requirements. To bridge this gap, we propose a multi-layered approach that combines the advantages of both types of intrusion detection. Our approach is centred around the idea that 1) a first layer of detection comprises flow-based intrusion detection, that makes a pre-selection of suspicious traffic, and 2)

additional packet-based intrusion detection is subsequently performed on a pre-filtered packet stream to facilitate in-depth detection. We demonstrate how this approach avoids the problem of a costly infrastructure, and obeys the various legal barriers on network traffic inspection.

Keywords: *Network Security, Intrusion Detection, High-speed Networks, Flow-Based Intrusion Detection, Legal Inspection*

1. INTRODUCTION

Network attacks have always been present since the birth of the Internet, but high link speeds and the ease of performing and participating in attacks have made this problem the order of the day. Internet insecurity is a worldwide problem that has generated a multitude of costs for businesses, governments, and individuals. When attacks are performed in a distributed fashion, their devastating power can easily overwhelm individual end hosts. For example, the Spamhaus project was targeted by Distributed Denial of Service (DDoS) attacks in early 2013 with more than 300 Gbps of traffic, enough to overload several Internet exchanges [1]. Throughout the last couple of years, in addition to DDoS attacks, in particular worms and botnets also represent special challenges for network operators, since they also tend to consume a great amount of resources [2-5].
To approach the detection of attacks, one of the well-established security solutions nowadays are Intrusion Detection Systems (IDSs). Intrusion detection is defined by the National Institute of Standards and Technology (NIST) as follows [6]:

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analysing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices.

Intrusion detection is usually been performed based on packet payloads. This approach, commonly referred to as Deep Packet Inspection (DPI), provides full visibility in the network traffic, which comes at the expense of scalability. As soon as intrusion detection has to be performed on links with speeds of 10 Gbps and higher, more complex/expensive hardware is needed to cope with the large amount of traffic.
To overcome the scalability problem of packet-based/payload-based intrusion detection, flow-based intrusion detection has been extensively researched [7]. This type of intrusion detection is performed on traffic aggregates, rather than individual network packets, but accuracy and detail are sacrificed for the sake of scalability. Many network operators have flow monitoring facilities at their disposal [8], so deploying them comes at almost no cost. We therefore consider flow-based intrusion detection a viable approach for operators of high-speed networks.
In this paper, we present an approach that exploits the advantages of both, flow-based and packet-based intrusion detection and overcomes many legal obstacles by operating in a multi-layered fashion; we use flow-based intrusion detection as the first layer of detection for identifying potential incidents, while more detailed intrusion detection is used as the second

stage for analysing only the part of the traffic stream that has been reported as suspicious by the first stage. In particular, we focus on backbone networks as a typical example of high-speed networks.

The remainder of the paper is structured as follows. In Section 2, we discuss background information in the field of intrusion detection. An example scenario that highlights the context of this work is described in Section 3. Our multi-layered architecture is discussed in Section 4, followed by an in-depth discussion of how the various architectural components are managed in Section 5. In Section 6, we discuss our first thoughts on the implementation. Section 7 provides an insight on our ideas regarding the evaluation. Finally, we close this work in Section 8 by discussing the next steps to be taken.

2. BACKGROUND

In this section, existing approaches to intrusion detection are briefly introduced. To be able to classify individual systems, we start by presenting a classification scheme for IDSs, which will serve as a basis to classify and evaluate existing approaches according to these criteria.

A. Classification Schemes for Intrusion Detection

Due to the fact that IDSs have been an active research area for several decades, quite a lot of work has been done on the classification of these systems. A classification or taxonomy is a hierarchical structure of a field of knowledge into groups [9]. Here, several properties have to be satisfied (see e.g., [10, 11]): mutual exclusiveness, completeness, traceability, conveniently, clarity and acceptance. However, no generally accepted taxonomy is available for the classification of IDSs and various classifications of very different levels of detail can be found in the literature [9]. The taxonomy published by Debar et al. [12, 13] is used widely [9]. Next to Debar et al., the taxonomy of Axelsson [14] is also generally considered to be a main contribution in this area [7]. In the following, we will briefly describe selected elements of Debar et al. and Axelsson, which are generally used to classify intrusion detection approaches [9]:

Detection Method: With regard to detection, three approaches can be distinguished [6]:

- *Knowledge-based techniques* are based on the idea of comparing currently observed activities (e.g., packets that pass the IDS) to investigate and examine them for the presence of already known attack traces (e.g., using string comparison operations).
- *Behaviour-based techniques* describe the process of comparing definitions of what activities are considered normal with the current events observed to identify significant deviations, using models to predict the expected state of a system. If the predicted and the measured value differ more than a specified threshold, an alert is raised.
- *Compound:* There are also approaches that form a compound decision in view of a model of both, the knowledge-based approach as well as the behaviour-based approach.

Behaviour on Detection/Response: If an IDS does not only monitor events and analyse them for signs of possible incidents, but also attempts to stop detected incidents, it is commonly referred to as an Intrusion Prevention System (IPS) [6]. IDSs are therefore considered as passive, while IPSs are considered reactive.

Audit Source Location: IDSs/IPSs can also be classified based on the audit source location. Although not in the scope of this paper, host-based IDSs, which monitor the characteristics of a single host and the events occurring within that host for suspicious activity, are also one way to classify IDSs. As this publication is focussing mainly on backbone network operators, in the following we focus on network-based IDSs, which monitor network traffic for particular network segments or devices and analyse the network and application protocol events to identify suspicious activities [6].

Time of Detection: Three main classes can be identified. Attempts that perform detection (i) in real-time or (ii) near real-time, and those that process data with a considerable delay, postponing detection; (iii) non-real-time.

Link Speed: This categories indicates whether an approach is able to work in high-speed environments. With this paper, connections of around 1 Gbps are considered as low link speed, whereas high-speed links usually have a data rate of 10 Gbps and higher.

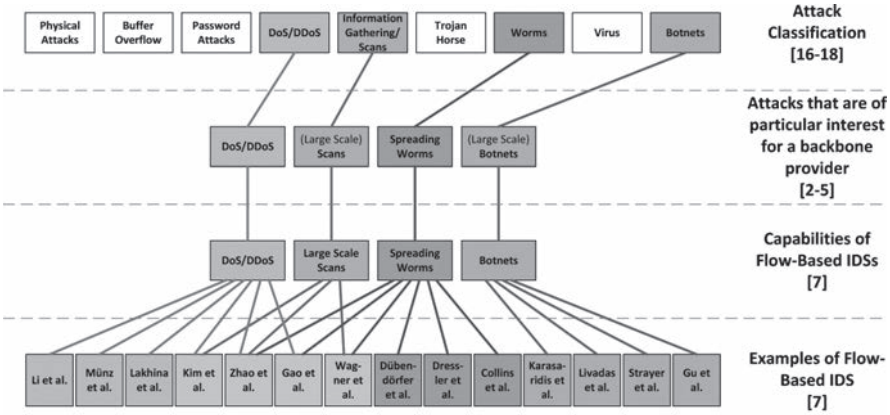
Layer of Detection: Although not considered by Debar et al. and Axelsson, IDSs can also be distinguished based on the layer on which the detection is performed. Header-based IDSs consider only header information, while payload-based IDS investigate both the header and the payload of a packet.

B. Overview of Existing Approaches to Intrusion Detection

We consider three existing approaches to intrusion detection in this work, which will be discussed in the remainder of this subsection:

Flow-based intrusion detection: Flow export technologies, such as NetFlow and IPFIX, are shipped with most high-end routers [7]. Traffic information is collected and stored in flow records that provide an overview of network usage at different levels of granularity. In [15], a flow is defined as *a set of IP packets passing an observation point in the network during a certain time interval; all packets belonging to a particular flow have a set of common properties.* Besides management purposes, flows can also be used to perform intrusion detection. With such an approach, the communication patterns within the network are analysed. Compared to traditional network-based IDSs, flow-based IDSs have to handle a considerable smaller amount of data, which is of advantage in terms of privacy and link speed (allowing to perform a detection in high-speed environments). This is mainly due to the aggregation of packets into flows, which comes at the expense of information granularity for the IDS. Figure 1 gives an overview of attacks that can be detected by flow-based IDSs. For the sake of clarity, it must be noted that in this classification, a virus is regarded as a worm that only replicates itself on the (infected) host computer and needs user interactions to propagate to other hosts [16-18].

FIGURE 1: CAPABILITIES OF FLOW-BASED IDS



As shown in Figure 1, on the one hand, flow-based IDSs are capable of detecting those attacks that are of special interest for a backbone network operator. On the other hand, quite a number of different approaches are available, each of them addressing specific aspects (see [7] for more details). However, the process of metering and exporting flows on a router, the collection of flows and the subsequent analysis consume a relatively large amount of time (up to several minutes [19]), introducing a certain delay within the intrusion detection process

Protocol-based/statistic-based intrusion detection: In contrast to flow-based IDSs, protocol-based/statistical IDSs are also performing decisions based on meta-data (i.e., packet header information), but here on *every* packet instead of an aggregated set of packets. One of the key advantages is that a decision is performed on a larger set of data. Furthermore, the process of generating the meta-data does not consist of multiple steps, but is performed by the IDS itself. Due to the fact that only packet headers are investigated, the approach is also capable of handling multiple Gbps *(medium link speed).*

Protocol-based IDSs monitor the dynamic behaviour and state of protocols. This method focuses on reviewing the strictly formatted data of network traffic (known as protocols) and searches for benign protocol activity for each protocol state to identify deviations. Unlike traditional behaviour-based intrusion detection, which uses host or network-specific profiles, protocol-based analysis relies on universal profiles that specify how particular protocols should and should not be used. *Stateful protocol analysis methods* (which is a synonym for protocol-based analysis) use protocol models, which are typically based on protocol standards from software vendors and standardization bodies (e.g., IETF) [6]. Each packet is wrapped in predefined layers of different protocols. A protocol-based IDS unwraps and inspects these layers, according to the protocol standards or RFCs. Anything that violates or is outside of these standards is likely malicious.

Statistical-based IDSs rely on statistical models such as the Bayes' Theorem, to identify anomalous packets. These statistics are based on actual usage patterns. As a consequence, statistical systems can adapt to behaviours and therefore create their own rule usage-patterns.

Anomalous activity is measured by a number of variables sampled over time and stored in a profile. In the course of this paper, the term statistical-based IDS is used to classify such behaviour-based approaches that only consider header information (or parts thereof) to generate their statistics (and to perform intrusion detection). Compared to flow-based IDSs, here, approximately the same time is needed for analysis. This is due to the fact that (i) in the case of stateful protocol analysis, the states of the protocol must be investigated for a certain time window, to have a clear indication, or (ii) in case of a statistical-based IDS, a *significant* deviation from the normal state is needed (large dataset). Thus, a near-real-time detection is considered as well.

Payload-based intrusion detection: Within this category, intrusion detection is usually preformed by checking a data stream (including the payload) for the presence of typical patterns, called signatures (knowledge-based approach). Typically, payload-based IDSs like Snort use rules for matching payload data. To this end, however, the entire package contents must be analysed, which slows down the process of intrusion detection, which in turn makes these systems less suitable for using them in high-speed environments. Typical representatives of open source IDSs are Snort, Suricata and Bro. In addition, several commercial products also perform intrusion detection with the use of knowledge-based DPI approaches.

C. Applicability of Existing Approaches for High-Speed Backbone Network Operators

Table 1 provides a brief overview of methods and approaches for intrusion detection. The first column lists previously discussed approaches. The second column lists the typical detection method of the respective approach. The third column provides information whether the approach relies on analysing the header/payload. The fourth column indicates whether the approach is feasible for a high-speed environment. Column five displays the time needed for detection. Finally, column six list the resource-intensiveness resp. the financial efforts for the corresponding approach.

TABLE 1: OVERVIEW OF IDS APPROACHES

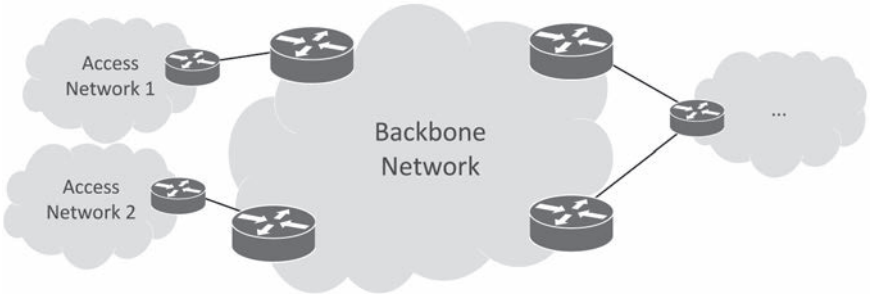
Approach	Typical Detection Method	Layer of Detection	Link Speed	Time of Detection	Financial Expenditure
Flow-Based	Behaviour	Header	High	Near Real-Time	Low
Protocol-Based	Knowledge	Header	Medium	Near Real-Time	Medium
Statistical-Based	Behaviour	Header	Medium	Near Real-Time	Medium
DPI-based	Knowledge	Payload	Low	Real-Time	High

Due to the large amounts of data in a backbone network, only flow-based IDSs can be used in practice. In addition, since customers do not explicitly pay network operators for security mechanisms, investments in IT security are very limited (low *Return on Security Investment (ROSI)*). Along with the ever-increasing data rates this in turn also leads to the fact that only flow-based IDS are used, since flow-based IDSs have by far the lowest financial expenditures [20].

3. SCENARIO

The primary focus of this work is on backbone networks, which we define as networks that do not provide network access to individual end hosts, and use links with speeds of 10 Gbps and higher. This is illustrated in Figure 2, where the backbone network has several edge routers that connect to other backbone networks and several access networks. These *access networks* can be residential Internet Service Providers (ISPs) or university campus networks, for example.

FIGURE 2: SIMPLIFIED BACKBONE NETWORK TOPOLOGY



Performing intrusion detection in backbone networks is subject to several challenges, both technical and legal. First, it is a resource-intensive process that requires expensive hardware to receive, pre-process, store and analyse the collected data. Second, backbone network operators often face legal constraints when performing DPI. DPI can be defined as scanning every byte of a packet payload and identifying a set of matching predefined patterns [21]. Although legislation in the area of packet inspection differs from country to country, the general tendency is that operators are not allowed to deal with data that can be traced back to individuals without permission. Exceptions are operational necessities, research, or court order. As a consequence, the backbone network operator in the context of this paper is generally not allowed to perform DPI, unless supported by a *clearly motivated occasion* or incident.

Many backbone network operators use flow export technologies for monitoring their networks. A recent survey among both commercial and research network operators has shown that 70% of the participants have devices that can export flows [8]. Flow export technologies, such as Cisco's NetFlow [22] or the recent standardization effort IPFIX [15], aggregate packets into flows. Deploying these technologies in backbone networks has several advantages. First, the aggregation of packets into flows significantly reduces the stringent requirements on data storage capacity and data analysis performance. Second, given that many high-end packet forwarding devices, such as routers and switches, already have flow export technologies embedded, deploying flow export comes at virtually no cost. And finally, backbone network operators have to save flow data anyway to comply with data retention laws. For example, network operators in Europe are forced to retain connection information for up to several years [23].

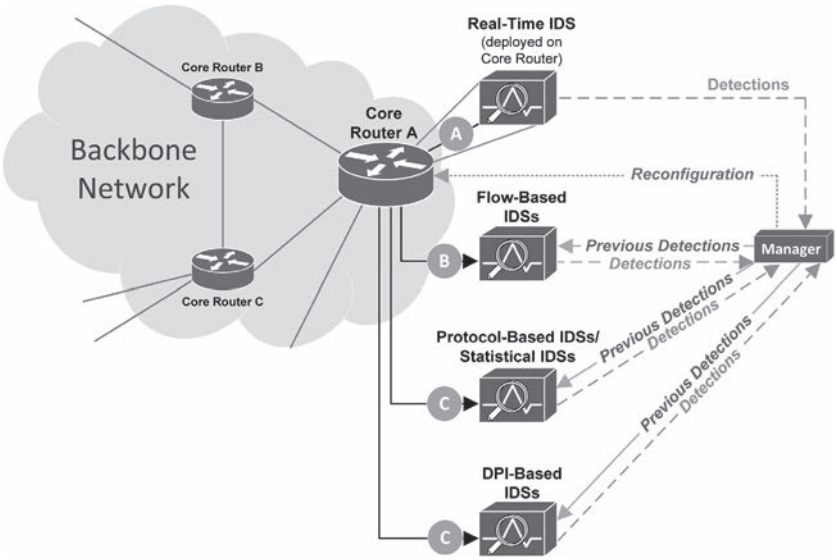
4. ARCHITECTURE

In this section we present our multi-layered architecture. We start by describing its main components and interactions in Section 4-A. In Section 4-B, we describe how existing systems can be integrated into our architecture.

A. Components and Interactions

The main components of our multi-layered architecture, together with their interactions, are shown in Figure 3. It has been designed with simplicity in mind and should be widely deployable.

FIGURE 3: COMPONENTS OF OUR ARCHITECTURE



The *Manager* controls all data-streams, and activates/configures the various IDSs. To make sure that every IDS receives the optimal data-stream, the *Manager* can reconfigure *Router A*. This router is equipped with a *Real-Time IDS* that performs the first layer of intrusion detection. Given that a router's main task is packet forwarding, this IDS is light-weight to not interfere with the router's critical operations.

Several data-streams can be identified in Figure 3:

- A Flow meta-data that can be retrieved directly from the router's Command-Line Interface (CLI);
- B Flow data, exported by means of Cisco's NetFlow [20] or the recent IETF standardization effort IPFIX [15];
- C Full packet streams, potentially pre-filtered by the router upon instruction by the *Manager*.

Key characteristic of the *Real-Time IDS* is that it constantly analyses the full traffic stream, without any form of sampling or filtering. In a previous work, we have shown that a similar approach is able to mitigate DDoS attacks in near real-time [23]. Upon detection of such an attack, the Real-Time IDS can reconfigure the router to drop the attack traffic, to make sure that neither the network itself, nor the monitoring infrastructure is overloaded. In addition, the *Manager* is informed about the attack by means of a standardized message exchange format, such as the Intrusion Detection Message Exchange Format (IDMEF); see [24] for an introduction and evaluation of IDS message exchange protocols.

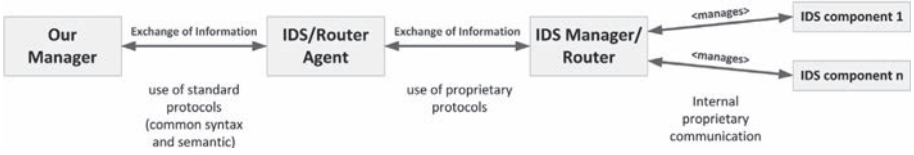
Besides the *Real-Time IDS*, the *Flow-Based IDS*s are also constantly monitoring the input data stream. Given that flow export technologies, such as NetFlow and IPFIX, aggregate packets into flows, such an IDS is usually capable of monitoring the aggregated traffic using commodity hardware. An example of a flow-based IDS is SSHCure, which detects SSH dictionary attacks and reports whether a host has been compromised [25]. The *Flow-Based IDS*s may be informed by the *Manager* about previous detections, and reports its own detections to the *Manager* again. Although not supported by current IDSs, the main idea of forwarding previous detection results to IDSs is to give as much information as possible and so to make the process of intrusion detection as reliable as possible.

In situations where the *Manager* decides to initiate a more extensive analysis of an attack, the *Protocol-Based IDS*s or *DPI-based IDS*s can be activated and instructed. The *Manager* decides which IDS/IDSs is/are most suitable for a particular attack. Before activating the other IDSs, the *Manager* has to reconfigure the router to pre-filter the traffic stream to only include the attack traffic. Analogously to the *Flow-Based IDS*s, these IDSs report their detections to the *Manager*. If an attack has been detected, the router is instructed to drop the attack traffic. If an attack could not be confirmed, the *Manager* will not dispatch any investigation about that particular traffic to the various IDSs anymore.

B. Use of Agents in Case of Proprietary Systems

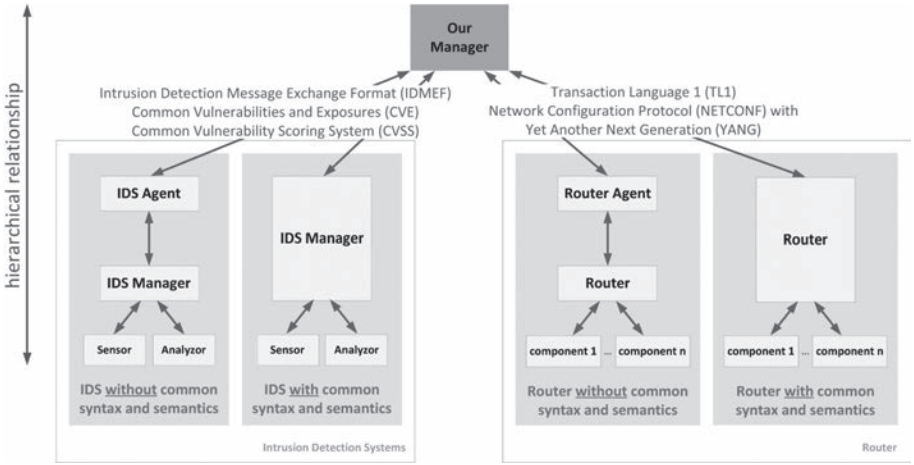
In this section, we discuss how existing systems that do not support standardized protocols for management (e.g., NETCONF) and information exchange (e.g., IDMEF) can be integrated into our architecture. The main idea, which is pursued in our approach, is to use specific agents (see Figure 4).

FIGURE 4: USING AGENTS IN CASE OF NON-STANDARDIZED PROTOCOLS



The agents are adapted for the individual system and thereby convert the standardized protocols used in our architecture into the proprietary counterpart used by the integrated system. This is done for the communication in both directions, i.e. from our *Manager* to the *IDS Manager* / Router, and for the reverse direction. The relationship between our *Manager*, the *IDS / Router Agent* and the *IDS Manager / router* is hierarchical. This means that our *Manager* uses the other Managers. Figure 5 visualizes this, as well as the standardized protocols and methods used.

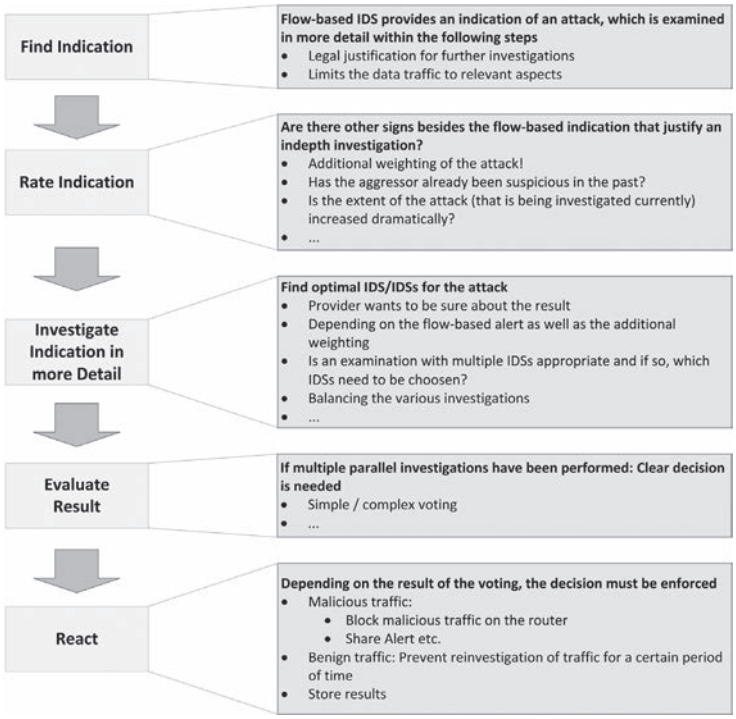
FIGURE 5: INTERACTIONS OF OUR MANAGER WITH EXISTING APPROACHES



5. MANAGER

The *Manager*, which is the architectural component that manages all other components, has a flow of operation as depicted in Figure 6. It consists of the following steps:

FIGURE 6: WORKFLOW OF THE MANAGER



Find Indication: The identification of the indication (*indication of an attack*) marks the beginning of a detailed investigation. For this, a flow-based IDS is used to look for signs of possible attacks. Since this investigation is not performed on packet payloads, both the individual privacy of the users is addressed in particular and the use of inexpensive IDS is made possible (especially since payload-based IDSs do have significant resource requirements). Hence, only few aspects of the data traffic are investigated.

Rate Indication: If an abnormality is detected, it is important to estimate the extent of the attack. Based on the alarm and the corresponding Common Vulnerabilities and Exposures (CVE)/Common Vulnerability Scoring System (CVSS), an assessment of the extent can be made.
While this gives a general assumption on the degree of damage an attack can cause, here, in addition to the scoring of the alert, supplemental criteria are used to estimate the specific severity. Such criteria are for example 1) whether an aggressor has already shown suspicious behaviour in the past, 2) whether the extent of the attack that is being investigated currently increases dramatically (e.g., the number of packets involved increases rapidly), or 3) whether a high number of similar attacks has been observed in the past. For this purpose, inter alia, a self-developed Geo-database is used in order to assist correlating attacks; see [26, 27].

Investigate Indication in more Detail: Depending on the overall scoring as well as individual aspects of the attack (type of attack), corresponding payload-based, protocol-based or statistical IDSs are to be identified. For example, if signs of an SSH-attack are observed by the flow-based IDS, the *Manager* may decide to investigate the relevant traffic by means of a statistical IDS (payload-based IDSs are not useful in this particular case, since SSH traffic is always encrypted). In contrast to this, when signs of a (non-encrypted) worm are detected, the *Manager* may directly involve a payload-based IDS.
As a backbone network operator wants to have a high-level of confidence before potentially mitigating an attack, involving multiple IDSs to investigate an incident may happen very often. The objective of the operator is to maximize the accuracy of the detection result and not to detect as many attacks as possible. However, the presence of several different types of IDSs does not necessarily imply that individual systems are very powerful. Since particularly transit customers don't spend a lot of money for security, an operator – as already mentioned – on the one hand wants to be sure that, if he blocks traffic that this decision in accordance to the law, but on the other hand, he will most likely not allocate powerful resources for that purpose. This leads to the situation that a relatively large number of systems may be available, but all of them with relatively little power. Therefore, it must be considered in advance, whether the specific request for an investigation can be carried out or not. This is mainly based on the scoring (see *Rate Indication*). The higher the score, the more important is a detailed investigation. If two investigations (with the same priority) are in conflict with each other, it is preferred to continue an on-going investigation, rather than to end and begin a new one.

Evaluate Result: Especially after several parallel investigations have taken place, the detection results need to be evaluated and compared. In case of contradictory results, an appropriate conflict resolution mechanism must be conducted. As a backbone network operator – as already stated – wants to be sure that the decisions made by him are solid, several models are conceivable.

On the one hand, this could mean that traffic is blocked only in the case of unanimity of all IDSs involved (which would subsequently lead to the fact that probably comparatively little traffic is blocked). On the other hand, a majority decision also seems to be conceivable. But also in this case, a clear vote seems to be essential, before a backbone network operator will make such a momentous decision like blocking traffic.

React: Once a decision is made, it must be enforced as well. In case of malicious traffic, corresponding packets must be blocked on the router. But even in the case of benign traffic, some actions need to be performed accordingly. E.g., it should be ensured that the traffic is not examined a second time (within a certain time period). In both cases, the result of the investigation is stored locally and also forwarded to other routers, which may include this result by means of the phase *Rate Indication*.

6. IMPLEMENTATION

For the realization of our architecture and implementation of a prototype, we use libraries and implement additional new modules and probes. As discussed before, the *Manager* is the central component of our architecture. It realizes the forwarding and selection of the network traffic, as well as the distribution based on the flow of operation presented in Section 5 as well as the configuration and assessment of alerts and their scores to the networks under consideration. The main routines of the controller are written in C programming language for the sake performance, combined with various open-source libraries.

The *Manager* contains a MySQL database, as well as different APIs to access and import data from various systems, such as CVSS and CVE details. With the help of the GUI of our Manager, the network security personnel can review and assess the relevance of the different threats. By this, the *Manager* is able to do additional weighting of possible attacks, including an estimation of the endangerment for the own network, and assigning examination orders to other IDSs. At the moment, the GUI is realized by a *ncurses* surface, but the upcoming prototype will be based on a Web interface. For further inspection of suspicious traffic, the *Manager* can forward the flows and network packets for a protocol analysis and further behaviour-based evaluations.

For our first prototype, we perform enhanced protocol analysis based on a special configured Snort IDS. Therefore, a standard Snort IDS is used with minimal functionality, disabling all signature-based detection schemes and only using the protocol analysis. In addition, we started to implement different modules for a behaviour-based protocol analysis. These modules are realized based on NFDUMP and the functionalities of the *nfreader* framework. Because of the comprehensive analysis of the protocols and the practical differences of their implementations in different operating systems, these modules will only be fully functional in a later release of our prototype.

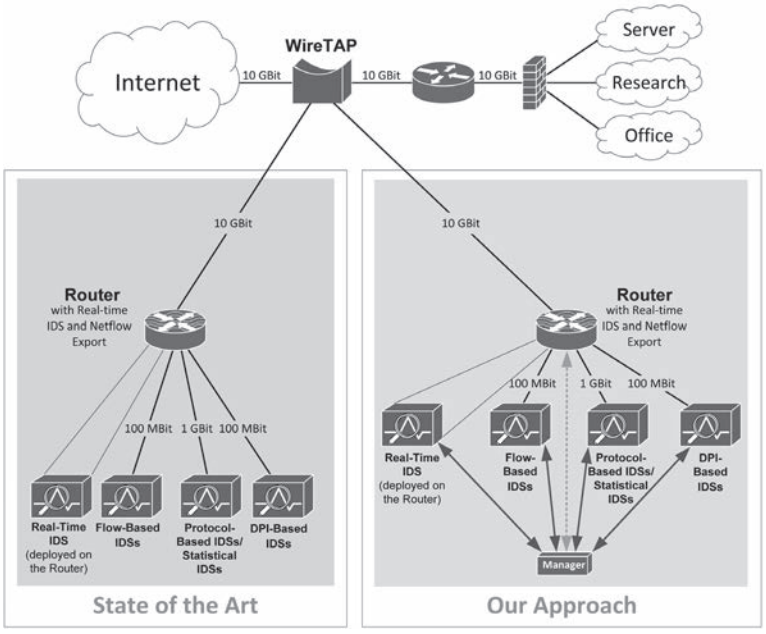
For the integration of knowledge-based and behaviour-based IDSs, a regular setup of Snort is used on the one hand, and a FlowMatrix system for the behaviour-based detection on the other hand.

The exchange of incident information between the different components and modules of our prototype is realized by IDMEF, for which the *LibIDMEF* is used [28].

7. EVALUATION

The first prototype is currently being tested extensively in our lab. For this purpose, a test set-up was built as described in more detail in Figure 7.

FIGURE 7: EVALUATION SETUP



With regard to our investigation, the department comprises three different networks: A server-network (with production systems), a research network (where various systems such as honeypots are tested, operated under specific conditions and evaluated) and a network for the office IT. By using a hardened system including a firewall and the application of additional protective measures, these three networks are separated intensively from each other (for more details see [29]).

With regard to state-of-the-art, the traffic is forwarded to the router. On the router itself, the *Real-Time IDS* is deployed and the Router also exports the data stream in the form of NetFlow V9 records, which are then analysed by the flow-based IDSs. In parallel, the protocol-based IDSs, statistical IDSs and DPI-based IDSs are supplied directly with data from the router. Here (as well as in our approach), the flow-based IDSs are connected with 100 Mbps (which is more than sufficient to handle the flow export records of the 10 Gbps link), while the protocol-based IDSs/statistical IDSs are connected with 1 Gbps and the DPI-based IDSs are connected at 100 Mbps. Of course, the flow export conditions are the same for state-of-the-art and our approach. In our approach the respective IDSs are connected using the same router model (Cisco 6513) and with the use of the *Manager* (as described in the previous sections).

Although it is too early to present results in detail, the first preliminary results are very promising. For the comparison, typical criteria such as 'probability of detection' (the ability of an IDS to identify positive results; proportion of malicious events that have been detected), false-alarm ratio (benign traffic that has been classified as malicious) and accuracy (proportion of true results, both true positives and true negatives) will be used. It may again be emphasized that the goal of our approach is not to identify as many positive results as possible (Probability of Detection), but to reduce the false alarm ratio.

8. CONCLUSIONS AND OUTLOOK

In this paper, we have presented a first step towards multi-layered intrusion detection, which aims both at reducing costs by being deployable on commodity hardware, and overcoming legal legislation with respect to traffic analysis (clearly motivated occasion in form of a flow-based alert is given before DPI is performed). Although a generic yet simple architecture has been defined and a first implementation realized, more steps have to be taken as future work before our IDS can be fully deployed in an operational environment. We shortly highlight these steps in the remainder of this section.

First, we plan to include more material on legislation in various countries with respect to network traffic analysis. As we want our multi-layer IDS to be as widely deployable as possible, this will be needed before finalizing the implementation.

The final design of our system will respect country-specific restrictions and possibilities. An auto-configuration based on the detected country will be provided, which can be tuned by the administrator. If modifications of the administrator violate the local restrictions, a warning will be given.

Second, after finishing the implementation, we plan to deploy it subsequently on campus-wide, region-wide and nation-wide scales. The goal of the various levels of deployment is twofold:

1. As operators of networks at different scales tend to use different devices and configurations, deploying our IDS in several networks allows us to validate its accuracy in multiple situations. For example, the flow data exported in campus networks is often exported with a sampling rate of 1:1 (i.e., everything is sampled), while nation-wide networks are often using sampling with a rate of 1:100, to reduce the data exported from the network. Our IDS should be able to cope with the difference in data granularity and should therefore be tested under these conditions, e.g., in terms of accuracy.
2. We have to get feedback from operators with respect to operational aspects. For example, we have to survey whether operators have technical facilities for deploying the various IDSs.

Third, we are trying to improve intrusion detection through inter-domain exchange of knowledge of attacks, both between "trusted partners" (in our case, within the so-called Joint Security Lab, consisting of various infrastructures operated by partners of Flamingo, a Network of Excellence project) and between partners with whom there is no special trust relationship. See [30] for an overview of our thoughts on this.

ACKNOWLEDGEMENTS

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:

[1] Ars Technica, “Can a DDoS break the Internet? Sure... just not all of it“ April 2013, accessed on 25 November 2013. [Online]. Available: http://arstechnica.com/security/2013/04/can-a-ddos-break-the-internet-sure-just-not-all-of-it/
[2] Arbor Networks, “Worldwide ISP Security Report“.
[3] Arbor Networks, “Worldwide Infrastructure Security Report“.
[4] Arbor Networks, “Worldwide Infrastructure Security Report - 2011 Volume VII“.
[5] Arbor Networks, ‘Worldwide Infrastructure Security Report - 2012 Volume VIII“.
[6] K. Scarfone and P. Mell, “Intrusion detection and prevention systems“ in Handbook of Information and Communication Security. Springer, 2010, pp. 177-192.
[7. 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, vol. 12, no. 3, pp. 343-356, 2010.
[8] J. Steinberger, L. Schehlmann, S. Abt, and H. Baier, “Anomaly Detection and mitigation at Internet scale: A survey,“ in Proceedings of the 7th International Conference on Autonomous Infrastructure, Management and Security, AIMS’13, Lecture Notes in Computer Science, vol. 7943. Springer Berlin Heidelberg, 2013, pp. 49-60.
[9] R. Koch, B. Stelte, and M. Golling, “Attack Trends in present Computer Networks,“ in Proceedings of the 4th International Conference on Cyber Conflict (CyCon). IEEE, june 2012, pp. 1-12.
[10] J. D. Howard and T. A. Longstaff, “A common language for computer security incidents“ Sandia Report: SAND98-8667, Sandia National Laboratories, http://www.cert.org/research/taxonomy 988667.pdf, 1998.
[11] S. Jin, Y. Wang, X. Cui, and X. Yun, “A review of classification methods for network vulnerability,” in Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on. IEEE, 2009, pp. 1171-1175.
[12] H. Debar, M. Dacier, and A. Wespi, “Towards a taxonomy of intrusion-detection systems,“ Computer Networks, vol. 31, no. 8, pp. 805-822, 1999.
[13] H. Debar, M. Dacier, and A. Wespi, “A revised taxonomy for intrusion-detection systems,“ in Annales des télécommunications, vol. 55, no. 7-8. Springer, 2000, pp. 361-378.
[14] S. Axelsson, “Intrusion detection systems: A survey and taxonomy“ Technical report, Tech. Rep., 2000.
[15] B. Claise, B. Trammell, and P. Aitken, “Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information,“ RFC 7011 (Internet Standard), 2013.
[16] V. Igure and R. Williams, “Taxonomies of attacks and vulnerabilities in computer systems“ Communications Surveys & Tutorials, IEEE, vol. 10, no. 1, pp. 6-19, 2008.
[17] N. Weaver, V. Paxson, S. Staniford, and R. Cunningham, “A taxonomy of computer worms,“ in Proceedings of the 2003 ACM workshop on Rapid malcode. ACM, 2003, pp. 11-18.
[18] S. Hansman and R. Hunt, “A taxonomy of network and computer attacks,“ Computers & Security, vol. 24, no. 1, pp. 31-43, 2005.
[19] R. Hofstede, V. Bartos, A. Sperotto, and A. Pras, “Towards Real-Time Intrusion Detection for NetFlow and IPFIX” in Proceedings of the 9th International Conference on Network and Service Management, CNSM’13, 2013, pp. 227-234.
[20] M. Golling and B. Stelte, “Requirements for a Future EWS - Cyber Defence in the Internet of the Future” in Proceedings of the 3rd International Conference on Cyber Conflict (ICCC). IEEE, June 2011.
[21] S. Kumar, J. Turner, and J. Williams, “Advanced Algorithms for Fast and Scalable Deep Packet Inspection,“ in Proceedings of the 2006 ACM/IEEE symposium on Architecture for networking and communications systems, 2016, pp. 81-92.
[22] B. Claise, “Cisco Systems NetFlow Services Export Version 9“ RFC 3954 (Informational), 2004.
[23] W. John, S. Tafvelin, and T. Olovsson, “Passive Internet Measurement: Overview and Guidelines based on Experiences,” Computer Communications, vol. 33, no. 5, pp. 533-550, 2010.
[24] R. Koch, M. Golling, and G. D. Rodosek, “Evaluation of State of the Art IDS Message Exchange Protocols” in International Conference on Communication and Network Security (ICCNS), 2013.

[25] L. Hellemons, L. Hendriks, R. Hofstede, A. Sperotto, R. Sadre, and A. Pras, “SSHCure: A Flow-Based SSH Intrusion Detection System,“ in Dependable Networks and Services. Proceedings of the 6th International Conference on Autonomous Infrastructure, Management and Security, AIMS’12, Lecture Notes in Computer Science, vol. 7279. Springer Berlin Heidelberg, 2012, pp. 86-97.
[26] R. Koch, M. Golling, and G. D. Rodosek, “Advanced Geolocation of IP Addresses” in International Conference on Communication and Network Security (ICCNS), 2013.
[27] R. Koch, M. Golling, and G. D. Rodosek, “Geolocation and Verification of IP Addresses with Specific Focus on IPv6” in 5th International Symposium on Cyberspace Safety and Security (CSS 2013). Springer, 2013.
[28] „LibIDMEF,“ accessed on 25 November 2013. [Online]. Available: http://sourceforge.net/projects/libidmef/
[29] R. Koch, and M. Golling, “Architecture for Evaluating and Correlating NIDS in Real-World Networks“ in Proceedings of the 5th International Conference on Cyber Conflict (CyCon), 2013.
[30] M. Golling, R. Koch, and G. D. R. Rodosek, “From Just-in-Time Intrusion Detection to Pro-Active Response by Means of Collaborated Cross-Domain Multilayered Intrusion Detection”, poster presented at the 9th International Conference on Cyber Warfare and Security ICCWS-2014.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Detecting and Defeating Advanced Man-In-The-Middle Attacks against TLS

Enrique de la Hoz
Computer Engineering Department
University of Alcala
Alcalá de Henares, Spain
enrique.delahoz@uah.es

Rafael Paez-Reyes
Spanish Navy
Cyberdefence Area
Madrid, Spain
rpaerey@fn.mde.es

Gary Cochrane
Cyberdefence Area
Indra
Torrejon de Ardoz, Spain
gicochrane@indra.es

Ivan Marsa-Maestre
Computer Engineering Department
University of Alcala
Alcalá de Henares, Spain
ivan marsa@uah.es

Jose Manuel Moreira-Lemus
Cyberdefence Area
Spanish Navy
Madrid, Spain
jmorlem@fn mde.es

Bernardo Alarcos
Computer Engineering Department
University of Alcala
Alcalá de Henares, Spain
bernardo.alarcos@uah.es

Abstract: TLS is an essential building block for virtual private networks. A critical aspect for the security of TLS dialogs is authentication and key exchange, usually performed by means of certificates. An insecure key exchange can lead to a man-in-the-middle attack (MITM). Trust in certificates is generally achieved using Public Key Infrastructures (PKIs), which employ trusted certificate authorities (CAs) to establish certificate validity chains.

In the last years, a number of security concerns regarding PKI usage have arisen: certificates can be issued for entities in the Internet, regardless of its position in the CA hierarchy tree. This means that successful attacks on CAs have the potential to generate valid certificates enabling man-in-the-middle attacks. The possibility of malicious use of intermediate CAs to perform targeted attacks through ad-hoc certificates cannot be neglected and are extremely difficult to detect.

Current PKI infrastructure for TLS is prone to MITM attacks, and new mechanisms for detection and avoidance of those attacks are needed. IETF and other standardization bodies have launched several initiatives to enable the detection of “forged” certificates. Most of these

initiatives attempt to solve the existing problems by maintaining the current PKI model and using *certificate pinning*, which associates certificates and servers on use. These techniques have significant limitations, such as the need of a secure bootstrap procedure, or pinning requiring some host-by-host basis.

This study proposes an evolution from pinning-in-the-host to pinning-in-the-net, by enabling mechanisms to validate certificates as they travel through a given network. Certificates would be classified as trusted or not trusted as a result of cross-information obtained from different sources. This would result in early detection of suspicious certificates and would trigger mechanisms to defeat the attack; minimize its impact; and gather information on the attackers. Additionally, a more detailed and thorough analysis could be performed.

Keywords: *certificate-pinning schemes, MITM attacks retaliation, SDN, OpenFlow*

1. INTRODUCTION

TLS [1] is an essential building block for securing virtually every application layer protocol and has also been successfully used to secure virtual private networks. A critical aspect for the security of any TLS dialog is authentication and key exchange, usually performed by means of X.509 certificates. An insecure key exchange can lead to an active third party (i.e. an attacker) being able not only to eavesdrop, but also to intercept and insert traffic in the communication in order to alter the setup process for the secure channel inserting himself effectively "in-the-middle" of the communication, thus hindering confidentiality and integrity.

Ideally, key exchange should only occur when there is certainty about the authenticity of the certificates involved. Trust in certificates is generally achieved using Public Key Infrastructures (PKIs), which rely on trusted third parties (called certificate authorities, CAs) to establish certificate validity chains [2], which are called certification paths. A communicating party assumes a certificate as authentic if the signature of the certificate can be traced back through a valid certification path up to a trusted CA. This method for validating certificates is the de facto standard in the Internet, and has been regarded as secure for decades.

Although the Public Key Infrastructure using X.509 Certificates (PKIX) [2] is meant to avoid the occurrence of man-in-the-middle attacks on TLS, recent incidents have clearly shown the weaknesses of the classical PKI model. The public CA model allows any trusted CA to issue a certificate for any domain name. A single trusted CA that betrays this trust, either voluntarily or by being compromised, can undermine the security provided by any certificates used in TLS just by issuing a replacement certificate that contains a rogue key, that is, a key not corresponding to the entity identified in the certificate.

A number of security concerns regarding PKIX usage have arisen in the last years, and it is foreseen that more incidents are likely to occur in the following years [3]. A certificate authority can issue certificates for any entity of the Internet, regardless of its position in the CA hierarchy

tree. A Spanish CA, for instance, can issue a certificate for a US government website, and vice versa. This was coherent with the decentralized nature of the Internet (avoiding single points of failure), but has turned instead into an "any-point-of-failure" problem. A successful attack on any CA in the hierarchy allows the attacker to generate valid certificates for any host in the Internet which will be blindly accepted by most users, browsers and Internet applications, thus enabling effective man-in-the-middle attacks. These attacks are not theoretical, but have been found in the real world. Comodo CA issued in 2011 certificates for major websites such as Google, Yahoo, Mozilla and Skype to an Iranian hacker [4]. The DigiNotar CA in the Netherlands was also removed as a trusted CA in most major browsers after issuing a Google certificate to a third party. Whether these incidents are the result of sophisticated attacks or poor security policies is irrelevant. The fact is that countries cannot just rely on the security of their own PKI infrastructures (or that of their allies). NATO can usually audit its own CA infrastructures and ensure their security. However, security breaches in an external CA can also jeopardize NATO own security. In addition, the possibility of malicious use of intermediate CAs to perform targeted attacks through ad-hoc certificates cannot be neglected [5], and these attacks are extremely difficult to detect. These rogue certificates can be used in man-in-the-middle attacks, which will not be detected by conventional mechanisms for PKIX certification path validation and revocation checks.

2. RELATED WORK

Current PKIX infrastructure for TLS is prone to MITM attacks, which are usually consummated by the use of forged certificates or by manipulating certificate path validation. IETF and other standardization bodies have launched several initiatives to enable the detection of "forged" certificates. Most of the proposals focus on minimizing the impact of certificate misissuance while maintaining the current PKI model almost unchanged in order to ensure compatibility, usability and low-cost deployment.

DNS-Based Authentication of Named Entities (DANE) [6] is a proposal to extend the secure DNS infrastructure DNSSEC [7] to store and sign keys and certificates which are used by TLS, so that clients can use this information to increase the level of assurance they receive from the TLS handshake process. Thanks to the use of DNSSEC, clients can verify that DNS information was provided by the domain operator and not tampered with while in transit.

The rationale behind DANE is that given that the DNS administrator for a domain name is authorized to provide identifying information about his jurisdiction zone, he should be allowed to make an authoritative binding between the domain name and a certificate that might be used by a host at that domain name. According to this line of thinking, the proper place to hold this information is the DNS database, securing the binding with DNSSEC.

This binding is done by means of a certificate association. A security association is composed by the domain name where the server application runs and some information from the certificate used to identify this application. A certificate association can also define the combination of a trust anchor and a domain name. This certificate association is represented by the TLSA

DNS resource record [6], which is used to associate a TLS server certificate or public key with a domain name. DANE defines several use cases, which allows to apply this binding information either to End Entities (EE) or to define new trust anchors that should be used to perform certificate path validation. A domain name administrator can even issue certificates for a domain without involving a third-party CA. A thorough description of DANE use cases can be found in [8].

Security associations are protected via DNSSEC. Taking into account that the deployment of DNSSEC infrastructure is still incomplete, any global proposal for certificate verification cannot rest solely on DANE. Moreover, certificate validation procedures will use only PKIX checks when no DANE information is available. An active attacker who is able to divert user traffic could block DANE traffic, so that he can bypassed these additional verifications. Moreover, there are situations where DANE information could fail to get to the End Entity due to server errors or broken intermediaries that filter DNSSEC errors. Under these circumstances, the End Entity performing the validation could assume an attack is undergoing and terminate the connection, or it could dismiss the error and proceed. The latter would mean that blocking DNSSEC traffic could help to bypass the DANE-defined procedures. Thus, in order for DANE to be effectively used to prevent MITIM attacks, a deployment of DNSSEC in clients, servers, DNS infrastructure and intermediaries (i.e., to avoid DNSSEC information filtering) is required. Taking into account the traditional resilience of network operators and manufacturers, we cannot rely solely on DANE to provide the kind of path validation we are looking for in this work. Finally, the verification of a key would require several DNSSEC queries that would introduce an undesired latency, unaffordable in some cases, e.g., SIP, XMPP.

In the short term, the basic technique that has been proposed to deal with this problem is known as *certificate pinning*, and relies on associating hosts with their *expected* X.509 certificates or public keys. *Pinning* is a way for clients to obtain a greater level of assurance in server public keys. By pinning a trusted known certificate (or public key), clients can detect any change either in the certificate or in the public key submitted by any server as part of any future TLS handshake.

There are two main problems related to pinning techniques. The first one is related to the process of bootstrapping the trust procedures, how we decide which associations are established. These associations can be set the first encounter with the host in a *Trust-On-First-Use* basis (TOFU), or can be defined by a list that is shipped with the application. The second one is the need for maintenance of the secure associations database, which is the secure creation of new associations and the revocation of existing ones if needed. Currently, there the two main proposals for certificate pinning are the Trust Assertion for Certificate Keys (TACK) Internet Draft [9] and the Public Key Pinning Extension for HTTP [10] promoted by Google.

In TACK, clients are allowed to pin to a server-chosen signing key (TACK signing key, TSK), which will be used to sign server's TLS keys. Given that the actual TLS keys are not pinned, the site is able to deploy different certificates and keys on different servers, without having the clients to renew its pins. Also since pins are not based on CA keys, there is no need to trust in CAs. TACK also defines a mechanism to activate pins. As part of the TLS handshake, a client

could request a compliant TACK server to send its TSK public key and signature. Once a client has seen the same hostname-TSK pair multiple times, it could decide to activate a time-limited pin for that pair. By time-limiting the pins, the potential impact of a bad pinning decision is bounded. The specification also mentions that pins could be aggregated and shared through a trusted third party but without defining either the infrastructure or the protocols required. This proposal, while promising, is still in a very early stage and accordingly not suitable for use in a production environment.

Public Key Pinning Extension (PKPE) for HTTP is conceptually quite similar to TACK but here the pins get delivered via a HTTP header and, accordingly, can only be applied to HTTP servers. This proposal defines a new HTTP header to enable a web host to tell browsers which public key should be present in the web host's certificate in future TLS connections. We can see this as a way to bootstrap public key pinnings. Once pinned, when connecting to a web server, the client can easily do PKIX checks and also can verify that one of the pinned keys for that server is present. The main drawback of this and other similar pinning techniques is that they do not protect the user against man in the middle attacks during the first connection attempt to the server. Also, such a MITM attack would not be detected until an update in the associations could be deployed to the hosts. This leaves an insecurity window that can be as long as one month in the PKPE case. To minimize this risk, a static list of pins is usually deployed with software packages. For instance, a total of 300 static pins are provided with the Google Chromium browser.

Another proposed solution is the 'sovereign keys' project by the Electronic Frontier Foundation [11] (EFF). This solution that uses a "semi-centralized, verifiably append-only data structure" containing the keys and revocations. These keys can only be added when it is strongly verified that the domain belongs to the requesting party. A browser would, when connecting to a TLS service, lookup the certificate from this key-store. Similarly to Certificate Transparency the existence of an append-only log with all CA-issued certificates is assumed.

Finally, pinning techniques require some configuration in a host-by-host basis and do not ship with a pre-established and well-defined mechanism for sharing pin information, even under the same domain. Currently very few sites publish pins, which limits the applicability of the proposal but it is expected that this situation will change in the near future, fuelled by the support by Google. Unfortunately, it is short-term solution and its scope is limited to HTTP so it is unable to help preventing MITM attacks against any other protocol secured by TLS.

The problem of verifying the authenticity of a given certificate can be affected by additional circumstances other than the presence of a rogue CA. For instance, a hostname can map to different servers, each with a different certificate and different CA chains, due to their dependence on different jurisdictions. Also, it is possible for a CA chain to change at any time, and this is out of the control of the administrators of the site. There is a proposal called *certificate transparency* [12], which tries to make the certificates that a certain CA has issued auditable and easy to track. This would make it easier for a site administrator to keep track of any new certificate issued for its site, usually a clear indication of a potential security breach. Participating entities should publish all certificates they issue so that clients could check

whether a certificate received by a server has a proof of publication. If the client is not able to obtain a cryptographic proof of publication, this could mean that the certificate has been forged. Note that this kind of verification can be provided by means of DANE.

Once again, the effectiveness of this technique is limited by the degree of deployment of the proposals. Certificate transparency can detect forged certificates issued by participating CAs but has none detection capabilities regarding non-participating CAs. This is an especially significant limitation, since usually a server is not concerned about misissuance by its own CA, but about the others (see, for instance, the TURKTRUST case [5],), these others CAs are out of its control. Finally, there is an inherent limitation derived from the very nature of the PKI model. Since the security of the whole PKI is the security of the weakest CA, and that these weakest CAs are not likely to be part of this initiative, the expected security improvement cannot be very significant.

There is a whole set of proposals that try to detect MITM attacks taking advantage from the fact that this kind of attacks are usually targeted attacks, rather than global scale attacks. This means that the attacker attempts to fool a specific target into believing the authenticity of the issued rogue certificate or key, while the rest of the Internet users are unaffected by the attack. Therefore, the victim will be receiving a certificate, which is different to the one seen by other Internet users. The *Perspectives* [13] and *Convergence* projects, in order to establish the validity of a received certificate, query designated nodes distributed over Internet, which act as *notaries*. A notary maintains a database of known server certificates. After the reception of a certificate by the client, she can check against the notary's version and flag mismatches as possible attacks. Notaries introduce a reputation scheme into the standard validation process.

Depending on the opinions received from these notaries, the certificate gets accepted or rejected. In practice, this voting scheme could be used to override the information used from the CA model. However, the client still has to trust these nodes (so they may become a point of failure if compromised), and there is a dependence on a pre-existing infrastructure.

There are some interesting initiatives in the Internet for sensing and mining information about the existing certificates, which could be used to produce a more valuable evaluation of the validity of a certificate. The most prominent examples are the ICSI Certificate Notaries Service [14] and the EFF SSL Observatory [15]. The ICSI Certificate Notaries Service passively collects certificates at multiple independent Internet sites, aggregating them into a central database almost in real-time. The ICSI Notary provides a public DNS interface allowing a client to query its database with the SHA1 digest of a certificate that it would like to check. The currently inactive project Certificate Catalogue by Google offered a service quite similar to this. The idea of deploying a set of sensors to passively detecting certificates in transit in order to identify common and uncommon patterns is also one of the key points of our proposal but, in our case, the set of provided parameters is richer in order to be able to perform a multigroup classification. Other initiatives like Crossbear or EFF SSL Observatory actively scan the Internet either by querying the TLS-enabled servers or asking the users to submit the certificates that they see.

The DetecTor Project [16] reuses the notary idea, but making every client to act as their own notary. To do this, the authors propose to use the Tor network to connect to the server under evaluation for the sole purpose of checking which certificate is seen when contacted from a different network location. This is essentially the same idea proposed in [17] but extended not only to HTTP but also to every protocol.

While the standardization work is progressing at a satisfactory speed, challenges remain. There is no common agreement of the design constraints and the types of threats that are supposed to be mitigated. The threat landscape is constantly evolving and an agreement about what threats need to be address does not exist.

In order to be actually effective, a widespread deployment is required by all this initiatives. This deployment can imply client, server and additional infrastructures (e.g., DNS infrastructure for DANE or CAs for Certificate Transparency).

3. MIDAS: A DISTRIBUTED PINNING IN THE NET APPROACH FOR AUTOMATED CERTIFICATE ASSESSMENT

The aforementioned proposals offer only partial solutions to the certificate assessment problem. They usually require full deployment of the initiative (given the 'weakest link in the chain' property of PKI). Finally, pinning needs to be performed in a host-by-host basis, which is hardly scalable.

This study proposes MIDAS (*Man-in-the-middle Distributed Assessment System*). MIDAS is an evolution from the pinning-in-the-host techniques to pinning-in-the-net techniques, by enabling mechanisms to validate certificates as they *travel through a given network*. Our idea is to classify certificates as trusted or not trusted as a result of cross-information obtained from different sources in an automated and distributed manner. While there have been some initiatives on using automated classification techniques for certificate assessment in the Internet [18, 19], they usually require centralized analysis of massive amounts of training data to become effective. While this large corpus of data including both legitimate and rogue certificates can be assumed as available for the Internet scenario, it is hardly achievable in internal NATO networks, which are not only more reduced in size and traffic, but also present a lower security incident rate than what we find in open networks. In the following, we propose an approach to pin certificates as they pass through the network, which takes advantage of collective intelligence techniques and does not require extensive training data.

A. Environmental Assumptions

Our approach assumes an environment of a secure internal NATO network. This implies a number of network segments compartmentalized by a set of (physical or virtual) switches and routers. It also implies the existence of several network management infrastructure elements.

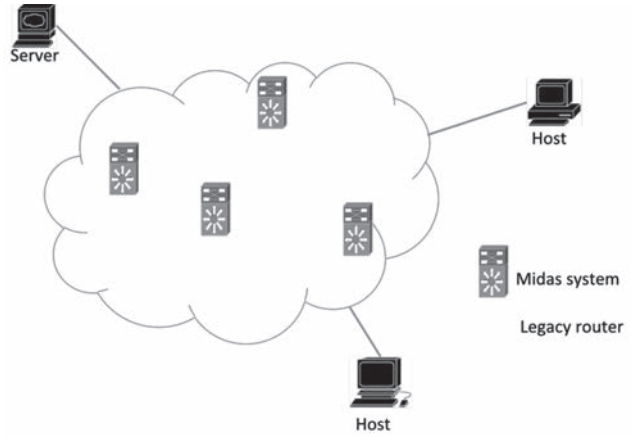
As for the threat model, we assume an insider attack scenario, since attacks from the outside will usually be handled using other techniques. We will assume the attacking entity to be an individual node or group of nodes, which are in minority with respect to the total nodes in the network. We will also assume that the targets of the MITM attacks (that is, the client and server between which the attackers intend to place themselves at) have not been completely isolated by the attackers (that is, both client and server are able to send data to other hosts in the network). We will later discuss techniques to ensure that these assumptions hold.

B. System architecture

Our system uses a distributed variation of the typical Intrusion Detection System Architecture [20], which encompasses the following elements:

- A distributed information source, consisting of a set of network probes. In our system, eventually any network element or host can act as a probe.
- A distributed analysis engine, which relies on Bayesian Networks to evaluate trust relationships according to information about the certificates involved and network history.
- A distributed reaction component, which allows to effective alter network topology in real time to transparently counter man in the middle attacks, thus ensuring network and service operation.

FIGURE 1: DISTRIBUTED MIDAS ARCHITECTURE



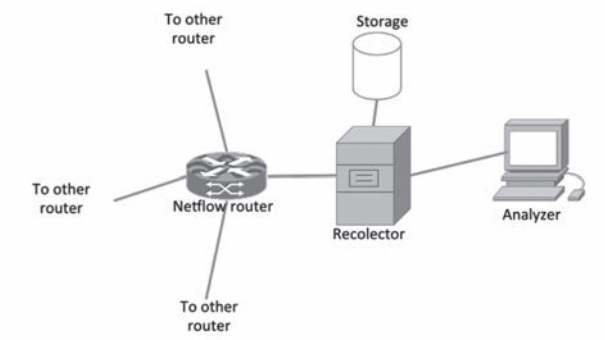
In this section we briefly outline each of these elements and their role in our proposal.

1) Network probes

As stated above, virtually any of the network elements within the communication infrastructure we want to secure may behave as a network probe. Basically, all that is needed is a network card that can act in promiscuous mode and capture packets from the network. Of course, such monitoring of traffic may have a serious impact in the performance of conventional hosts, so

we do not expect that all hosts in the network will act as probes. However, we assume that there are a sufficient number of probes distributed throughout the network. In particular, we rely on the existence of network management devices, which are specifically designed to be network probes. For instance, devices using NetFlow [21] or similar technologies for flow data analysis are especially suitable as probes in our system. These devices could, for instance, gather information about the TLS flows being established in the network, aggregating not only data about the certificates being used, but also the network path from source to destination or even the traffic patterns observed (e.g. an asymmetric flow with 80% of the traffic flowing from server to client).

FIGURE 2: ARCHITECTURE OF MIDAS SYSTEM



2) Distributed analysis engine based on bayesian networks

The MIDAS analysis engine, again, relies on distribution. Any node in the network can act as an analyzer, provided that it has information to analyze. Therefore, the most usual scenario is that network probes themselves act as analyzers in the case of probes residing in hosts, whereas Netflow collectors or analysis consoles act as analyzers in the case of probes residing in network management elements.

Analysis itself will be performed by using Bayesian Networks. A Bayesian network is a model that encodes probabilistic relationships among variables of interest. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data [22]. Each analyzer will have a built-in Bayesian network which is tailored to the specific scenario (given the usage model of the scenario), and which probability values are automatically adjusted during system life to adapt to the evolution of the network. The idea is that, for a given assessment query (e.g. “does this TLS handshake appear to be trustworthy?”), any analyzer can issue an assessment value which directly derives from the probabilities resulting from the network evaluation. Queries will typically occur when a given host needs to evaluate the trustworthiness of a given TLS exchange. The host will run an assessment using its own Bayesian network, and will also query a random set of nodes for their assessments on the validity of the same exchange. The host will then integrate all received values and its own to get

a final assessment, and will use this assessment to decide whether to accept the TLS exchange as valid or to flag it as an intrusion. The fact that the set of analyzers is chosen randomly by the evaluating host will make it harder to manipulate the receiving assessments, provided that there are different network paths used in the communications between host and analyzers and that a majority of analyzers have not been compromised.

Apart from assessments derived from queries, some of the analysis engines (typically, the ones in network management devices) will be entitled to provide automated detection. In this way, even if an assessment has not been requested on a given TLS exchange, a network element could flag it as anomalous (e.g. if a router captures a TLS flow with a certificate belonging to a server which is known to be in a different part of the network). This will allow also to react to events not directly related to certificate forging which could enable a MITM attack, such as the isolation of a given client or server.

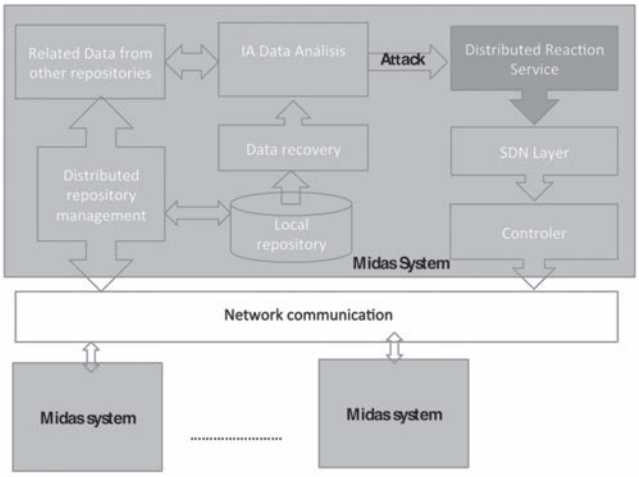
3) Reaction subsystem based on SDN

From the information obtained from the aforementioned analysis, MIDAS will be able to automatically define and put in place a restoring and reconfiguration plan of the network elements involved. This will allow, for instance, for traffic to be rerouted via an alternative path avoiding the attacking nodes, or to isolate compromised network segments. The reaction subsystem will be designed and implemented according to the novel, emerging architectural model called SDN (Software Defined Networking) that separates the control plane from the data plane in network switches and routers.

OpenFlow [23] is the first standard communications interface defined between the control and forwarding layers of an SDN architecture. It provides a singular point of control over the network flow routing decisions across the data planes of all OpenFlow-enabled network components. Taking advantage of this, security app can implement complex quarantine procedures, or malicious connection migration functions that can redirect malicious network flows in ways not easily perceived by the flow participants.

With SDN providing control over the forwarding, we can then isolate any malicious traffic to the quarantined network while all other traffic continues to operate as normal after being cleaned up. Upon detection of a potential attack, the traffic is placed in an isolated network segment that closely monitors the activity giving the attacker the perception that they’re interacting with a real system when, in reality, it’s a system that records their actions, decisions, and reactions, giving insight into their methodology.

FIGURE 3: LOGICAL ARCHITECTURE OF MIDAS SYSTEM



Although this component has not yet been implemented, similar approaches have shown its viability. In order to simplify the development and deployment process, we plan to use an approach similar to the FRESCO framework [24]. FRESCO is an OpenFlow security application development framework designed to facilitate the rapid design, and modular composition of OpenFlow-enabled detection and mitigation modules. Recently, the first security enforcement kernel for the OpenFlow controller Floodlight [25] has been released. The combination of FRESCO framework and SE-Flodlight provides a reference framework to rapidly prototype and field innovative security applications, which is makes it suitable for the kind of application that we want to develop.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced MIDAS, our proposal for a distributed certificate assessment system intended to thwart advanced Man-in-the-Middle attacks. This system builds on existing network monitoring and management technologies to provide a *pinning-in-the-net* approach enabling hosts to effectively assess the validity of the certificates they encounter during TLS interactions. The system relies on the existence of a set of network probes located in different elements of the network (either hosts or switches or routers), a distributed analysis engine based on bayesian networks and a reaction subsystem which makes use of SDN technologies.

Right now we have fully implemented the network probes and developed a proof-of-concept scenario of the complete architecture. Although the system looks promising, there is still considerable work to be done to build realistic Bayesian networks specifically tailored to realistic high-sensitive network scenarios. This would result in early detection of suspicious certificates and would trigger mechanisms to defeat the attack, minimize its impact, and gather information

on the attackers. Additionally, a more detailed and thorough analysis could be performed. This would be achieved through the use of Software Defined Network (SDN) techniques, allowing a much more accurate and efficient response to man-in-the-middle attacks, and mitigating damage in highly sensitive communication networks.

5. ACKNOWLEDGMENTS

The author would like to thank Álvaro Felipe Melchor for his work in the implementation of this proposal.

REFERENCES:

[1] T. Dierks and E. Rescorla. The transport layer security (TLS) protocol version 1.2. *(5246)*, 2008. Available: http://www ietf.org/rfc/rfc5246.txt,.

[2] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley and W. Polk. Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile. *(5280)*, 2008. Available: http://www.ietf.org/rfc/rfc5280.txt,.

[3] R. Oppliger. Certification authorities under attack: A plea for certificate legitimation. *Internet Computing, IEEE PP(99)*, pp. 1-1. 2013. . DOI: 10.1109/MIC 2013.5.

[4] *Comodo Report of Incident*. Available: http://www.comodo.com/Comodo-Fraud-Incident-2011-03-23.html.

[5] S. B. Roosa and S. Schultze. Trust darknet: Control and compromise in the internet's certificate authority model. *IEEE Internet Comput. 17(3)*, pp. 18-25. 2013. . DOI: http://doi.ieeecomputersociety.org/10.1109/MIC.2013.27.

[6] P. Hoffman and J. Schlyter. The DNS-based authentication of named entities (DANE) transport layer security (TLS) protocol: TLSA. *(6698)*, 2012. Available: http://www.ietf.org/rfc/rfc6698.txt,.

[7] R. Arends, R. Austein, M. Larson, D. Massey and S. Rose. DNS security introduction and requirements. *(4033)*, 2005. Available: http://www.ietf.org/rfc/rfc4033.txt,.

[8] R. Barnes. Use cases and requirements for DNS-based authentication of named entities (DANE). *(6394)*, 2011. Available: http://www.ietf.org/rfc/rfc6394.txt,.

[9] M. Marlinspike and T. Perrin. Trust assertions for certificate keys. Internet Engineering Task Force. 2013Available: http://tools.ietf.org/id/draft-perrin-tls-tack-02.txt,.

[10] C. Evans, C. Palmer and R. Sleevi. Public key pinning extension for HTTP. Internet Engineering Task Force. 27Available: http://www.ietf.org/internet-drafts/draft-ietf-websec-key-pinning-09.txt,.

[11] EFF. (feb). *The Sovereign Keys Project*. Available: https://www.eff.org/sovereign-keys/.

[12] B. Laurie, A. Langley and E. Kasper. Certificate transparency. Internet Engineering Task Force. 2013Available: http://www.ietf.org/internet-drafts/draft-laurie-rfc6962-bis-00.txt.

[13] D. Wendlandt, D. G. Andersen and A. Perrig. Perspectives: Improving SSH-style host authentication with multi-path probing. Presented at USENIX 2008 Annual Technical Conference on Annual Technical Conference. 2008, Available: http://dl.acm.org/citation cfm?id=1404014.1404041.

[14] *The ICSI Certificate Notary*. Available: http://notary.icsi.berkeley.edu.

[15] EFF. The EFF SSL observatory. Available: https://www.eff.org/observatory 2013.

[16] *DetecTor.io*. Available: http://detector.io.

[17] M. Alicherry and A. D. Keromytis. DoubleCheck: Multi-path verification against man-in-the-middle attacks. Presented at Iscc. 2009, Available: http://dblp.uni-trier.de/db/conf/iscc/iscc2009.html#AlicherryK09.

[18] J. Braun, F. Volk, J. Buchmann and M. M\uhlh\auser. Trust views for the web PKI. *Proceedings of the 10th European Workshop on Public Key Infrastructures, Services and Application (EuroPKI 2013)* 2013.

[19] M. Abadi, A. Birrell, I. Mironov, T. Wobber and Y. Xie. Global authentication in an untrustworthy world. Presented at Proceedings of the 14th USENIX Conference on Hot Topics in Operating Systems. 2013, Available: http://dl.acm.org/citation.cfm?id=2490483.2490502.

[20] R. G. Bace. *Intrusion Detection. Ed Mc.Millan* 2000.

[21] B. Claise. Cisco systems NetFlow services export version 9. *(3954)*, 2004. Available: http://www.ietf.org/rfc/rfc3954.txt,.

[22] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández and E. Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Comput. Secur. 28(1–2)*, pp. 18-28. 2009. . DOI: http://dx.doi.org/10.1016/j.cose.2008.08.003.

[23] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker and J. Turner. OpenFlow: Enabling innovation in campus networks. *SIGCOMM Comput.Commun.Rev. 38(2)*, pp. 69-74. 2008. Available: http://doi.acm.org/10.1145/1355734.1355746. DOI: 10.1145/1355734.1355746.

[24] S. Shin, P. A. Porras, V. Yegneswaran, M. W. Fong, G. Gu and M. Tyson. FRESCO: Modular composable security services for software-defined networks. Presented at Proceedings of the 20th Annual Network & Distributed System Security Symposium. 2013, Available: http://dblp.uni-trier.de/db/conf/ndss/ndss2013.html#ShinPYFGT13.

[25] *The FloodlightProject. Floodlight*. Available: http://www.projectfloodlight.org.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Inter-AS Routing Anomalies: Improved Detection and Classi cation*

Matthias Wübbeling
Fraunhofer FKIE
& University of Bonn
Bonn, Germany
wueb@cs.uni-bonn.de

Michael Meier
Fraunhofer FKIE
& University of Bonn
Bonn, Germany
mm@cs.uni-bonn.de

Till Elsner
Fraunhofer FKIE
& University of Bonn
Bonn, Germany
elsner@cs.uni-bonn.de

Abstract: Based on the interconnection of currently about 45.000 Autonomous Systems (ASs) the Internet and its routing system in particular is highly fragile. To exchange inter-AS routing information, the Border Gateway Protocol (BGP) is used since the very beginning, and will be used for the next years, even with IPv6. BGP has many weaknesses by design, of which the implicit trust of ASs to each other AS is the most threatening one. Although this has been topic on network security research for more than a decade, the problem still persists with no solution in sight. This paper contributes a solution to stay up to date concerning inter-AS routing anomalies based on a broad evidence collected from different publicly available sources. Such an overview is necessary to question and to rely on the Internet as a basis in general and must be a part of every cyber defense strategy. Existing methods of detecting inter-AS routing anomalies result in large sets of real time routing anomalies, based on the evaluation of routing announcements collected from different viewpoints. To decide, whether a detected anomaly is harmful or not, each of them has to be classified and correlated to others. We combine various detection methods and improve them with additional publicly available information. The improved outcome of the implemented routing anomaly detection system is used as input for our classification algorithms.

Keywords: *Internet, Routing, Anomaly Detection, BGP, Autonomous Systems*

* The work presented in this paper was partly funded by the German Federal Ministry of Education and Research under the projects MonIKA (BMBF 16BY1208A)

1. INTRODUCTION

The *Border Gateway Protocol* (BGP) [22] defines the exchange of IP routing information between interconnected Autonomous Systems (ASs) in computer networks. It is the only used routing protocol in the Internet and it is topic of security research since the late 90's. Therefore, itself and its inherent weaknesses are well known. Implicit trust between connected ASs results in the possibility for any AS to inject invalid and malicious routing information with very little effort. Wrong routing information, distributed from one or more ASs over the whole Internet could lead to large scale connectivity problems. The existence of contrary routing information at different locations is called a routing anomaly. Routing anomalies like Multiple Origin AS (MOAS) [25, 8] conflicts, where two or more ASs claim to own the same range of IP addresses, occur regularly. This situation is not only intended to cause harm, based on malicious intention. It can also happen as a result of misconfiguration inside an AS. Countermeasures against IP prefix hijacking, the advertisement of the same IP address space from a foreign AS, still do not exist. Legitimate owners of IP addresses are able to announce longer, more specific IP subnets than the causing/attacking AS, because they are preferred, when the route for a packet is chosen. Only few of these events are publicly known, usually those involving large internet companies such as YouTube or Google [23].

Although MOAS conflicts are easy to detect, they could be used intentionally by prefix owners to implement load balancing or to minimize the routing distance for connections to/from different locations. Thus, the distinction between legitimate and illegitimate conflicts is hard to make. Due to several reasons, e.g. performance issues on large routing systems or impracticability of approaches like S-BGP [14, 13], the threats still exist nowadays. The improvement of routing security brought by origin authentication [6] and asymmetric cryptography, e.g. RPKI [17] is currently small, because it is not yet implemented in broadly used hardware and business processes of ASs. Unless most parts of the Internet support origin authentication or RPKI, the routing system in general is as vulnerable as before. In contrast to prefix hijacking, routing anomalies, that are based on invalid topological information propagated in routing announcements, are significantly harder to detect and to classify.

Several approaches were made to detect and classify routing anomalies based on information gathered from inside the routing plane. They provide systems to identify prefix hijacking events [16, 7, 21]. None of those solutions really classify all found conflicts properly. Classification is necessary to determine whether an occurring conflict is legitimate or illegitimate to derive a level of criticality for a conflict. One common shortcoming of all these solutions is that the assumed ground truth, the data used to train and measure the detection and classification systems, is just based on inherent information exchanged via BGP itself [4, 18].

This is questionable because the exchanged routing information is not reliable, as discussed above. To determine facts of actually existing peering relations and legitimate IP address owners, it is necessary, to query other sources to increase the data used as ground truth evidence. Ground truth evidence in this context is the amount of reliable data to be used to find and to classify occurring routing anomalies.

Our contribution is (1) the collection of a broader data base of reliable information on peering relations between autonomous systems and therefore higher accuracy at finding and classifying routing anomalies, (2) an approach, based on existing systems named above, providing evidence to the ground truth used to find and classify routing anomalies, (3) a selection of reliable sources for this enrichment and (4) a crawling system, gathering information from different viewpoints inside the Internet routing layer, internet exchange points (where ASs can, an mostly do, peer with each other), and AS specific web services such as looking glass, a service to query information from running routers inside an AS.

This paper is structured as follows: first we describe the background and challenge of our research in section 2, then we present used detection and classification methods in section 3 and move on with the presentation of our approach to extend the assumed ground truth as argued above to improve the handling of routing anomalies in section 4. Our applied approach to classify routing anomalies is discussed in section 5 followed by the evaluation in section 6 and the last section 7 includes discussion and future work.

2. BACKGROUND

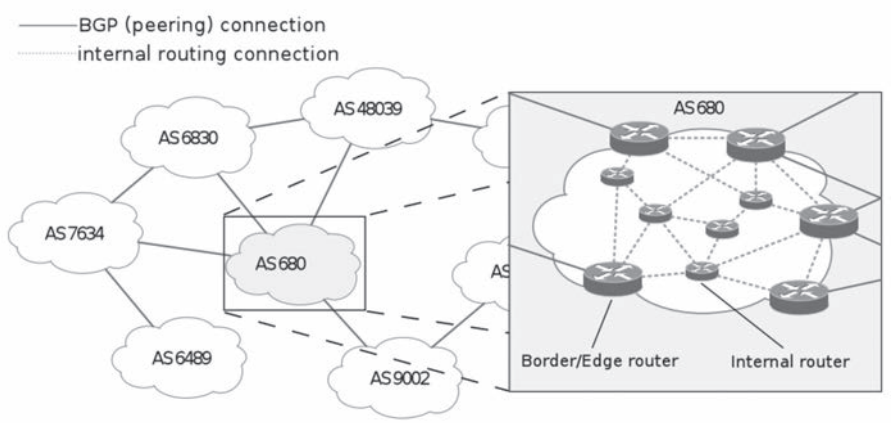
This section describes backgrounds of the Internet and its routing plane followed by an introduction and an explanation of Internet routing anomalies.

Internet routing

The current structure of the Internet is the result of massive growth in the last two decades, mostly driven by civil usage of the World Wide Web and other services such as IP-Telephony and IP-Television. The majority of Internet participants, either they use it for private or business purposes, has a limited view of the techniques behind the Internet. Although the Internet is seen as an abstract item, it is in fact just the interconnection of different independent networks, called Autonomous System (AS), as illustrated in Figure 1.

Each AS belongs to one administrative domain, most of them are large enterprises (e.g. ISPs, IT-Services), governments, organizations or universities. The interconnection of these networks is possible because of physical links between them. Routers connected to other ASs' routers are called border router or gateway. Thus, the Internet is not more than a network of networks. The connection between Autonomous Systems is called *neighborship* or *peering*. Each neighborship is related to at least one (commercial) agreement between the two parties. There are provider-customer, peering and transit relationships between ASs.

FIGURE 1: INTERNET ROUTING



To operate an AS as part of the Internet it is necessary to register a unique AS number. AS numbers are assigned by regional internet registries (RIRs) on behalf of the Internet Corporation for Assigned Names and Numbers (ICANN). Additionally, each AS needs at least one subnet of the global IP address space so it can be addressed by other ASs. These subnets are also regulated by ICANN and distributed by RIRs, e.g. RIPE NCC [1] for European customers. An AS announces owned IP addresses as *subnets* (also named *prefixes*; the prefix of an IP address defines the subnet to which an address belongs to) to each neighbor. The *announcement* (or *advertisement*) contains the served prefix in classless interdomain routing (CIDR) notation [11] together with the owners AS number as *origin* and additional information as described in the BGP [22]. Prefixes, reachable by a neighbor of a receiving AS, are then re-announced by that AS to all other neighbors, with the own AS number prepended to the origin AS. The concatenation of all AS numbers between an AS and the owner of a prefix builds the AS *path* for the prefix. When the receiving AS already knows another path to a prefix, only the best path will be chosen and sent to the neighbors.

Announcements of prefixes and AS paths are routing information used to deliver IP packets to their destination. To exchange routing information between AS border routers, BGP is used. BGP is the first and only routing protocol used in the Internet, so it is the de-facto standard, implemented in all participating border routers. Besides static routing protocols used inside ASs, BGP encounters the dynamics of a global and rapidly changing Internet. Border routers establish BGP connections to border routers of other ASs and exchange routing information with them. Since the Internet originally was an interconnection of trusted universities and research facilities, BGP assumes unlimited trust between neighbors regarding
routing information provided by them. Hence, BGP has no built-in verification mechanisms to check for the validity of routing announcements.

To provide access to the Internet, an AS needs routing information for every addressable Prefix. Because CIDR allows different prefix lengths, it is possible to aggregate them into shorter prefixes containing all subnets to decrease the number of necessary routing entries. To prevent routing failures, the most specific (longest) prefix determines the route to a destination address, if two or more prefixes overlap. Due to the implicit trust and a missing global authority, it is possible for ASs to provide invalid routing information. Thus, an AS can announce the reachability of an IP prefix, although it is not the legitimate owner nor does it have the advertised routing abilities.

BGP alongside Internet routing in general is subject of research activities for more than a decade [20, 15, 10]. Problems resulting from the weakness of implicit trust between neighbors, no matter whether they are in a provider-to-customer, a peering or transit relationship, cannot finally be solved. It is possible to filter announced routes from customers or peers but that is not sufficient to secure BGP routing as only few of the prefixes are originated by an AS's peers. Research projects and routing hardware vendors [14, 19] from time to time propose BGP optimizations or BGP successors to secure Internet routing [13] but none of them has been emerged to secure every day routing. Besides the goal to solve this issue, the research community accepts it as a fact and tries to find other ways to allow trustworthy inter AS routing. One of the main goals of network and internet security research is to provide reliable internet connectivity to end users, organizations and enterprises. To achieve this, the BGP-state of the Internet is continually monitored by different institutions and companies. BGPMon net [2], as an example, offers services to inform victims of IP hijacking, in case of another AS illegitimately announcing any of their prefixes.

Most of the named research projects are built upon information collected by routing archives such as RIPE RIS [4] or routeviews [18]. Those archives peer with volunteer ASs and collect announced routes or received routing announcements from a route reflector, a border router that just reflects all received announcements to designated clients, inside different ASs around the globe. Relying on information derived from the routing layer itself is one of the handicaps all these projects have in common.

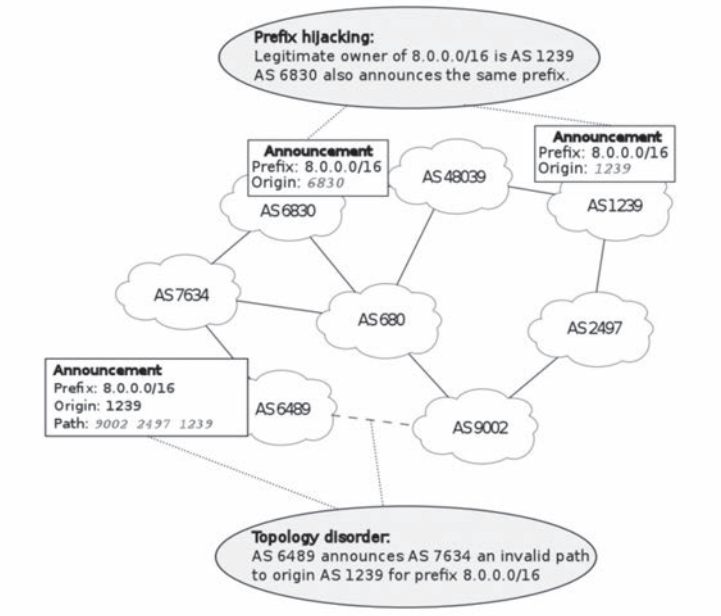
Routing anomalies

Anomalies within the routing plane of the Internet occur regularly and they last from only a few seconds to several months [7, 8, 16, 23]. This section will give a short summary of how anomalies can happen and how to react on them.
BGP is a message based protocol. Border, or *edge routers* of ASs send messages to their physically connected neighbors in other ASs to inform them about a) their own IP prefix and b) transitively reachable IP prefixes of other ASs. Beside other information, those messages at least contain the registered AS number of the peering AS, the AS number of the originating AS, meaning the AS that owns the announced prefix, and the AS numbers of all the ASs between the receiving and the originating AS, namely the AS path.

According to Lad et al. [16] and Qiu et al. [21], we consider a routing update an anomaly, when at least one of the following conditions is met:

- An invalid AS number is used.
- One or more invalid or reserved IP prefixes are used.
- The IP prefix is not owned by the originating AS.
- The same IP prefix is originated (announced) by two or more ASs.
- The given AS path has no physical equivalent.
- The provided AS path does not match common routing decisions.

FIGURE 2: ROUTING ANOMALIES: PREFIX HIJACKING AND TOPOLOGY DISORDER



The consequences (or incidents) of routing anomalies are commonly categorized into *blackholing*, *rerouting/path spoofing* and *hijacking* [21, 24]. For our research we only need to distinguish routing anomalies into two different types: prefix hijacking and *topology disorder*, as shown in Figure 2 and described below.

Prefix hijacking

Prefix hijacking occurs, when the given origin inside a BGP announcement, i.e. the owner of an IP prefix, is not the legitimated and registered AS itself. Prefix hijacking can affect a whole subnet or only parts of it with a larger prefix, which we then call *subprefix hijacking*. Subprefix hijacking differs from the common understanding of sub MOAS conflicts, as long as the subnet is legitimately assigned to another AS. The route selection process prefers paths with the longest prefix to determine the route to a specific IP address. In case of equal length, a MOAS conflict would match our understanding of prefix hijacking. Prefix hijacking could cause blackholing, when the wrongly announced prefix is not routed (or served) within the causing AS. It does not affect the whole Internet, it rather divides it concerning the announced prefix, one part of the Internet uses the benign and the other the bogus route, depending on AS specific routing decisions.

Topology disorder
Topology disorder happens, when an announced path is invalid, i.e. has no corresponding physical equivalent that could be traversed or violates reasonable routing decisions. Such a disorder could lead to longer and also shorter paths, hence influences the route selection process of other ASs. While prefix hijacking caused by accidental misconfiguration, a manipulated AS path can only happen intentionally save those caused by bugs in router firmware, but latter are rather unlikely.

Routing anomalies are not necessary harmful. Large service providers might enforce anomalies to realize geographical load balancing or multi homed ASs. An AS is multi homed, when it has two or more relations to ASs where it is customer in a provider-to-customer relationship, i.e. needs other ASs to address the rest of the Internet, to increase its own routing abilities and to have a backup path if one provider fails. That means, occurring anomalies caused by topology disorder have to be examined in a special way to classify them as legitimate or illegitimate ones. The results of this classification should be reliable enough to send proper alerts to legitimate owners of prefixes or administrators of causing and affected ASs to be informed about the anomaly and to solve it.

Conclusion
As a matter of fact, no real countermeasures to routing anomalies exist. It is thinkable, e.g. in case of a race for a specific IP prefix, to announce longer prefixes than the causing AS. This game stops at least at 24 Bits length because longer prefixes are not valid in the Internet routing. Thus, the conflict remains. Unless BGP could be totally replaced, AS operators and researchers have to deal with its weaknesses.

3. ANOMALY DETECTION

Since BGP routing weaknesses and anomalies are still topic of active research, various mechanisms and algorithms for anomaly detection have been proposed and developed by the research community. This section describes our applied approach to detect routing anomalies based on already existing solutions.

Our anomaly detection incorporates already existing approaches, which we combine to gain benefits from all solutions [26, 16, 21]. Based on these, we examine current routing announcements from the beginning of 2013 until the end of October 2013. We evaluate our results against a list of known anomaly routing events from Team Cymru and BGPMon net [5, 2].

The named systems are mainly based on historical routing information derived from routing archives [4, 18]. To improve detection rates, detection runtime and in order to detect anomalies not yet in these lists, we filter the routing announcements prior giving them to the anomaly detection. Based on our broader knowledge we filter announcements that are reliably proper so that there is no need to run each classifier on it. Our contribution is that not only anomalies are classified on broader ground truth evidence, but additionally to confirm information found in the announcements prior the detection.

To improve the reliable data our solution is based on, we gather additional reliable routing (and especially peering) information from different (primary) sources of the Internet. How we achieved this is shown in the next section 4 of this paper.

While parsing retrieved BGP archives each contained announcement is evaluated before being inserted into the analysis database. As mentioned earlier, the database contains all announcements from the beginning of the regarded interval, i.e. January 2013 in this case, until the receive date of the examined announcement. If a database entry holds an announcement that is still vital and provides the same prefix but is originated by another AS, a MOAS conflict is detected. Such conflicts are calculated per prefix and reported with the affected prefix and all participating Autonomous Systems.

Afterwards, the AS path of each announcement is examined and checked for known and confirmed AS peering relations. Those peering relationships are derived from the database containing historical announcements. As this information is not sufficient, paths shall be examined based on the database created as a result of this papers research, see the following sections for further details on how the data is collected and evaluated. When no such peering relation can be confirmed for each contained AS link, an anomaly is raised with the affected announcement and the corresponding ASs. When an anomaly is detected, additional actions are triggered, such as querying the corresponding ASs or an internet exchange point both ASs are connected to.

Conclusion
Anomaly detection is primarily based on publicly available data and has to be improved by additional collected data as evidence of ground truth. Detected anomalies are stored inside a separate database for further use such like BGPMon net or end user warning systems [27].

4. IMPROVING THE GROUND TRUTH EVIDENCE

This section describes our contribution and the steps we make to collect further information on routing relationships from other primary and reliable sources, in order to enlarge the assumed ground truth of our detection system.

To improve the basis of the classification of routing announcements, we need to obtain reliable information about peering and other business relationships of ASs, but unless such information is publicly available and it is known how to retrieve it, there is currently no way to take them into account. Confidential information aside, there is a lot of publicly available and usable information about AS relationships.
Existing approaches obtaining AS relationships [12, 9] use information gathered from within the routing system itself, based on collected BGP announcements and derived node degrees.

In the context of our project, the examined autonomous systems are restricted to those, located in countries of the European Union (EU). Having a number of 28 countries, we retrieved a number of about 11.500 ASs from the RIPE whois database located in respective countries.

This represents about 25% of the registered ASs worldwide. The number of registered and announced IPv4 prefixes is about 70.000 at the time of writing, what is around 14% of the globally assigned 510.000 prefixes we found in a recent table dump in October 2013 [4].

Our goal is to collect additional information on those EU-located ASs in order to improve the routing anomaly classification. Several sources exist, where such information could be found. We start with a naive approach and collect whois data from RIPE [1] first. A RIPE whois database entry contains information of a registered ASs, its AS number, the name, description, contacts and various other. The number of queries at RIPE is generally limited to 1.000 queries per 24 hours and IP address, when contact information is contained. Additional information on RIPE's whois database usage is given below.

Reaching this limit quickly leads us to look for other sources containing similar information. The website peeringdb.com [3] contains specific information about inter-AS peering and holds a list of known Internet Exchange Points (IXPs) in Europe. An Internet Exchange Point is a datacenter with special focus on network peering. To get an AS connected to many other ASs with little effort, AS operators rent special network ports in that datacenter. Depending on an ASs peering policy connections between different ASs can be established and used for BGP peering. Hence, it is feasible to establish peering connections to many other ASs located at the same datacenter with just one physically network connection. Due to the ease of establishing peering, ASs located in large IXPs commonly have many peers.

The peeringdb.com database does not claim to be complete but it gives a good starting point for further research. We extracted EU-located IXP datasets, 118 in number, including their website addresses and a list of ASs peering at them. An AS peering at a specific IXP is referred to as *member* of this IXP.

An AS entry from peeringdb also contains the address of looking glass servers provided by the AS, when it is publicly accessible. The utilization of information obtained from looking glasses is described in this section below. To get listed inside the peeringdb database it is necessary for an AS to register and provide sufficient information for the entry. Consequently, not all IXPs, ASs and peering relationships are listed there. As part of their peering policy, some ASs require the existence of a database entry of the peering partner at peeringdb.com to peer with them. Due to the fact that database entries are manually maintained by each AS itself, the database can't be regarded up to date.

Starting with the list of EU-located IXPs [3] we collect information from the IXPs' websites directly. Most of them provide a list of members and some additionally a detailed peering matrix. This is valuable and reliable information on actual peering relationships between listed ASs since IXPs get paid for peering services and therefore update information of their members regularly based on their business processes. If there is no peering matrix provided, the majority of IXPs at least list peering policies of their members showing whether it is open, selective or closed. The usage of peering policy information is also described below.

Looking glass

In order to get reliable information about peering relations, BGP specific information such as full routing tables and next hops for various routes are of interest. To ensure, that derived information is reliably and correct, we collect information from AS border routers directly by accessing them through their looking glass service. Looking glass servers provide access to live routing information of an AS itself. A looking glass service directly queries the routers involved in BGP operation to provide up to date information about actual relationships between BGP nodes. Based on settings and restrictions set by an AS's network operation center (NOC), different information can be requested from looking glass services.

Automated querying of looking glass servers is a great challenge. Where BGP routers provide direct access to the routing devices (e.g. via telnet), a more or less consistent interface is available to query the nodes participating in questionable routing by automatic means.
Looking glass servers, however, usually provide web interfaces to access the required information. Such a web interface provides access to at least partially the information available from border routers. The type of information differs between most of these web interfaces, as well as the web interfaces itself. Although usually optically similar, the technical differences of the provided web interfaces make automated querying and information parsing a complex task that often requires human intervention.

Our system queries as much EU located looking glass servers as possible and tries to reach a large coverage. Based on gathered looking glass information, the first hop of each route can be used to verify an indicated physical connection between ASs. Thus, the derived information will be used to mark peering relationships gathered from BGP announcements as confirmed. As another contribution, we provide additionally collected views on the Internet routing and can use them as another source for routes and AS paths for anomaly detection in addition to those, collected from routing archives. When an unavailable looking glass server is found in the list, we inform the provided AS's network operation center to inform them about the orphaned database entry and ask for an alternative address for looking glass access.

Neighbor information from a looking glass interface can help to decide whether routes are valid or not by providing information if those routes can actually exist. Invalid routes can be filtered, if the announced route has no physical counterpart, valid routes can be verified on each edge. Additional services provided from looking glasses such as ping or traceroute provide additional information about the connectivity and availability of BGP infrastructure outside of the Border Gateway protocol itself. As all of these information, BGP and non-BGP, are available from different viewpoints, verification or falsification of announced routes is easier and more reliable by making automated use of these information.

Peering policies

66 of 119 examined IXPs publish information about peering policies of their members. Information about peering policies are used if yet unknown AS paths are announced and examined by our anomaly detection tool. Topological details are required when it comes to a decision, whether a newly announced path is reasonable or not. It seems more likely that ASs peer with each other when they share the same IXP. The data from peeringdb [3] and the

information gathered from the EU IXPs directly indicate that many ASs are located at several IXPs. This increases the number of possible peering relationships in case of an open peering policy. If the peering policy is selective or closed, new peerings are less likely but more stable in general. When we find information on peering restrictions/conditions, e.g. up to date entries within the peeringdb.com database, they are additionally checked by our system. An open AS's peering policy indicates a smaller or non-profit AS, since larger Tier-1 or Tier-2 provider earn money to act as a smaller AS's upstream and have case-by-case or closed policies. We check this information with existing AS relationship classification [12] and obtain more reliable information on ASs' relationships to be used by our anomaly detection and classification.

Whois information

Whois information from regional registries such as RIPE [1] contain details about the owner of an AS. Data gathered from whois services is primarily used by our system to determine the country an AS is located in. For this purpose we firstly request all *descr* and *address* fields and parse them for country information. If no contrary information is found, the AS is counted to the corresponding country. There are few ASs with opposing country information in the address fields of the whois entry. This is likely, when a corporation that operates an AS has several business units, responsible for network operation, located in several countries. Those entries have to be checked and added to the database manually in the current implementation.

Secondly, whois information is used to classify occurring MOAS conflicts. For each AS participating in the conflict the given company or administrator is checked and an affiliation or relationship factor between them is calculated. If our heuristics indicate closer relationship, e.g. the same company name or equal responsible email addresses, a conflict is rather expected legitimate.

Thirdly some ASs provide peering hints in their whois entry. *Import* fields reflect which ASs and Prefixes are imported and accepted from which peer. *Export* fields name the corresponding outgoing rules. If paths are announced that violate those given rules, they are suspicious but not sufficiently illegitimate since a connection between the ASs might yet exist. The information contained in both fields are considered by the heuristic classifier.

In general it should be mentioned, that whois information from RIPE should be considered outdated and unreliable. Nonetheless, this information should not be ignored when classifying hijacking events as long as it is not contradicted by another more reliable source.

Conclusion

Based on the set of derived information we mark announcements (and especially contained routes within them) as confirmed, when it is evidence through our collected data, that such a peering really exists. Our database contains reliable information on peering from publicly available sources, facts on peering policies and historical data to be used by our heuristics.

5. CLASSIFICATION

This section provides a short overview on the improved classification and detection of routing anomalies.

The Classification of each routing announcement takes place before and after the anomaly detection itself. As stated earlier, anomalies can be legitimate or illegitimate. To differentiate between both classes, our classifier uses the additional information that has been gathered in the *improving the ground truth evidence* process.

Prefix hijacking
For prefix hijacking events, additional information from whois services provided by Internet registries is necessary to determine the legitimate owner of an affected prefix. When the legitimate owner is known, all other ASs, involved in this anomaly, will be checked for a (non-routing) relationship between each of them and the legitimate owner. It will be estimated from the information found in the internet registries whois database and on IXP websites as described earlier. If a relation is found and considered reasonable, i.e. AS operators are named similar or the contacts are equal, the anomaly will be classified as rather legitimate.

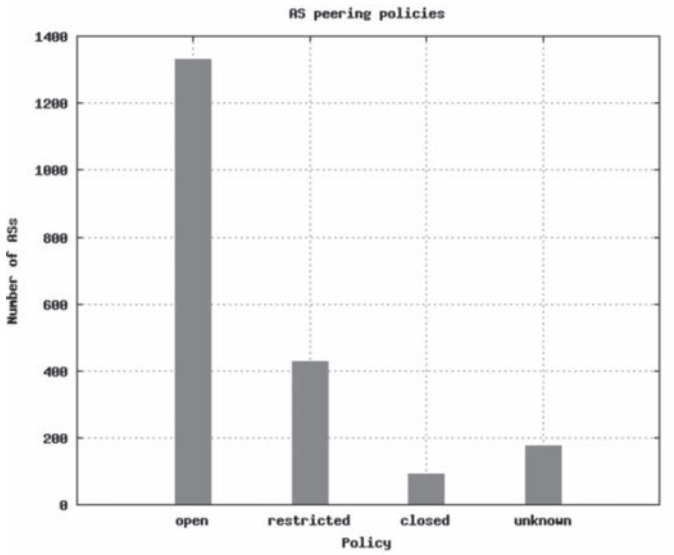
Topology disorder
If a topology disorder is detected in a BGP announcement, the corresponding path will be examined in a special way. First, all the ASs on the path will be checked for historical suspicious behavior and the relationship to predecessor and successor. Additional information about peering relationship between ASs can be used, to mark newly created links harmless. Wrong topological information like an attack against targeted ASs or prefixes can be used to influence routing decisions and lead to the usage of unpredicted paths for affected prefixes.

6. EVALUATION

We collect routing and peering information as described above for our studies. This section describes the collected data in detail and evaluates the impact of enriched ground truth evidence to existing methods and algorithms to observe Internet routing anomalies. One of the most valuable achievements is the decreased number of suspicious peering relationships through reliable evidence of actual connectivity between ASs.

The list of known IXPs located in the EU is used to gather peering relations of the participating members. The database at peeringdb.com lists 3065 ASs connected at these IXPs. We collected 66 lists from websites of 119 IXPs. Using our approach, we have found 5185 ASs as member of these. The difference between our AS list derived from IXPs directly and the list provided by peeringdb for large IXPs is presented in Table 1. Related to the IXP member provided by peeringdb, our system collected 74% more entries in total with assumable higher evidence.

FIGURE 3: AS PEERING POLICIES



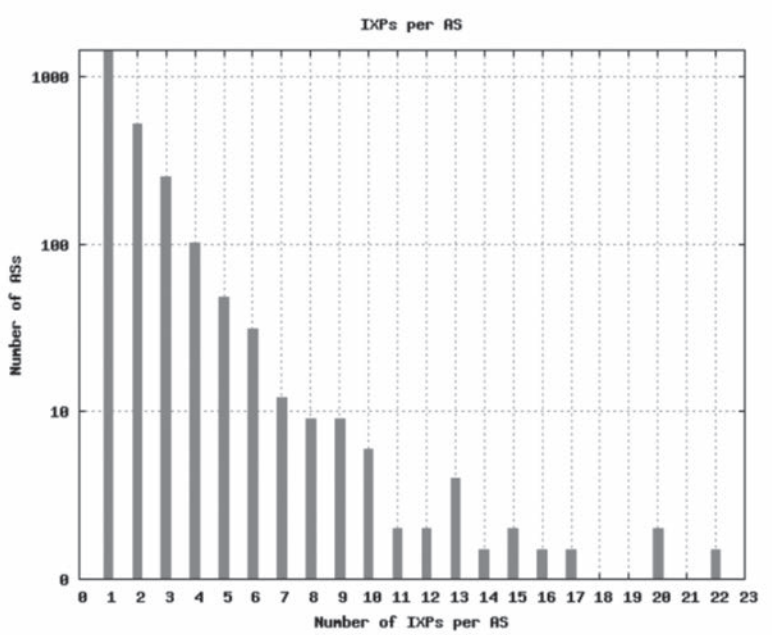
17 of those member lists contained the peering policy of 2024 different ASs used for the classification heuristics as described in the sections above. This additional information increases the number of conflicts being classified as legitimate ones.

TABLE 1

IXP	peeringdb.com	our database
AMS-IX	564	627
DE-CIX	453	515
France-IX	191	449
NL-IX	173	390
V-IX	87	120
Netnod	14	88
DIS-DK	41	42

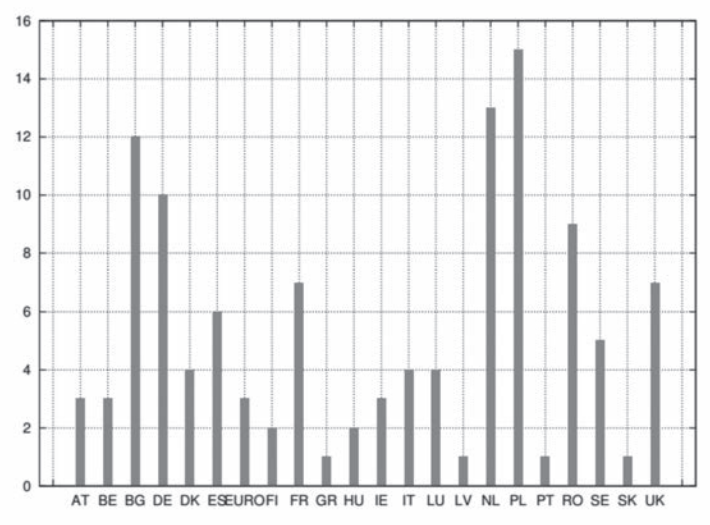
According to Figure 3, 1452 ASs have an open, 454 a restricted, 92 a closed and 175 a currently undeterminable peering policy. The number of IXPs, an AS is located at can be used as an indicator for its size or role inside the Internet routing system and increases the number of potential peering ASs. Therefore we determined the number of IXPs an AS is located at. As shown in Figure 4, most of the ASs are located at few IXPs.

FIGURE 4: NUMBER OF IXPs PER AS



The collection of looking glass URLs and the responses of them allows to gain evidence on direct AS peering relationships. When an unknown peering is found in the analyzed BGP announcements the looking glass servers of both related ASs are queried. If confirmed by the existing looking glasses the information is added into our database. During our research, we examined 116 looking glass servers run by EU-based network operations centers, from which we found 97 to be reachable for querying. Figure 5 shows the number of looking glasses by country. Although all operated within the European Union, most of these NOC's operate BGP nodes around the whole globe, having peering connections with major international operating ASs. Therefore, even a European effort to gather live BGP data from looking glasses provides a useful data base to make sense of anomalies detected throughout the whole Internet. The challenges of gathering those data could be overcome by a common approach supported by the network operations centers.

FIGURE 5: FOUND LOOKING GLASS SERVERS IN THE EU BY COUNTRY



The adoption of looking glass interfaces is still ongoing work.

Especially for the classification of prefix hijacking events the data gathered from RIPE's whois database is used. In case of a hijacking anomaly, the whois data regarding to the owners of affected ASs is considered. When equally or similarly named organizations own all those ASs, the conflict is rather classified as legitimate by the heuristics. Whois data of all 11687 ASs we located inside the EU has been pulled from RIPE to be used in our prefix hijacking classifier.

7. DISCUSSION AND FUTURE WORK

The state of Internet routing is still hard to determine continuously and thus, still vague. The number of involved autonomous systems increase and the number of IP prefixes will massively increase when IPv6 is implemented by all of them. That is why adjusting anomaly detection mechanisms is yet necessary. Our contribution is a larger data basis gathered from primary sources, that are trustworthier sources as those only based on information from within the examined routing system itself, i.e. routing archives, used for identifying and classifying routing anomalies. We created a system to increase evidence of routing information derived from these publicly available sources. This enrichment leads to more reliable detection and classification mechanisms and allows to decrease the number of decisions made on unreliable information. There is no final solution in sight to secure Internet routing at all. Thus network operators and security engineers have to work with continuously improved tools. The work on detection and classification of anomalies is not finally done and we will adjust our solution in the future to become more efficient and to collect more reliable information from primary sources such as IXPs and ASs themselves. To allow statements on the Internet routing state as a whole the restriction to EU-located ASs should be weakened and the number of monitored ASs shall be increased.

REFERENCES:

[1] Ripe network coordination centre (ncc). http://www.ripe.net (29. Nov 2013)
[2] BGP monitor, Nov 2013. http://www.bgpmon.net (29. Nov 2013)
[3] Peering database, http://www.peeringdb.com (29. Nov 2013)
[4] Ripe routing information service (RIS), http://www.ripe.net/data-tools/stats/ris (29. Nov 2013)
[5] Team cymru, http://www.team-cymru.org/Monitoring/BGP/ (9. Nov 2013)
[6] W. Aiello, J. Ioannidis, and P. McDaniel. Origin authentication in interdomain routing. In *Proceedings of the 10th ACM conference on Computer and communications security*, CCS '03, pages 165–178, New York, NY, USA, 2003. ACM.
[7] H. Ballani, P. Francis, and X. Zhang. A study of prefix hijacking and interception in the internet. *SIGCOMM Comput. Commun. Rev.*, 37(4):265–276, August 2007.
[8] K.W. Chin. On the characteristics of BGP multiple origin AS conflicts. In *Telecommunication Networks and Applications Conference*, 2007. ATNAC 2007. Australasian, pages 157 –162, December 2007.
[9] G. Di Battista, T. Erlebach, A. Hall, M. Patrignani, M. Pizzonia, and T. Schank. Computing the types of the relationships between autonomous systems. *IEEE/ACM Trans. Netw.*, 15(2):267–280, April 2007.
[10] R. Dube. A comparison of scaling techniques for BGP. SIGCOMM *Comput. Commun. Rev.*, 29(3):44–46, July 1999.
[11] V. Fuller and T. Li. Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan. RFC 4632 (Best Current Practice), August 2006.
[12] L. Gao. On inferring autonomous system relationships in the internet. *IEEE/ACM Trans. Netw.*, 9(6):733–745, December 2001.
[13] S. Goldberg, M. Schapira, P Hummon, and J. Rexford. How secure are secure interdomain routing protocols. In *Proceedings of the ACM SIGCOMM 2010 conference*, SIGCOMM '10, pages 87–98, New York, NY, USA, 2010. ACM.
[14] S. Kent, C. Lynn, and K. Seo. Secure Border Gateway Protocol (Secure-BGP). *IEEE Journal on Selected Areas in Communications*, 18(4):582–592, April 2000.
[15] C. Labovitz, G. R. Malan, and F. Jahanian. Internet routing instability. *IEEE/ACM Trans. Netw.*, 6(5):515–528, October 1998.
[16] M. Lad, D. Massey, D. Pei, Y. Wu, B. Zhang, and L. Zhang. PHAS: a prefix hijack alert system. In *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15*, USENIX-SS'06, Berkeley, CA, USA, 2006. USENIX Association.
[17] M. Lepinski and S. Kent. An Infrastructure to Support Secure Internet Routing. RFC 6480 (Proposed Standard), February 2012.
[18] David Meyer. Route views archive project.
[19] J. Ng. Extensions to bgp transport sobgp ertificates. Internet-Draft, May 2005.
[20] V. Paxson. End-to-end routing behavior in the Internet. In *Conference proceedings on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '96, pages 25–38, New York, NY, USA, 1996. ACM.
[21] J. Qiu, L. Gao, S. Ranjan, and A. Nucci. Detecting bogus BGP route information: Going beyond prefix hijacking. In *Third International Conference on Security and Privacy in Communication Networks and the Workshops*, pages 381–390, 2007.
[22] Y. Rekhter, T. Li, and S. Hares. A Border Gateway Protocol 4 (BGP-4). RFC 4271 (Draft Standard), January 2006.
[23] T. Wan and P.C. van Oorschot. Analysis of BGP prefix origins during Google's May 2005 outage. In *Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*, page 8 pp., april 2006.
[24] Z. Zhang, Y. Zhang, Y. C. Hu, and Z. M. Mao. Practical defenses against BGP prefix hijacking. In *Proceedings of the 2007 ACM CoNEXT conference*, CoNEXT '07, pages 3:1–3:12, New York, NY, USA, 2007. ACM.
[25] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. An analysis of BGP multiple origin AS (MOAS) conflicts. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, IMW '01, pages 31–35, New York, NY, USA, 2001. ACM.
[26] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. Detection of Invalid Routing Announcement in the Internet. In *Proceedings of the 2002 International Conference on Dependable Systems and Networks*, DSN '02, pages 59–68, Washington, DC, USA, 2002. IEEE Computer Society.
[27] M. Wübbeling. Visibility of Routing Anomalies for End Users. In Christoph Pohl, Sebastian Schinzel und Steffen Wendzel, editors. Proceedings of the Eight GI SIGSIDAR Graduate Workshop on Reactive Security (SPRING). Technical Report SR-2013-01, GI FG SIDAR, Munchen, Februar 2013

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Elastic Deep Packet Inspection

Bruce W. Watson
Dept. of Information Science
Stellenbosch University
Stellenbosch, South Africa
bruce@fastar.org

IP Blox
Kelowna, Canada
bruce@ip-blox.com

Abstract: Deep packet inspection (DPI) systems are required to perform at or near network line-rate speeds, matching thousands of rules against the network traffic. The engineering performance and price trade-offs are such that DPI is difficult to virtualize, either because of very high memory consumption or the use of custom hardware; similarly, a running DPI instance is difficult to 'move' cheaply to another part of the network. Algorithmic constraints make it costly to update the set of rules, even with minor edits.

In this paper, we present *Elastic DPI*. Thanks to new algorithms and data-structures, all of these performance and flexibility constraints can be overcome – an important development in an increasingly virtualized network environment. The ability to incrementally update rule sets is also a potentially interesting use-case in next generation firewall appliances that rapidly update their rule sets.

Keywords: *deep packet inspection (DPI), speed/memory performance, incremental defense*

1. INTRODUCTION

In this paper, we describe a new approach to deep packet inspection (DPI) – known as *Elastic DPI (EDPI)*. Next-generation firewalls (NGFW's, which include intrusion-detection and -prevention systems) usually consist of two main architectural components:

1. *Sensors*, which inspect network traffic (standard TCP/IP traffic, but also network area storage traffic, etc.), reporting back on which 'rules' matched. Sensors have previously been *shallow* packet inspectors for performance reasons, but DPI has now reached line-rate performance – and the *deep* inspection of network traffic makes it an obvious choice of sensor. A typical system will involve DPI instances deployed at various points in the network, perhaps with different rule sets to gain various types of insight into the traffic.

2. *Aggregators* and *correlators*, which take input from many sensors, assembling a broader picture of a threat. There may be multiple levels of aggregation or correlation, each feeding upwards to a more general level that eventually signals an alarm or threat.

Throughout this paper, we focus on DPI sensors. A good overview of the technical details or the field is given in (Varghese, 2005). Ongoing advances in this field are typically covered in conferences such as RAID[1].

Present day DPI is typically static in *where* it is run – either because of large computational needs or semi-custom hardware. Furthermore, while the rules recognized by a particular DPI instance are changeable, such changes (even for a single rule add, edit, or delete) are not fast enough to be done while processing a packet.
These aspects of current DPI (discussed in more detail in the next section) hint at the following problematic use-cases:

1. *Virtualization*. Software defined networking, cloud computing, and in-house virtual machine servers mean that network traffic may be in a virtual network. The soft (as in *software*) nature and dynamic topology of such networks means the traffic is not easily piped through semi-custom (or non-virtualizable) DPI hardware. Even when this is possible, the scale of such networks may overwhelm the DPI instances, making them a performance bottleneck. Ideally, DPI would be performed in COTS hardware that is also virtualizable.
2. *Fine-grained rule updates*. Current DPI offerings can make rapid rule updates as network administrators discover problems. Rule updates typically involve some actions (e.g. 'compiling' the rule set) offline before an update; as such, the updates are not performed *in-place*. The updates are usually not fast enough to be performed *while processing* traffic – i.e. there is some visible latency effect. Such fast updates could have a role in systems where rules are being *learned* automatically; that scenario would involve much more rapid rule generation/editing than done by human network administrators.
3. *Mobile DPI*. For DPI implemented on COTS hardware, load-balancing can be achieved by moving the DPI instances themselves as hardware becomes over-/under-loaded. This is doable today, but typically involves either already having a DPI instance at the destination, or bundling the DPI executable and compiled rules. Ideally, this would be lightweight and fine-grained, where a DPI instance can be partially moved downstream (in the network) – perhaps with the traffic it is already processing.

Our contribution is Elastic DPI (sensor implementation, consisting of algorithms, data-structures and implementation techniques), with several novel characteristics:

1. *Elastic resource consumption*. Most DPI implementations are built upon some form of regular expression ('regex') engine – all of which suffer from large memory or time consumption. Recent advances in data-structures and algorithms for regex processing allow us to adjust the DPI instance dynamically (both upwards and downwards), trading memory for time and vice-versa, even while processing a single packet.

1 Research on Attacks, Intrusions and Defenses, formerly Recent Advances in Intrusion Detection.

2. *Dynamic rule set*. A side effect of elasticity (in particular, JIT) is that the rule set can be edited on-the-fly, also within inspection of a single packet. Other attempts at this have involved recompiling at least part of the rule set, whereas our solution allows for incremental and in-place updates while processing traffic.
3. *Movable DPI engine*. Part of the elastic implementation involves a *domain specific virtual machine* (DSVM) for regular expressions[2]. The DSVM, along with the ability to reduce the memory footprint under EDPI, allows for the physical relocation (migration) of the DPI engine to other locations in the network, for robustness and load balancing. Such moves can be fine-grained, in which a DPI engine which has been partially run over a packet can then be migrated (perhaps after encryption and signing) with the packet itself, after which the DPI run can be completed.

A. Structure of this paper

We begin in the next section with an overview of present-day DPI, focusing on the architectures, rules for inspecting network packets, algorithms in use, implementation technologies, and perhaps most importantly: the performance trade-offs and constraints that arise from these.

The main new results – technical aspects and advances of elastic DPI – are described in the next section, with a focus on how it solves the problematic performance aspects of current solutions. This paper ends with the conclusions and future work.

The reader is expected to have a passing familiarity with regular expressions and/or finite state machines, or programming languages that use them, such as Awk, Perl, Python, etc. For a good introduction, see (Friedl, 2006), which covers both the user perspective and the under-the-hood workings of regex implementations.

2. DEEP PACKET INSPECTION TODAY

This section gives a brief overview of current DPI implementations, with a specific focus on Snort – see (Cox & Gerg, 2004) and (Cisco/SourceFIRE/Snort, 2014). While numerous other systems exist, they bear a general resemblance to Snort. Further coverage of DPI can be found in standard references such as (Varghese, 2005), (Nucci & Papagiannaki, 2009) and (Lockwood, 2008).

A. Architectural overview

Today, all academic and commercial/production DPI systems consist of the same basic architectural building blocks and interactions. These closely resemble a programming tool-chain, with a *programming language*, a *compiler*, and an *execution engine*. Indeed, current DPI systems are a form of *domain-specific language* (DSL[3]) and tools:

- *Rule sets*. These specify precisely what the system is to look for in the traffic. The rules are usually written in some domain specific language in which rule experts express interesting patterns or relationships between portions of the packet, session, flow, etc.

2 Virtual machines in this context are not new – see for example Russ Cox's work in this area (Cox R. , 2009). A comprehensive coverage of virtual machines can be found in (Smith & Nair, 2005), while (Fick, Kourie, & Watson, 2009) covers domain specific virtual machines.

3 Fowler provides a good overview of domain specific languages in (Fowler, 2010), while (Hudak, 1998) is one of the first papers explicitly treating DSL's.

- *Rule set compiler*. The human-readable rules are *compiled* (transformed) to a set of data-structures that are optimized for processing against the traffic. Compilation is usually an offline (batch) task, rerun for each change to the rule set. The compiler may support some form of incremental update, in which minor changes are much less time-consuming. Compilation is run by a network administrator, after which the new data-structures are downloaded to the matching engine.
- *Matching engine*. The precompiled rules (by now often consisting of hundreds of megabytes of data) are run against the traffic by an 'engine', in many cases consisting of specialized hardware to keep up with current network line-rates.

The following subsections consider each of these in some detail.

B. Rule sets

The rule language is a domain-specific language in which a rule engineer (a networking threat expert) can express the patterns in traffic corresponding to a threat. The best known such system is Snort[4], whose rules contain one or more clauses of the following types

- *IP addresses and ports*. A rule can apply to a specific IP address, a range of address, or a mask of addresses. Similarly, ports may be selected.
- *Flags*. A rule can apply to packets with certain flags (un)set.
- *Strings*. A specific string of bytes may be required in a packet; additionally, an offset range can be given, specifying where the bytes must appear.
- *Regex*. Regular expressions can express byte sequences that must appear in the packet, including repetitive sequences, alternative subsequences, etc.
- *Actions*. Most rules include some type of action, such as logging a message, raising an alarm, etc.

The set of such clauses making up a rule can be combined in a Boolean expression, indicating when the rule has matched.

Several observations can be made, given that the rule language is a DSL:

- The structure of all such rule languages is not only a reflection of the domain (as captured during a domain modeling phase before designing the domain specific rule language), but also of the underlying algorithmic and computational model used in the matching engine. We will return to this later as a point for optimization.
- Rules can vary dramatically in granularity, meaning that some rule authors use a one-to-one mapping between threats and rules ('coarse' rules), whereas others favor fine-grained rules in which a threat is made up of several such smaller rules. While this is often a question of style (as in other programming languages), coarser ('fatter') rules can be so complex as to also impede optimization, and therefore performance. The total number of rules in current systems is well over 1000, even with relatively coarse-grained rules.
- Often, rule sets consist of several subsets, each of which are actually written for different applications – e.g. intrusion detection rules, load-balancing rules, and quality of service rules. In the interests of performance, these application-specific

4 We occasionally refer to Snort, however, all major vendor's systems bear a close resemblance to Snort. Snort is used as an example here because it has both open source and commercial versions. (Cisco/SourceFIRE/Snort, 2014)

rules are often combined into a single large rule set for deployment in a single DPI instance. Later, we consider the performance implications of such combinations.

C. Matching engine

(We discuss the matching engine before the rule compiler, as the engine choices determine the compiler's characteristics.) The matching engine is designed with three competing engineering requirements:

1. *Speed*. The maximum bandwidth of the network is a given, and the engine must typically deal with full line-rates. In addition, only limited latency is permitted.
2. *Accuracy*. The engine must faithfully match rules. If the engine becomes overloaded with network traffic, some applications allow for lossy matching, in which some false positives or negatives are allowed.
3. *Price*. Balancing speed versus accuracy is also a price tradeoff. High speed and accuracy is computationally and memory intensive and may require semi-custom hardware.

The rule set (which is a domain specific language and its underlying domain model) in some sense dictate an abstract computational model for the engine – in some sense a domain specific virtual machine (Smith & Nair, 2005). In the case of Snort (which is representative of most DPI systems), there is a combination of two things:

1. *Finite state machines (FSM's, also known as automata)*. These are an efficient representation of string patterns and also regex's[5]. The finite state machines are run over the packet, indicating matches of regex's or strings. There are several types of FSM, with the best known being6:
   a. Deterministic: fast, predictable, but potentially massive memory consumption, or
   b. Nondeterministic: can be much slower and less predictable, but with modest memory needs. Additionally, there are efficient bit-parallel versions of these, which can be fast but require wide bit-vector machines or custom hardware.
2. *Decision trees*. Most of the other clauses in Snort rules are best compiled into decision trees, which are then evaluated by the engine in conjunction with what is found by the finite state machines.

These data-structures are loaded into the engine at startup time, and are not easily modified on-the-fly. The engine is typically so performance constrained (barely enough clock cycles for line-rate traffic) that on-the-fly optimizations and modifications are rarely done. This means any data-structure optimization is necessarily pulled into the compilation phase. One optimization that would be ideal (though not realized in current DPI implementations) is *incremental construction* of the data-structures in the match engine – only fleshing out those parts that are actually needed while processing traffic.

5 Regex's are typically compiled into FSM's using any of a number of algorithms, many of which are found in compiler textbooks such as (Cooper & Torczon, 2012). To date, the most comprehensive (and only taxonomic) treatment of such algorithms is (Watson, 1995). Dozens of regex compilation algorithms were devised in the years 1958-1995, with only modest advances since then. As a result, such a taxonomy remains a good overview of the field, despite its age.

6 Most interesting alternatives require additional hardware support, such as TCAM memory, etc. Given that one of our requirements is to push for COTS implementation, we avoid such state machines here.

In the engine, FSM's and decision trees can both be implemented along the following spectrum: pure/portable software, accelerated software (GPU[7]), FPGA[8], ASIC[9]. That spectrum is increasing in price, performance and time-to-market, but decreasing in flexibility. With the aforementioned performance and accuracy requirements, line-rate DPI engines often involve FPGA's or ASIC's, as well as highly optimized data-structures, significantly reducing flexibility.

D. Rule compilers

With the rule language defined and the matching engine's computational model selected, rule compilation is a straightforward problem of producing the correct data-structures. As with general purpose programming languages, the optimizer in a rule compiler is the most time consuming component: ideally, all of the rules are compiled together and co-optimized. Editing, adding or removing, even a single rule therefore requires an incremental recompile step and perhaps a global re-optimization step. Compilation is also very ill suited to running on the hardware hosting the matching engine (which is geared to high performance traffic stream processing) – reinforcing the notion that this is a task for the network administrator. An overview and taxonomy of algorithms involved in regex compilation can be found in (Watson, 1995), though several conferences cover new developments in such algorithms (e.g. the *International Conference on Implementations and Applications of Automata*).

E. Performance tradeoffs and constraints

The performance tradeoffs in current systems can be summarized as:

1. Current rule sets consist of >1000 rules, and growing. Rule sets often consist of subsets for different application areas. In practice, they are compiled together, yielding data-structures. An alternative solution would be to separate them and compile and deploy separate match engines – leading to at least partial duplication of data-structures.
2. Deterministic FSM for regex's: fast, can be implemented on standard hardware, but can require exponential memory against the number of rules.
    a. Cheap execution unit (can be standard CPU) for the engine.
    b. Potentially exponential memory costs.
    c. Can require exponential running time and memory for compilation, giving very slow update time when rules are edited.
3. Nondeterministic FSM for regex's: fast, but only when implemented with bit-parallelism on wide bit-vector custom hardware; memory requirements linear in the size of the rule set.
    a. Expensive execution unit consisting of custom hardware.
    b. Cheap memory for linear-sized FSM.
    c. Compilation is usually quadratic in the rule set size – still too slow for incremental updates after rule edits.
4. With both types of FSM, compilation is a network administrator task, and the resulting data-structures are relatively static once moved to the match engine. As such, *all rules* in the set are present (in compiled form), even if they are not used,

7 Graphics Processing Unit – e.g. from NVIDIA. Numerous papers have been written on network processing acceleration using GPU's.

8 Field Programmable Gate Array – a 'soft' silicon chip which is 'programmed', e.g. from Xylinx or Altera. Network processing acceleration using FPGA's is covered in (Lockwood, 2008).

9 Application-Specific Integrated Circuit – essentially a custom silicon chip. ASIC solutions to DPI are typically proprietary or secret.

they conflict, or are from different rule application areas. This can be a significant system overhead, given that practical situations see only a fraction of the total rule set in use while processing typical network traffic. (Of course, that is alleviated when DPI runs on a system with virtual memory and not all data-structures are in physical memory at a time.)

5. Many rule languages use Perl-compatible regex's (PCRE's). Pure regex's compile and optimize very well for FSM's, but PCRE's contain numerous features (such as backtracking, greedy operators, etc.) that impede the match engine's implementation and performance. As a result, rule writers shy away from regex unless absolutely needed, preferring to use the other rule clauses – making the rules very heterogeneous and difficult to optimize (Friedl, 2006).

These tradeoffs have some DPI-system-wide implications:

- DPI is not suited to a virtualized environment:
  - Deterministic FSM: the match engine uses COTS hardware, but with high memory consumption (incompatible with virtualization, in which the virtual machines are expected to not appropriate all resources).
  - Nondeterministic FSM: the match engine uses custom hardware not found in a virtual environment.
- For similar reasons, it is not movable, even in a virtualized environment. Either the system is consuming large amounts of memory (making it costly to move), or using custom hardware (impossible to move).
- In the deterministic FSM scenario (the most common one in practice), rule set edits do not allow for incremental compilation (where only the impacted parts of the data-structures change). The illusion of incremental compilation is given by some systems – though this is accomplished by compiling a separate set of tables for the rules that have changed, thereby further raising system overheads as those new FSM's must also be run over the packet.

3. ELASTIC DPI

Elastic DPI uses recent advances in algorithms and data-structures (for regex's and FSM's) to provide solutions to the problems sketched in the last section.

A. Simplifying the rule language

As mentioned earlier, two of the performance penalties in DPI systems are the use of: elaborate rule structures (e.g. thanks to the different clause types in Snort rules) that require decision trees, and regex's, specifically PCRE.

In EDPI, we have chosen to only use regex's and actions in rules:

- Regex's can be used to express IP address, port and flag aspects that must match. In the match engine, the regex is run against the entire packet, including any headers and trailers containing such information.

- Strings, including their offsets within the packet, are written as regex's. Indeed, also in Snort string clauses are actually a form of regex in a different notation, as offsets are readily written in regex's as well using counting quantifiers (Friedl, 2006).
- The dialect of regex's chosen is much purer than PCRE, leaving out the computationally heavy backtracking and capture mechanisms[10]. In return, our dialect allows for exotic extended regex operators such intersection, negation, shuffle, cut, etc., which gain more than enough expressive power. Those operators allow us to directly combine what would previously have been multiple clauses and Boolean expression in the rule, yielding a single regex for the rule. In fact, the rule compiler merges all of the rules' regex's into a single large regex (of the form Expression1 | Expression2 | …). See (Brzozowski, 1964) for more on compiling extended regular expressions to FSM's.

This unification of rule notation, and underlying computational formalism is both elegant (rule writers can think in one formalism) and also computational efficient, as discussed below.

B. Speed versus memory

In this section, we detail three groups of algorithmic, implementation, and optimization techniques that, independently, are already significant advances, but together are key enablers for Elastic DPI.

1) On-demand construction

As mentioned earlier, most current compilers from regex's to FSM's are *batch compilers*, meaning they compile the entire regex (in our case, the composite regex consisting of all rules) into a single massive FSM without regard to which parts of the FSM will actually be used. At run time, usually only a fraction of the FSM is used (because not all DPI rules match over the traffic) – imposing an unfortunate system overhead. Ideally, we would like to only build those parts actually in-use – a kind of *hot state/path* optimization. Such algorithms have been known since the early days of regex and FSM implementation (Thompson, 1968). In DPI systems, for performance the match engine is often running on hardware highly tuned for the matching process, or COTS hardware fully devoted to DPI – not the environment in which to run the compiler or perform such on-the-fly construction. EDPI rests on a new class of algorithms and data-structures that are fast enough for on-the-fly construction and optimization while simultaneously performing matching.

A regex/FSM co-representation is presented in (Watson, Frishert, & Cleophas, 2005) and (Frishert & Watson, 2004), and we have extended that work for EDPI. The algorithm (our *continuation engine*) takes two parameters: a regular expression to be matched, and an input byte of the traffic. It returns another regular expression, known as the *continuation*. Essentially, the continuation[11] encodes the 'remainder' of the pattern to be matched in the input, and computing the continuation is equivalent to taking a transition in an FSM corresponding to the regex. Continuations date to Janusz Brzozowski's original work in this area in the late 1950's (Brzozowski, 1964), though the algorithm has been oddly underused in compilers and other applications.

10 Those mechanisms are not only computationally heavy, but also nondeterministic, making them problematic when making real-time performance promises, as are required in DPI.

11 Also known in the literature as the derivative.

In our *continuation engine (CE)*, we have made two important optimizations over Brzozowski's original work:

1. The continuations (over all possible input bytes) of a regex share most subexpressions with the original regex. As such, we can apply *common subexpression sharing* – a well-known technique in compilers (Cooper & Torczon, 2012) – to dramatically reduce space. In addition to this effect in continuations, many rules in a rule-set share subexpressions (Massicotte & Labiche, 2011) – leading to further savings under EDPI.
2. As continuations are generated (by processing traffic), they are *cached* in lookup tables and do not need to be recomputed. Whenever a continuation is needed which has not yet been computed, the cache entry is empty and it is computed once-off relatively cheaply.

These two techniques allow us to process the input traffic (taking transitions in FSM terms, but actually computing continuations in CE terms) while effectively only building those parts of the FSM that are actually in-use.

There are two performance implications:

1. *Startup costs*. With an initially empty cache, every traffic byte processed triggers a continuation computation in the CE. This continues until a the cache consists of the 'hot states/path' – a critical mass of reusable cache entries is reached. In many DPI applications, this occurs within the first megabyte of input traffic. With suitable traffic profiles, such *cache preloading* is something that can be done offline, saving startup costs.
2. *Processing costs*. Most traffic bytes processed result in a cache lookup – essentially an FSM transition, making this as fast as any other FSM-based solution, providing the cache implementation is highly tuned. Occasionally, a cache miss occurs, giving some overhead in building a new continuation and cache entry. With buffer management in EDPI, the latency from a cache miss is smoothed, and this does not cause any throughput or latency issues. In the worst case, the CE can 'cache thrash', consuming as much memory as a traditional DPI system and having some startup latency[12].

These caching techniques were explored in (Thompson, 1968), but became less interesting as available memory grew. More recently, the performance has been quantified in (Ngassam, Watson, & Kourie, 2006).

These performance characteristics make EDPI competitive with traditional DPI in practice[13], also because hot path optimization (computing only those FSM parts that are actually in-use) reduces total memory load and improves processor cache (not to be confused with CE cache) utilization. To contrast, EDPI can have as little as a few kilobytes in use (representing the regex rule set and some caching) whereas traditional DPI has megabytes of memory in use at a given time for a comparable rule set.

12 This worst-case scenario would amount to pulling some of the compilation costs of traditional DPI into the match engine area of the system, since one of EDPI's architectural advantages is to support compilation in the match engine via the CE.

13 despite the overhead of the continuation engine.

2) Restricting memory

Like other caches in computational systems (e.g. memory and disk caches), the CE's cache can be flushed without errors, but with a computational penalty for rebuilding. The match engine's memory budget may be reduced during processing (that is, in a 'hot/live' system). The CE will discard the cache entries, leaving the entries 'undefined' and triggering recomputing the continuations later. In memory constrained systems, such cache flushing can also be done selectively – when memory is full and a new continuation is being constructed, the least-recently used cache entry is flushed. Least-recently used is tracked using the time-stamp (clock) counter present on most modern processors.

Flushing some cache entries frees up memory used for the transitions, but additional memory may also be freed. In particular, the representation of the additional derivatives (the continuations) consumes memory – even with common subexpression sharing. The CE marks the original regex (as opposed to the continuations, which are *derived* from the original regex), and can discard the non-original regex's (the continuations) when reclaiming memory, since the continuations are easily reconstructed by the CE. This is particularly useful for reducing the state and regex set to a *kernel* that can then be moved to a new compute location (perhaps with the packet being processed), the CE then reconstructing the cache at the new location.

3) Approximate EDPI

Using cache management techniques similar to those in the previous section, EDPI also allows for *approximate DPI* (also known as lossy matching) in very memory constrained systems. That is not presented here, but may be found in (Watson, Kourie, Ketcha, Strauss, & Cleophas, 2006).

4) Stretching and jamming

Stretching and jamming are two additional optimization techniques (they are the reverse of each other) that can move the FSM along the speed versus memory axis (de Beijer, Cleophas, Kourie, & Watson, 2010). DPI typically processes the traffic an 8-bit byte at a time, implying that regex's are also expressed as bytes, and the FSM is represented with transitions on bytes. Stretching allows us to process 4-bit nybbles at a time – each byte of the traffic is separated into a high- and a low-order nybble; the high-order nybble is processed through an FSM transition, followed by the low-order nybble. (The order of the two nybbles can be swapped for processing, giving an endianness optimization which sometimes yields faster processing – though this has not yet been quantified.)

Splitting traffic bytes into nybbles is done on-the-fly – a very fast operation on modern processors. The FSM, however, needs some preparation, with each transition on a byte being stretched into two transitions on the corresponding pair of nybbles. The FSM must typically also be made deterministic again: an FSM state with transitions on bytes b_0 (= nybbles n_{0high} and n_{0low}) and b_1 (= nybbles n_{1high} and n_{1low}) is deterministic when b_0 and b_1 are different, but in the stretched transitions we may have $n_{0high} = n_{1high}$, making the FSM nondeterministic. Stretching doubles the number of steps required to process the traffic, so what does stretching gain us? The narrower alphabet (4-bit nybbles) yields much narrower transition tables: 16 columns now compared to the 256 columns for the full 8-bit byte alphabet, and this is often

a significant space savings despite doubling the number of transitions and adding new states. Jamming is the opposite transformation, changing the alphabet from 8-bit bytes to 16-bit short-words. This equates to processing two adjacent traffic bytes at a time by merging two subsequent transitions – halving of the processing time[14]. This speed win is traded against the fact the transition tables may now have $2^{16} = 65536$ columns compared to 256 – a massive increase, despite the halving of transitions and reduced number of states. Both stretching and jamming may be applied again, respectively yielding transitions on 2-bit half-nybbles or on 32-bit words, etc.

The optimization sweet spot for stretching and jamming is difficult to find a priori, though some benchmarks are presented in (de Beijer, 2004). The CE in EDPI allows us to dynamically stretch and jam, by rebuilding the FSM in stretched or jammed form (under the hood, the FSM is actually modified in places where it has already been built), based on speed or memory requirements at that moment. A key future optimization is for the CE to locally stretch and jam – an optimization for only part of the regex and FSM where it may be particularly profitable.

C. Incremental rule set modifications

The co-representation of the regex set with the FSM, along with CE, has a critical side-effect: the regex's may be edited on-the-fly. Parts of the rule regex's (which are, in turn, parts of the combined regex) can be added, deleted, or modified. The CE observes this and discards those parts of the continuation data-structures that are no longer valid; they are then rebuilt as needed. This elegant solution brings incremental rule set modification to EDPI, even in running ('hot') systems. Of course, if the modified regex is to be rerun on the packet, this requires backtracking to the beginning of the packet – a relatively small penalty for incremental rules.

D. Location flexibility

The ability to move a running DPI match engine to another machine is a natural side effect of two EDPI aspects:

- EDPI can be virtualized (thanks to more virtualization-friendly memory consumption), and the resulting virtual stack is easily migrated within existing hypervisor products.
- Even without virtualizing EDPI, the data-structures are shrinkable to a kernel – as mentioned earlier. Without harming the matching process, the data-structures can be shrunk to the size of the original regex set (usually measured in *kilobytes*), which can then be moved along with the partially-processed packets, and restarted at a new location. This allows for a form of data-flow architecture with the computation (EDPI instance consisting of the CE) traveling with the data to more appropriate locations (in terms of load balancing, for example).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we gave an overview of the current state of deep packet inspection (DPI) systems, with a particular focus on their engineering tradeoffs and potential performance problems in an increasingly virtualized environment. Against that backdrop, we presented Elastic DPI as a new approach with some key differentiators:

14 There are some nontrivial issues that we do not discuss here, for example: a packet consisting of an odd number of bytes must be specially handled with the last byte which has nothing to jam with.

1. The amount of memory in use can be grown or shrunk dynamically, trading speed against memory consumption. This is key to enabling virtualization.
2. The set of DPI rules may be edited on-the-fly, allowing for highly dynamic systems.
3. The actual DPI engine can be sufficiently shrunk (in service, while processing) to be moved efficiently to another computing resource.
4. The domain specific language for expressing rules can be made uniform, in terms of extended regular expressions which capture all of the presently used clauses in other rule languages.

These aspects are significant advances in this field and are made possible by recent advances in the algorithmics of pattern matching, as well as new implementation techniques.
The primary direction for future work is to integrate and measure the Elastic DPI system in a production environment, yielding benchmarking data. Additional foci are on parallelism in the Elastic DPI algorithms – especially given the current trends towards multicore hardware.

Acknowlegements:
The anonymous referees provided particularly strong, useful and well thought out feedback, which is appreciated.

BIBLIOGRAPHY:

Watson, B. W. (1995). *Taxonomies and Toolkits of Regular Language Algorithms* (Ph.D dissertation ed.). Eindhoven: Eindhoven University of Technology.

Thompson, K. (1968). Regular expression search algorithm. *Communications of the ACM , 11* (6), 419-422.

Brzozowski, J. (1964). Derivatives of Regular Expressions. *Journal of the ACM , 11* (4), 481-494.

Varghese, G. (2005). *Network Algorithmics*. Morgan Kaufmann.

Watson, B. W., Frishert, M., & Cleophas, L. (2005). Combining Regular Expressions With Near-Optimal Automata. In A. Arppe, L. Carlson, K. Linden, J. Piitulainen, M. Suominen, M. Vainio, et al , & A. Copestake (Ed.), *Inquiries Into Words, Constraints And Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday* (Online ed., pp. 163-171). Stanford: CSLI Studies in Computational Linguistics, Stanford University.

Cox, K , & Gerg, C. (2004). *Managing Security with Snort and IDS Tools*. O'Reilly.

Smith, J. E., & Nair, R. (2005). *Virtual Machines*. Morgan Kaufmann.

Nucci, A , & Papagiannaki, K. (2009). *Design, Mesurement and Management of Large-Scale IP Networks*. Cambridge University Press.

de Beijer, N., Cleophas, L., Kourie, D , & Watson, B. W. (2010). Improving Automata Efficiency by Stretching and Jamming. In J. Holub, & J. Zdarek (Ed.), *Prague Stringology Conference* (pp. 9-24). Prague: Czech Technical University.

Watson, B. W., Kourie, D , Ketcha, E., Strauss, T., & Cleophas, L. (2006). Efficient Automata Constructions and Approximate Automata. In J. Holub (Ed.), *Prague Stringology Conference* (pp. 100-107). Prague: Czech Technical University.

Cox, R. (2009, December). *Regular Expression Matching the Virtual Machine Approach*. Retrieved from Regexp2: http://swtch.com/~rsc/regexp/regexp2.html

Cooper, K. D., & Torczon, L. (2012). *Engineering a Compiler* (Second ed.). Morgan Kaufmann.

Frishert, M , & Watson, B. W. (2004). Combining Regular Expressions with Near-Optimal Brzozowski Automata. In K. Salomaa (Ed.), *Conference on Implementations and Applications of Automata*. Kingston: Queen's University Press.

Fick, D., Kourie, D. G., & Watson, B. W. (2009). A Virtual Machine Framework for Constructing Domain Specific Languages. *IEE Proceedings - Software*, 3 (1).

Ngassam, E. K., Watson, B. W., & Kourie, D. G. (2006). Performance of Hardcoded Finite Automata. *Software - Practice & Experience*, 35 (5), 525-538.

Ngassam, E. K., Watson, B. W., & Kourie, D. G. (2006). Dynamic Allocation of Finite Automata States for Fast String Recognition. *International Journal of Foundations of Computer Science*, 17 (6), 1307-1323.

Massicotte, F., & Labiche, Y. (2011). An Analysis of Signature Overlaps in Intrusion Detection Systems. *41st International Conference on Dependable Systems & Networks* (pp. 109-120). Hong Kong: IEEE/IFIP.

Lockwood, J. W. (2008). Network Packet Processing in Reconfigurable Hardware. In S. Hauck, & A. Dehon, *Reconfigurable Computing* (pp. 753-778). Morgan Kaufmann.

Friedl, J. E. (2006). *Mastering Regular Expressions* (Third ed.). O'Reilly.

Cisco/SourceFIRE/Snort. (2014). *Snort Homepage*. Retrieved March 17, 2014, from Snort: www.snort.org

Fowler, M. (2010). *Domain Specific Languages*. Addison-Wesley.

Hudak, P. (1998). Domain Specific Languages. In P. H. Salas, *Handbook of Programming Languages Little Languages and Tools* (Vol. 3, pp. 39-60). MacMillan.

de Beijer, N. (2004). *Stretching and Jamming of Automata* (M.Sc thesis ed.). Eindhoven: Eindhoven University of Technology.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

An Automated Bot Detection System through Honeypots for Large-Scale

Fatih Haltaş
Cyber Security Institute
The Scientific and Technological Research
Council of Turkey
Ankara, Turkey
fatih.haltas@tubitak.gov.tr

Abdulkadir Poşul
Cyber Security Institute
The Scientific and Technological Research
Council of Turkey
Kocaeli, Turkey
abdulkadir.posul@tubitak.gov.tr

Erkam Uzun
Computer Engineering
TOBB University of Economics
& Technology
Ankara, Turkey
euzun@etu.edu.tr

Bakır Emre
Cyber Security Institute
The Scientific and Technological Research
Council of Turkey
Kocaeli, Turkey
bakir.emre@tubitak.gov.tr

Necati Şişeci
Cyber Security Institute
The Scientific and Technological Research
Council of Turkey
Kocaeli, Turkey
necati.siseci@tubitak.gov.tr

Abstract: One of the purposes of active cyber defense systems is identifying infected machines in enterprise networks that are presumably root cause and main agent of various cyber-attacks. To achieve this, researchers have suggested many detection systems that rely on host-monitoring techniques and require deep packet inspection or which are trained by malware samples by applying machine learning and clustering techniques. To our knowledge, most approaches are either lack of being deployed easily to real enterprise networks, because of practicability of their training system which is supposed to be trained by malware samples or dependent to host-based or deep packet inspection analysis which requires a big amount of storage capacity for an enterprise. Beside this, honeypot systems are mostly used to collect malware samples for analysis purposes and identify coming attacks.

Rather than keeping experimental results of bot detection techniques as theory and using honeypots for only analysis purposes, in this paper, we present a novel automated bot-infected

machine detection system BFH (BotFinder through Honeypots), based on BotFinder, that identifies infected hosts in a real enterprise network by learning approach. Our solution, relies on NetFlow data, is capable of detecting bots which are infected by most-recent malwares whose samples are caught via 97 different honeypot systems. We train BFH by created models, according to malware samples, provided and updated by 97 honeypot systems. BFH system automatically sends caught malwares to classification unit to construct family groups. Later, samples are automatically given to training unit for modeling and perform detection over NetFlow data. Results are double checked by using full packet capture of a month and through tools that identify rogue domains. Our results show that BFH is able to detect infected hosts with very few false-positive rates and successful on handling most-recent malware families since it is fed by 97 Honeypot and it supports large networks with scalability of Hadoop infrastructure, as deployed in a large-scale enterprise network in Turkey.

Keywords: *Botnet, honeypots, NetFlow analysis, machine learning*

1. INTRODUCTION

Attackers, astutely, perform their attacks in a well-organized and automated way by leveraging infected zombie machines, for which, enterprise network is preferable basin [1], [2]. Since, infected machines are key players in Cyber-attacks, cleansing them is one of main goals for Active Cyber Defense Systems, thereby, initial step is inevitably, diagnosis. In other words, effectuating a practical and scalable system with capability of withstanding expeditiously growing and enhancing malwares to identify infected machines in an enterprise network has high priority in technologies to be improved within the technical domain of Active Cyber Defense.

There has been extensive work on identifying infected machines, mostly rely on host-based analyses that are feeble against today's malwares with complicated hiding techniques. Acknowledging that they retain a significance role in intrusion analysis, resting only on them can be imprudent as many intruders run wild inside the network in which host machines are armored with at least a couple of host-based security solutions, while those solutions do not provide any clue to system administrators.

In the meantime, numerous researches suggest the use of network related data to detect infected machine or benefit them as auxiliary to host-based systems [3], [4] and [5]. Some detection methodologies might require deep-packet inspection that is overcharge for an enterprise and not successful in the scenario of encrypted communication preferred as command and control channel by malwares. Availability of raw data and time-scalability of processing DPI data are important obstacles to deploy an automated detection system within an enterprise network. To surmount these issues, some of detection system methodologies ([6], [7], [8]) are developed to identify infected machines by using NetFlow standard data that is widely stored in an enterprise network [9].

Because of the limited information within NetFlow data, researchers should conduct a wise statistical analysis to conclude it with malicious activity detection. For that matter, some researchers suggest to include malware families' statistical NetFlow values by leveraging the machine learning techniques and training their systems through beforehand-created models [6], [8]. Some of the challenges here are creating a successful model by using malwares that might feed detection system featly and feature selection that creates utilitarian models. Moreover, they are mostly lack of being deployed in an enterprise network because of modeling unit that requires to be trained by most recent malwares and should be kept up-to-date.

Aforementioned limitations of current automated bot-detection technologies and stealthiness of recently introduced bots, which are not only send spam or conduct DoS attack but also steal sensitive data over encrypted C&C channels [1], [10], inspire us to design a more applicable, scalable and self-updated automated individual bot detection system with high detection rate, indeed, it was a corollary of a need for such system to an enterprise network in Turkey.

In this paper, we present BFH (BotFinder through Honeypots) automated bot-infected machine detection system, based on BotFinder [8], relying on exclusively NetFlow data and leveraging the capability of Honeypots on collecting topical malware samples and utilizing the scalability of Hadoop infrastructure and MapReduce programming logic. In particular, our system consists of three important units, which are Cyber Threat Monitoring Unit, modeling and matching units which trade on Hadoop system.

Cyber Threat Monitoring System (CTMS) unit is, basically, a comprehensive system, developed by our team within the scope of European Unions SysSec project [11]. In BFH system, we benefit its capability of collecting and classifying most recent malwares through 97 honeypots, beside this; it is cultivated with the extension of an aptitude for NetFlow generation of malware families. In a nutshell, this produces preliminary data in order to feed modeling unit.

Modeling and matching units of BFH are implemented based upon BotFinder's methodology with an additional feature analysis. Multi-faceted models are acutely crafted after execution of samples for each malware family in a controlled environment, handled as component of CTMS, through using NetFlow-based features that characterize a malware family communication pattern and by identifying similarities in following demeanors; (i) temporal behavior of flows, (ii) outgoing and incoming data size characteristics, (iii) duration of connections, (iv) communication regularity, (v) data accumulation regularity. These features are also calculated, during trace extraction part, on NetFlow data of investigated enterprise network and used in matching unit and worked out to identify bot-like machine activities. Since an enterprise network consists of a numerous number of hosts and large amount of flow records that should be stored long for better results, our system leverages the Hadoop infrastructure and map-reduce programming logic[12].

An extensive evaluation of BFH is provided in a large-scale enterprise network in Turkey, BFH is deployed. Modeling unit of BFH is automatically trained by caught and classified malwares which are still active, at least in Turkish Networks, as they are caught through 97 Honeypots, that are live more than four months. Based on models, BFH runs over subjected enterprise

network whose pcap data is logged for affirmation purposes for a month. Our evaluation demonstrates that BFH is able to detect malicious activity in the network traffic of bot-infected machines with high accuracy in a reasonable scale for an enterprise. In substance, contributions of this paper are as follows:

- We introduce BFH (BotFinder through Honeypots); a vigilant automated bot-detection system, which leverages capability of Honeypots on collecting recent malwares and scalability of Hadoop infrastructure to increase applicability to an enterprise.

- We present BFH (BotFinder through Honeypots) that strengthen BotFinder's model generating approach with extra feature analysis, examining similarities of stolen data size over time in C&C communication of a bot family.

- We consolidate that C&C communication traffic of bot families has some similarities, even in most recent bot families in the wild as they are caught lively and exploit these similarities on detecting bot-infected machine by only analyzing NetFlow data that provides successful detection even on encrypted or obfuscated traffic.

2. RELATED WORK

Botnet detection studies over network data include multiple approaches. However, to our knowledge, honeypots are not actively involved in the individual bot detection systems though yet they have been benefited. BotMiner [13], BotGrep [14] and BotTrack [7] typify the approach on correlating NetFlow data and detect P2P bots through their C&C topology. They propose to identify the hosts that build up P2P networks by clustering them and discriminate rogue and benign groups benefiting the information on infected machines, gathered from several sources such as IDS and honeypots. On that sense, they are instances of bot detection systems, utilizing the use of Honeypots. However, they are restricted with IDS signatures, which may be insufficient as attackers evolve bots shrewdly to be more disguised.

Aforementioned studies do not only capitalize on NetFlow analysis. Indeed, there exist only a few papers, specifically focusing on it. For instance, Livadas et al. [15] focuses on IRC-based botnets through classification methodology based on machine-learning. Francois et al. [7] leverages PageRank algorithm on NetFlow-based approach to detect P2P botnets [16]. They both focus on particular type of traffic.

On the other hand, BotFinder detects malware infections by exploiting traffic patterns characteristics of them, yet it should be extended to detect malwares, performing non-periodic communication patterns. While not conclusive, BFH proposes a way of smoothing this out by additional feature analysis. BFH provides practicality in an enterprise network by keeping training module updated via Honeypots. It is also a significant illustration of applicability of BotFinder that BFH is a live system with some improvements, deployed to an enterprise network. Furthermore, BFH upgrades infrastructure for scalability concern to Hadoop and processes data.

Lastly, Disclosure suggests a distinct approach to detect botnet over large-scale NetFlow analysis [6]. Disclosure exhibits similar approach to BotFinder on which BFH is based upon, yet, it detects C&C servers. Giroire et al. [17] has similar approach to BotFinder, albeit, it differs in malware detection methodology.

3. SYSTEM OVERVIEW

BFH operates in two phases as training and investigation. Detection models based on statistical features are generated for each malware families in training phase. Investigation phase includes extracting same statistical feature extraction for test data and matching unit which compares test data with each of the malware family models to detect whether incoming data belongs to an infected machine or not.

Figure 1 shows an overview of the system architecture. In the training phase, after collecting malwares honeypots, a classification unit classifies the malwares in different families. Then, NetFlow data is generated after executing samples of each malware families. Afterwards, the trace extraction is conducted to the NetFlow data of each malware family and ordered connections are listed between internal and external IP addresses on a given destination port. After extracting trace data, six statistical features are calculated for each member of families. These features are the average time between the start times of two consecutive flows in the trace, the average duration of a connection, the number of bytes on average transferred to the source and to the destination, the Fourier Transformation over the flow start times in the trace and the ratio of outgoing data difference over time difference between the start times of two subsequent flows [8]. In the modelling unit, multiple binary classification models for each malware families are created by combining all the feature vectors of the members of corresponding family.

Finally, in the matching unit, each of the produced feature vectors for evaluated traces is subjected to classification via all detection models that are created in the training phase with a particular clustering algorithm in a sequential fashion. If any of the detection models raise an alert for an examined trace in the matching unit, it is assumed that the internal IP in this particular trace is infected.

4. SYSTEM DETAILS

A. Cyber Threat Monitoring System (CTMS)

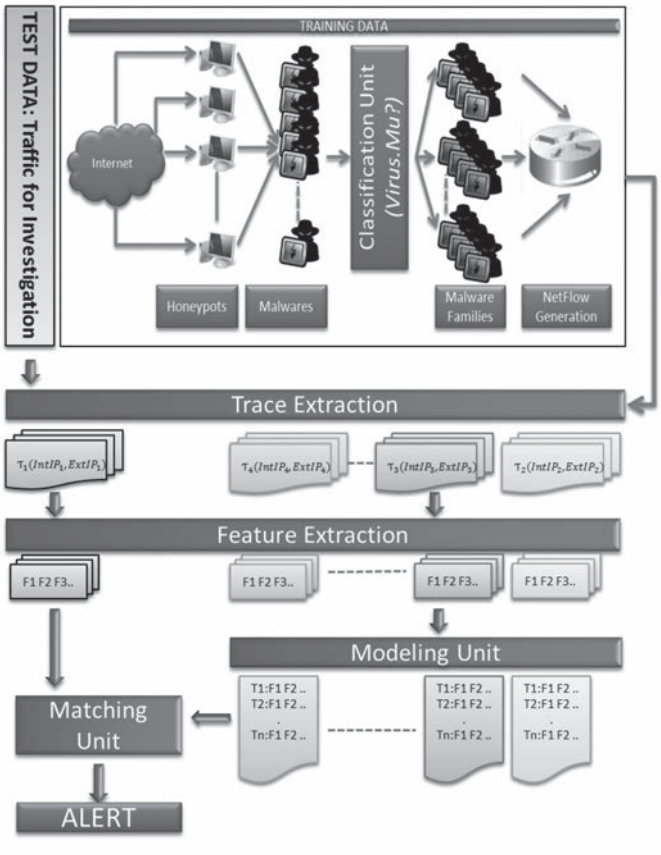
The input data we need for our detection models is collected through distributed sensors located in wide area as traffic capture. For collecting this input data we propose an infrastructure (CTMS) which comprises to two main parts; distributed sensors and malware detection centre. Malware detection centre is composed of sub modules such as virtualization servers which are hosted low and high interaction honeypots, network traffic monitoring systems such as NetFlow collection and aggregation unit, IDS and anti-virus scanners. In this step, it is important to correctly classify the collected input data through honeypots so that different samples of same malware family are analysed together. Thus, an actual classification unit which includes anti-virus scanners is used in this work.

B. Honeypots, NetFlow Generation and Classification Unit

1) Honeypots

The main feature of a honeypot is to collect attack records and malware samples by imitating networks, network services, operating systems or applications. Honeypots are classified depending on their abilities as low and high interaction or their roles as server sided and client sided. In our work, we use four types of honeypots as collectors (Col_A , Col_B) and generators (Gen_A , Gen_B) which are responsible for catching malwares from internet and generating malware communication, respectively. While URLs and attachments of spam mails and web crawlers are used as source for Gen_A , Gen_B and Col_A honeypots, other malwares are captured by Col_B honeypot. Gen_A honeypot is responsible for the execution of spam mail attachments whereas Col_A honeypot runs detected URLs from spam mails. Detailed explanations about these honeypots are as follows;

FIGURE 1: SYSTEM OVERVIEW



- Gen_A: Windows XP operating system running on a virtualization environment. Three Gen_A high interaction honeypots are used in our environment.
- Gen_B: Windows XP and Windows Vista operating systems running on a server with 2.5 GHz CPU and 2 GB RAM. Three Gen_B machines are used as sandbox.
- Col_A: A client side low interaction honeypot aimed at mimicking the behaviour of a web browser in order to detect and emulate malicious contents [18].
- Col_B: A server side low interaction honeypot that captures attack payloads and malwares [19].

2) NetFlow Generation

Gen_A executes each malware samples five or more times in CTMS environment. After executing and generating trace of flow, we restore virtual machines to clean state and then Gen_A prepares itself for new malware execution. Same process is repeated for Gen_B. Nevertheless, this process is more complicated than the cleaning state in Gen_A because of requirement of operating system reinstallation on the server.

Since some malware families are virtual machine (VM) aware that can recognize the virtualized environment and alters behaviour accordingly, we alter the settings by changing original manufacturer information of the VM with a pseudo one, removing or changing registry keys containing VM keyword, changing MAC address identified as VM Ethernet cards, changing disk settings such as serial number, firmware number etc. and killing particular service threads which indicate VM existence to delude the VM-aware malwares.

3) Classification Unit (Virus.Mu?)

A custom malware classification module called *Virus.Mu?*, (meaning *"Is it virus?"* in Turkish) similar to VirusTotal which is multi engine online virus scanner [20], is implemented by using actual versions of various antivirus products from different vendors on isolated VMs [11]. After appending malware samples and suspicious documents gathered by honeypots to a queue, each antivirus product scans the queue. If a suspicious file in the queue is identified as malicious, it is tagged based on common keyword in virus naming scheme of corresponding vendor. Then, different naming scheme correlated with the same malware family are used to get exact family name. For instance, a *Waledac* malware sample is tagged as *Email-Worm.Win32.Iksmas.gen*, *Mal/WaledPak-A* and *Trojan:Win32/Waledac.gen!A* by Kaspersky, Sophos and Microsoft, respectively. In addition to *Virus.Mu?*, we use Suricata-IDS with the Emerging Threats Pro Ruleset (ETPro) which delivers network based malware threat detection rule set [21], [22]. This rule set contains newly detected malwares' signatures, thus, we can validate *Virus.Mu?* and IDS alerts in our development network.

C. Features

1) Trace Extraction

As a preliminary phase for some statistical and computational features we extract traces from NetFlow data. Traces, representing consequent flows in terms of chronological order are the most commonly used concepts in bot detection algorithms. Since we apply trace extraction unit both training and investigation data, we have to whitelist common Internet services such as

Microsoft, Google, Akamai, update services, file-sharing services such as SharePoint, DropBox etc. Another filtering process is applied by comparing the destination IP with most known C&C servers. Flows are eliminated and our trained models are more likely to capture only bot traffic. As a result, the filtering process in trace extraction has a mediate impact on malware detection results.

2) Feature Extraction

Later, we utilize statistical features such as average time interval, average connection duration, average number of source bytes per flow, average number of destination bytes per flow, communication regularity and outgoing data accumulation regularity. Each statistical feature is computed on subsequent flow pairs. Features are briefly as follow: ([8] gives detailed explanations for first five ones):

- **Average Time Interval:** It reflects the average time interval between two subsequent flows in the trace. This measure detects the periodic characteristics occurred in C&C connections. Most of the malwares intent to use a constant time interval or a random interval time within a constant value between two connection periods.
- **Average Duration of Connections:** Since a malware runs same process in each connection, it is expected that the duration of different connections of a malware might be similar and different than human-computer interaction. Therefore, computing this statistic helps to distinguish a malware connection from normal ones.
- **Average Number of Source and Destination Bytes per Flow:** As the same motivation with the previous feature, it is expected that a specific C&C server will send same commands to a target machine. Thus, the average number of bytes has a characteristic structure in a C&C trace. Similar consideration will be in charge in destinations bytes. A target machine will give a fixed response to a particular C&C server.
- **Communication Regularity:** We apply Fast Fourier Transform to the binary sampled C&C communication to detect communication regularities. While doing this we sample our connection start time as 1 and 1/4th of the smallest connection interval slops as 0. Afterwards, we compute the Power Spectral Density (PSD) of the Fast Fourier Transformation over our sampled trace and extract the most significant frequency. This helps us to detect even randomly varied C&C connections within a certain range to an extent.
- **Data Accumulation:** We apply a new feature in addition to [8] for detecting malwares with randomly changed duration within two subsequent flows in a trace. This measure is calculated as average value of the each ratio of data size difference between two subsequent flows to difference of start times of them. Since the connection times of such flows may be extended because of communication problems with C&C, the accumulated data amount, which is produced by victim and stolen by an attacker, in the following connection in such a case will grow up, especially in malware with keylogger payload. Thus, characterizing the accumulated data amount per second between two connections might exhibit similarities

D. Model Creation and Detection Unit

The basic assumption behind the usage of a machine learning algorithm in detection module is that malwares leave proprietary patterns of traffic or behaviour, which could be tracked over traces, within the target machine. Our desired outcome is to raise an alert if the NetFlow data gathered from investigated traffic includes a known pattern belongs well-known and actual malwares. Thus, we use a supervised machine learning algorithm based on several statistical features, explained in previous section, instead of one of the unsupervised algorithms which do not need any training data and are mostly used to cluster similar data within isolated groups.

A supervised machine learning algorithm in malware activity detection has to address several concerns such as generality, robustness on evasion techniques, stealthiness and timely detection [23]. Firstly, the generality of the detection module represents the ability of covering a wide spectrum of malware types in the training data. Secondly, the robustness refers to the ability of recognizing different and new types of smuggling methods. Thirdly, stealthiness requires detecting a malware attack without revealing ourselves to the attacker. Moreover, the detection algorithm has to operate in on-line fashion with a reasonable respond time and high detection accuracies. Since our system upgrades itself with daily collected data through a number of honeypots, classification models cover recent malware types and are getting robust on their evasion techniques. In our method, since we analyse the trace data in passive fashion without establishing an interaction with attacker, it is not possible to draw information about detection process to the attacker. Finally, the investigation data are gathered as NetFlow data and it is a trivial operation in terms of time consumption to extract traces and statistical features. Thus, the detection system in this work is suitable for on-line operation.

In the last decade several supervised machine learning algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision tree classifiers, Bayesian classifiers and random forest algorithms have been proposed in botnet detection and C&C server identification [6], [24]. On the other hand, similar algorithms like these ones could be customized for botnet detection with specific feature space as applied in [8]. In this case, such techniques require a clustering phase for creating classification models in training while they need a weighted scoring methodology to identify the cluster of the investigation data. In what follows next, we introduce our modelling and matching algorithms based on six statistical features. Detailed explanation about our detection algorithm is given in [8].

1) Model Creation

In common supervised machine learning algorithms, the size and attributes of the classes in the classifier model should be introduced before triggering the training process. For instance, labels represent the malware families should be included the detection model by associating them with the feature vectors created via the traces belong that malware family in the first place before training the model in SVM algorithm, and like so many others. However, this limits to introduce the actual and new malware families to the classifier model while dynamically updating that with daily incoming data from honeypots in our situation. Therefore, we use an un-supervised machine learning approach, CLUES (CLUstEring based on local Shrinking) algorithm [25], to create detection models for each malware families. We first calculate our six statistical features separately for each trace of the training data. Then, the trace-features are

clustered by using CLUES algorithm which allows dynamic sized clustering without selecting number of clusters. A cluster which includes large number of trace-features for a particular malware family, identified through malware classification unit, represents the one of the expected values for this feature for this particular malware family. A weight is associated with each cluster in the degree of representing that malware family. Eventually, six sets of weighted clusters are created for each malware families. The average value of all of each cluster weights for a family is assigned as *cluster quality*.

2) Detection

Each features of a trace belongs to investigation data is compared individually with each of the clusters of each of the detection models which represent malware families. For instance, the first feature of a trace (*T*) is in the scope of values belong to one of the clusters in a model (*M*), then, it counts a hit. Then, the weight associated with this cluster is added to that feature's *total hit score*. If another cluster for this feature in model-M raises a *hit*, its weight is added in the same way. Then, if this feature's *total hit score* exceeds the same feature's *total hit threshold*, it counts that this feature belongs to model-*M*. Same calculations are conducted for other features of trace-*T*. Eventually, if majority of the features of trace-*T* belongs to model-*M*, an alert raises about detection of infected machine by the malware family which has the classification model as model-*M*.

E. Distributed Processing

Hadoop Distributed File System (HDFS), is a purposefully developed system for handling large files through write-once and read many data-access patterns. It has two components; name node, which is responsible for metadata of file system and management and data node that is for block storage and retrieval of data. Hadoop provides MapReduce software framework. MapReduce programming model utilizes input and output (key, value) pairs to manage processing data on different nodes.

BFH processes exclusive traces and does not require correlating any of two, thus, calculating statistical feature is easy to be implemented in distributed way. Since, its modeling and matching unit focuses on traces between $IP_{internal}$ and $IP_{external}$ entities, MapReduce programming logic is a perfect match for our system as they can be used for key values.

MapReduce methodology provides grouping and partitioning utilities to manage to group flows based on multiple entities at the same time. BFH manipulates it to store the flows that have same ($IP_{internal}$, $IP_{external}$) entities, meaning once flow start times are sorted, it extracts traces automatically. Main overhead for Hadoop is moving data over network, reading and writing to disk, yet, this type of data storing, maximizes the possibility of keeping traces in one data node, minimize the possibility of moving data over network. Performance evaluation of our system is a complete work for another paper; thus, it is not discussed in this paper. However, [26] provides ground truth on how Hadoop can outperform for enough large scale networks.

5. EXPERIMENTATION

BFH is deployed in a part of large-scale enterprise network in Turkey which has about 15000 hosts as an extension of CTMS, actively running in a production environment. NetFlow data over this network is directly extracted from Cisco devices and stored on Hadoop clusters after dumping them to text file.

For evaluation purposes, we evaluate BFH on a part of system, a network with ~8200 hosts and in daily measurement ~6300 concurrently active, which are more vulnerable to be infected as they provide services over internet (Table I) for three months. This network will be referred as "experiment network".

TABLE I: EXPERIMENT NETWORK INFORMATION

Total Number of Flows	322920000
NetFlow Size (GB)	41.4
Internal Host Count	~8200
Concurrently Active	~6300
Start Date	01-07-2013
End Date	30-09-2013
Length (Days)	92

A. Training Dataset

As SysSec Report [11] details the information on malwares caught by CTMS, our system is able to perform on a large amount of malware samples, however, to provide better estimation on performance, as Table II shows, six different malware families are discussed over time period of 15 days. Classified Malwares, caught via 97 honeypots are used to train our system. On each 15 days, traces and models are updated via accumulated malwares till that date. Table II shows sample and trace details of families over time.

1) Malwares

Carberp - Sophisticated, modular and persistent malware utilizing advanced obfuscation techniques to evade detection, removal and the ability to disable antivirus.

Hesperbot - A Trojan horse that opens a back door on the compromised computer and may steal information.

Tinba - Tiny Banking Trojan that steals information from the compromised computer.

Ramnit - A multi-component malware family which infects Windows executables as well as HTML files.

Gamarue - A malware that can download files and steal information about compromised computer.

Cridex - A malware that may be delivered via spammed malware. It captures online banking credentials entered via web browsers, downloads and executes files, and searches and uploads local files.

Aforemntioned malwares are most observed malwares within Turkish Network, thus, they have been selected in experiments.

B. Experiment

Experiment is conducted on experiment network after whitelisting for some external services that might exhibit regular behavior and increase FP rate, such as; Microsoft, Google, Akamai, update services, file-sharing services; SharePoint, DropBox etc. A BFH generated alert is analyzed by using full traffic capture, if symptoms are explicitly matched than it is signed as true alert. Meanwhile both network-based and host-based IDS/IPS alerts are also used for double-check. If there is no explicit symptom from neither full packet capture nor IDS/IPS solutions then blacklisting services are used to determine [27-31]. In case, none of these controls provide any infection implication, it is signed as False Positive, while this might not be completely true.

TABLE II: MALWARE FAMILY INFORMATION CAUGHT BY HONEYPOTS

	Start Date - End Date	01 Jul - 15 Jul	01 Jul - 31 Jul	01 Jul - 15 Aug	01 Jul - 31 Aug	01 Jul - 15 Sep	01 Jul - 30 Sep
		Number of Samples / Traces					
Malware Family	**Carberp**	3 / 8	4 / 9	18 / 18	32 / 24	42 / 31	52 / 35
	Hesperbot	4 / 4	6 / 6	9 / 10	11 / 13	14 / 21	19 / 27
	Tinba	20 / 24	32 / 30	34 / 38	38 / 45	46 / 52	49 / 62
	Ramnit	11 / 21	14 / 25	18 / 29	25 / 36	33 / 46	37 / 55
	Gamarue	25 / 24	28 / 29	31 / 35	34 / 39	38 / 43	43 / 51
	Cridex	12 / 20	16 / 25	21 / 33	25 / 39	32 / 46	36 / 50

1) Test Dataset

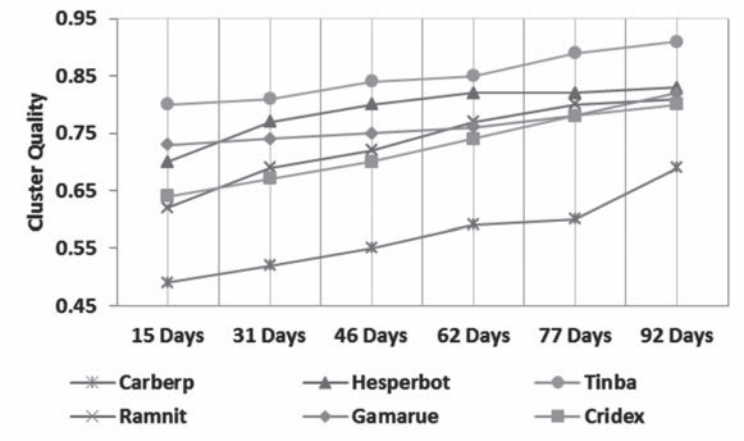
Bot Detection systems, mostly focus on off-line dataset analysis and one dataset of a large-scale enterprise network. However, in real scenarios, actively running bot detection systems are most likely to be analyzing weekly or monthly changing dataset. In our active system, created models via accumulated malwares are used to detect bots on NetFlow traffic that belongs to last four months. Since our NetFlow data changes over time, we focus on diverse dataset, which is NetFlow of each month. Consequently, our test dataset consists of three different NetFlow, stored in months: July 2013, August 2013, and September 2013.

Besides, complete traffic captures of this particular network are stored for 30 days to verify generated alerts, but, for storage limitations, it is deleted monthly. Therefore, in our experimental setup, detection rates and infected host are analyzed by using accumulated malware samples and traces after each 15 days to provide better understanding for contribution of Honeypots. More precisely, accumulated traces are used to train the system then created models are applied on subjected month's NetFlow data.

6. DISCUSSION

Figure 2 summarizes training dataset characterization for each family over time. This graphic illustrates, when number of samples increases, cluster quality for each family rises. This graphic implies that BFH, wisely, manipulates honeypots to increase cluster quality.

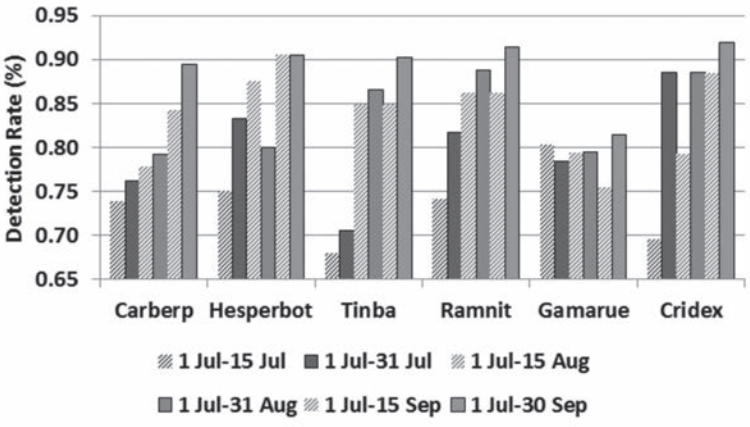
FIGURE 2: CLUSTER QUALITY OVER TIME



Interestingly, cluster quality of Carberp malware family is less than other malware families; main reason for this is that Carberp produces different number of traces from one sample. For example, in the first half of July, three samples are captured and eight traces are generated out of them. Beyond that, two factors can be considered as cause for this, one is that classification unit identifies some of the malwares as Carberp, yet, it belongs to a different family. Second, Carberp might have different variants, exhibiting diverse network characteristics.

Figure 3 is the BFH detection results. In this graphic, detection rate of each experiment on same dataset is highlighted with same color. First and foremost, Figure 3 reveals that BFH is able to detect bot-infected machines in worst case 68%, in which NetFlow data is limited to two weeks and number of samples of this particular family is less than a half of number of samples in September. Although, there is not false negative evaluation opportunity, for a detection system, having a few false-positives among a significant number of alerts (Figure 4) is an important indication of success, where BotFinder has detection rate from 49% to 87%, except Banbra family.

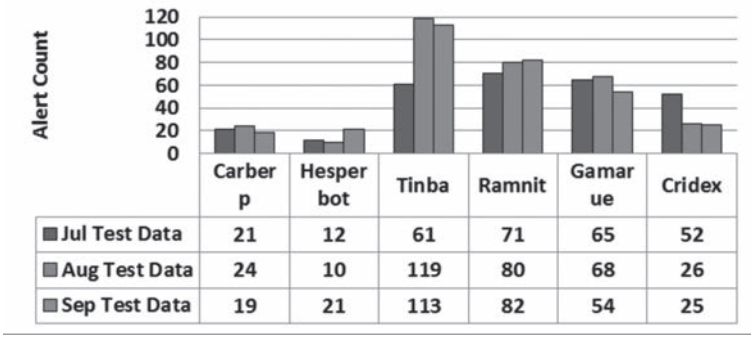
FIGURE 3: DETECTION RATE (EACH COLOR INDICATES RESULT OF EXPERIMENT ON SUBJECTED MONTH'S NETFLOW)



Secondly, detection rate for same dataset (tone-in-tone dyed) indicates, the more traces used in training, the more accurate detection, except Hesperbot in August. This family generates only 8 alerts with 1 FP on first half of August whereas it generates 10 alerts with 2 FPs. In real, it detects more bot-infected machines. Consequently, it highlights the vigilance of BFH on integrating Honeypots to bot-detection system.

Furthermore, when we compare detection rates for different datasets, in Figure 3, dashed columns of each family should be considered so as to infer that detection rates increase in monthly by improvement of samples and traces with a few exceptions, discussed on later section. Indeed, detection rate is expected to increase between second half of a month and first half of a month because system is trained with higher number of traces, yet datasets are different but hosts within the network same. However, Ramnit and Gamarue families have statistics that contradict to it. For instance; BFH has higher detection rate on end of August than beginning of September. Since experiment network involves around 8000 hosts with approximately 6300 concurrently active hosts, and active hosts are most likely to be different within different months while matching unit runs.

FIGURE 4: INFECTION ALERTS ON EACH DATASET OVER TIME



	Carberp	Hesperbot	Tinba	Ramnit	Gamarue	Cridex
Jul Test Data	21	12	61	71	65	52
Aug Test Data	24	10	119	80	68	26
Sep Test Data	19	21	113	82	54	25

7. CONCLUSION

This paper presented BFH, a live BotFinder-based automated bot-infected system through Honeypots. BFH does not require any host-based information, deep-packet inspection or any support from other network-based security deployments such as IDS/IPS. Instead, it relies on NetFlow data, uses behavioral and training-based approach so as to detect encrypted communications and avoid storage overhead, thus, it provides solution for large-scale. BFH is vigilant system, since training module of BFH is fed by samples caught via sophisticated honeypot system. BFH is deployed to a large-scale enterprise network in Turkey on Hadoop that provides scalability. Our experiment on subjected network shows that BFH is able to detect centralized bot-infected machines with high-accuracy; indeed, similar approach can be improved to detect P2P bots as future work.

8. ACKNOWLEDGEMENT

This work was an extension of European Union SysSec project and it is funded by The Scientific and Technological Research Council of Turkey (TUBITAK).

REFERENCES:

[1] Cooke, Evan, Farnam Jahanian, and Danny McPherson. "The zombie roundup: Understanding, detecting, and disrupting botnets." Proceedings of the USENIX SRUTI Workshop. Vol. 39. 2005.
[2] Freiling, Felix C., Thorsten Holz, and Georg Wicherski. Botnet tracking: Exploring a root-cause methodology to prevent distributed denial-of-service attacks. Springer Berlin Heidelberg, 2005.
[3] Bayer, Ulrich, et al. "Scalable, Behavior-Based Malware Clustering." NDSS. Vol. 9. 2009.
[4] Bayer, Ulrich, Christopher Kruegel, and Engin Kirda. "Anubis: Analyzing Unknown Binaries." (2009).
[5] Coskun, Baris, Sven Dietrich, and Nasir Memon. "Friends of an enemy: identifying local members of peer-to-peer botnets using mutual contacts." Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010.
[6] Bilge, Leyla, et al. "Disclosure: detecting botnet command and control servers through large-scale NetFlow analysis." Proceedings of the 28th Annual Computer Security Applications Conference. ACM, 2012.
[7] Franćois, Jérôme, Shaonan Wang, and Thomas Engel. "BotTrack: tracking botnets using NetFlow and PageRank." NETWORKING 2011. Springer Berlin Heidelberg, 2011. 1-14.
[8] Tegeler, Florian, et al. "BotFinder: finding bots in network traffic without deep packet inspection." Proceedings of the 8th international conference on Emerging networking experiments and technologies. ACM, 2012.
[9] Claise, Benoit. "Cisco systems NetFlow services export version 9." (2004).
[10] Franklin, Jason, et al. "An inquiry into the nature and causes of the wealth of internet miscreants." ACM conference on Computer and communications security. 2007.
[11] (SysSec) A European Network of Excellence in Managing Threats and Vulnerabilities in the Future Internet: Europe for the World.
[12] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
[13] Gu, Guofei, et al. "BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection." USENIX Security Symposium. 2008.
[14] Nagaraja, Shishir, et al. "BotGrep: Finding P2P Bots with Structured Graph Analysis." USENIX Security Symposium. 2010.
[15] Livadas, Carl, et al. "Usilng machine learning technliques to identify botnet traffic." Local Computer Networks, Proceedings 2006 31st IEEE Conference on. IEEE, 2006.
[16] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).
[17] Giroire, Frederic, et al. "Exploiting temporal persistence to detect covert botnet channels." Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2009.

[18] Thug: The Honeynet Project [Online]. Available: http://www.honeynet.org/taxonomy/term/218
[19] Provos, Niels, and Thorsten Holz. Virtual honeypots: from botnet tracking to intrusion detection. Pearson Education, 2007.
[20] VirusTotal [Online]. Available: https://www.virustotal.com/
[21] Jonkman, M. "Suricata IDS available for download." Message posted to marc. info (2009).
[22] "Emerging Threats," [Online]. Available: http://www.emergingthreats.net/.
[23] Stevanovic, Matija, and Jens Myrup Pedersen. "Machine learning for identifying botnet network traffic."
[24] Nogueira, António, Paulo Salvador, and Fábio Blessa. "A botnet detection system based on neural networks." Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE, 2010.
[25] Wang, Xiaogang, Weiliang Qiu, and Ruben H. Zamar. "CLUES: A non-parametric clustering method based on local shrinking." Computational Statistics & Data Analysis 52.1 (2007): 286-298.
[26] Lee, Youngseok, Wonchul Kang, and Hyeongu Son. "An internet traffic analysis method with mapreduce." Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP. IEEE, 2010.
[27] [Online]. Available: http://www.mtc.sri.com/live_data/attackers/.
[28] [Online]. Available: http://isc.sans.edu/sources html.
[29] [Online]. Available: http://www.projecthoneypot.org/list_of_ips.php.
[30] [Online]. Available: http://mirror1.malwaredomains.com/files/BOOT.
[31] [Online]. Available: http://www.malwaredomainlist.com/hostslist/hosts.txt.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Botnet over Tor: The Illusion of Hiding

Matteo Casenove
VrijeUniversiteit
Amsterdam, The Netherlands
m.casenove@gmail.com

Armando Miraglia
VrijeUniversiteit
Amsterdam, The Netherlands
a miraglia@student.vu.nl

Abstract: Botmasters have lately focused their attention to the Tor network to provide the botnet command-and-control (C&C) servers with anonymity. The C&C constitutes the crucial part of the botnet infrastructure, and hence needs to be protected. Even though Tor provides such an anonymity service, it also exposes the botnet activity due to recognizable patterns. On the one hand, the bot using Tor is detectable due to the characteristic network traffic, and the ports used. Moreover, the malware needs to download the Tor client at infection time. The act of downloading the software is itself peculiar and detectable. On the other hand, centralized C&C servers attract a lot of communication from all the bots. This behaviour exposes the botnet and the anomaly can be easily identified in the network.
This paper analyses how the Tor network is currently used by botmasters to guarantee C&C anonymity. Furthermore, we address the problems that still afflict Tor-based botnets. Finally, we show that the use of Tor does not, in fact, fully guarantee the anonymity features required by botnets that are still detectable and susceptible to attacks.

Keywords: *Botnet, Tor, Command-and-Control, Malware, Anonymity, Resilience.*

1. INTRODUCTION

Nowadays, one of the main threats that the Internet users face are *botnets*. Botnets are employed for many kind of malicious activities; examples are DDoS, personal data theft, spam, bitcoin mining, and cyber-espionage [19][9]. In the last ten years, the main antivirus vendors have reported a constant growth of botnets in the wild [1][2].

Traditionally, botnets are centralised overlay networks where the Command-and-Control (C&C) servers act as single point of control. Centralised botnets are easy to manage and maintain due to their centralised structure. A botmaster has a clear overview of the overlay network and she manages the bots, which, in turn, connect to the C&C servers to be reachable. Nevertheless, this architecture has an important drawback: the C&C servers are exposed and represent a single point of failure. Hence, by taking down the C&C servers, the whole botnet is defeated. In order to overcome this problem, botmasters have moved to more resilient unstructured P2P

networks for their bots. In this manner, P2P botnets remove the single point of failure, building a completely distributed network. In this network structure, the bots exchange commands among themselves. Ultimately, this new architecture achieves resiliency, making the disruption of the botnet significantly harder.

Alternatively, some botmasters have opted to keep the centralized structure of the botnet but, at the same time using improved techniques to decrease C&C servers detectability. In fact, the simplicity of the protocols and of the network organization are desirable properties. One of the most interesting techniques to achieve this goal is the use of the *Tor network*. By means of the Tor network, the botmaster can anonymously locate their C&C servers, which, in turn, are contacted by the bots which join the botnet. Tor is a network that provides anonymity. It creates an encrypted routing system to avoid traffic analysis and allows to publish services without revealing their locations. To do so, Tor provides the so-called *hidden services*. Hidden services [21] are characterised by services like web servers, shell providing services and others which are accessible only via Tor. In this manner, the client using the service does not require the actual address, and hence the actual location of the service, guaranteeing the service anonymity. In turn, botmasters can configure the C&C servers as hidden services. In this way, it is not possible to detect the C&C locations and, consequently, take down the botnet. Additionally, even though this technique is yet not being actively deployed for P2P botnets, it is still fully applicable and could, in fact, provide further resilience to take down attempts. Unfortunately, while the botmaster tries to hide her network, she also exposes it due to peculiar properties of the Tor network.

In this paper, we describe the weaknesses, overlooked by botmasters, which derive from the deployment of botnets over Tor when aiming for stealthiness. We present the use of Tor by the botnets providing the following contributions. Firstly, (a) we argue that botnets over Tor are still subject to the very same attacks used to defeat botnets that do not use Tor. Afterwards, (b) we argue that in some situations moving to Tor is counterproductive for the botmaster since this creates an anomaly in the normal network flow, attracting the observers attention. Finally, (c) we discuss the vulnerabilities of the Tor network that can expose the botnet, and mine the anonymity offered.

The remainder of the paper is organised as follow. In Section 2, we provide an overview of the background knowledge required. Section 3 describes the use of Tor in real-life botnets. After presenting the current state of these botnets, in Section 4 we analyse the vulnerabilities of this approach. Moreover, in Section 5 we discuss the related work and finally Section 6 summarises our work and provides a conclusion of the paper.

2. BACKGROUND

In order to better understand the problems that arise from building botnets over Tor, we first clearly describe the way in which botnets are structured and what type of features they use to achieve resiliency. Secondly, we present the Tor infrastructure and all its actors involved in the management of the system. These are crucial concepts, which are required before analysing that the combination of these two strong systems does not necessarily produce a stronger one.

A. Botnets

A Botnet is an overlay network of compromised machines called bots that are controlled by an attacker (botmaster). In order to connect the bots together, the machines are infected with a malware. There are different ways to infect a machine such as *0-day* exploits and spam. However, the most effective method used nowadays is *drive-by-download*. Using this last vector of infection, the attacker compromises a site which, in turn, is visited by a user. When visiting the page, the user will unknowingly download and install the malware, hence becoming infected. The botmaster uses the botnet to control the bots. By issuing commands, she can instruct the bots to perform malicious activities, namely *DDoS*, spam campaigns, credential theft, cyber-espionage, bitcoin mining, and others.

Botnets can be distinguished based on (a) the kind of malicious activities they perform, (b) the protocol they use, and (c) their architecture. Traditionally, the botnets have a very basic structure where every bot is connected to a central server controlled by the botmaster. In order to simplify the control of the bots, botmasters have deployed botnets with IRC-based (Internet Relay Chat) communication. The central server is also called C&C. This structure makes the botnet very easy to control but also easy to attack. The C&C is the only central node that connects all the bots. Hence, by taking down the server, the whole botnet is disrupted. For the last years, botmasters have continually updated the protocol and the architecture of their botnets. Botmasters have replaced a single C&C with multiple C&C servers or have used the *Fast-Flux* technique in order to achieve a better resiliency. They have also implemented *Domain Generation Algorithms* (DGA), which allow the malware to generate a random domain name at runtime. The random generated domain locates the C&C server[1]. Despite all the efforts from the botmasters, the centralised structure makes the botnets weak and easy to take over. In fact, the central C&C remains the single point of failure.

Since the structure is the weakness of such botnets, botmasters have moved to a more resilient structure: the P2P network [18]. The P2P architecture replaces the central C&C with a completely distributed network of bots. Bots exchange information between each other, transmitting commands and overlay management information using custom protocols. They also use common protocols, such as HTTP, DNS, and others, in order to be as stealthy as possible for operations like downloading new versions of the malware. Of course, this also makes the botnet more difficult to manage and to monitor. The P2P structure makes the botnets more resilient but not invulnerable against attacks or even disruption [19]. The P2P botnet can be identified using a method called *crawling*. With crawling it is possible to enumerate all or almost all the bots in the botnet. Furthermore, disruption can be achieved using *sinkholing*. Sinkholing is a technique that allows disruption by injecting crafted information in the list of peers of every bot. This modifies the structure of the network, turning it into a centralised network. The injected node can be controlled by the defender or can be inexistent, making all the bots point to a black hole.

Crawling and sinkholing are not always successful, since they require a deep analysis of the botnet protocol. Additionally, botmasters improve their botnets over time to prevent these attacks. Currently there is a competition going on between botmasters and researchers where botmasters try to make the botnets resilient and powerful while the researchers try to take them

1 Contextually, the botmasterhas to register the samerandom domain name and link with the C&C server.

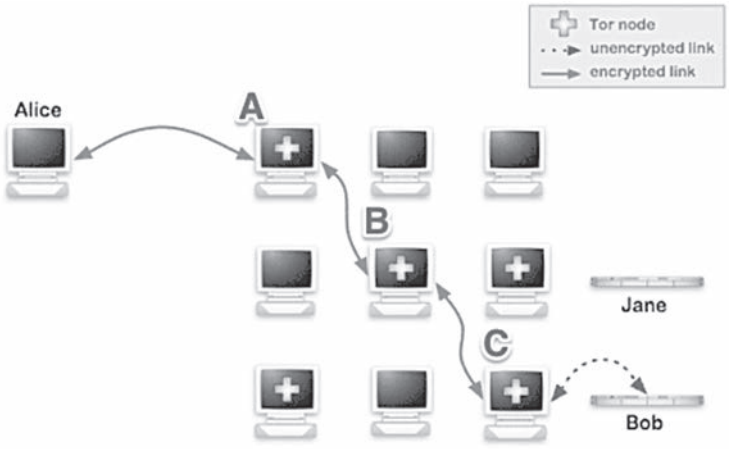
down. Lately, a new trend is arising: botmasters have started using Tor to hide the C&C servers [4] [5] [8].

B. Tor: Third-generation Oniuon Router

In a world where governments make use of monitoring and censorship, Tor [20] allows users to evade these invasive governments activities by providing anonymity. It is a network of volunteers, which provide thousands of relays used to route communication in the Tor network. The onion name comes from the multilayer encryption used by the relays in order to provide confidentiality. In fact, the encryption protocol is designed to avoid giving the relays access to the data they are routing.

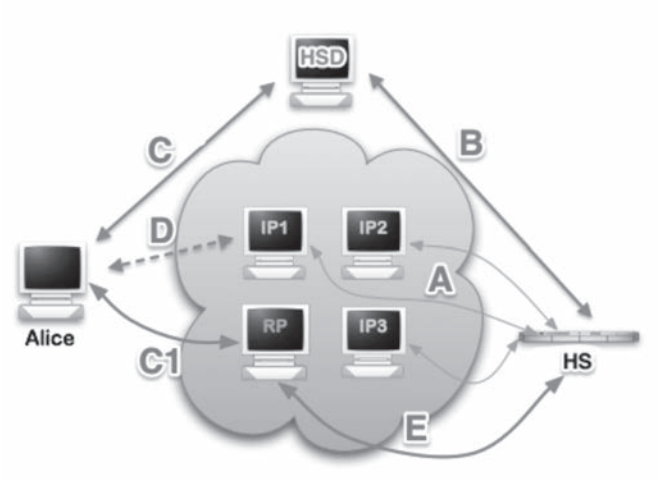
Consider the scenario where Alice wants to communicate with Bob in an anonymous way as in Fig. 1. At first, Alice randomly selects three different relays, which represent the entry point (A in Figure 1), the exit point (C in Figure 1) and the middle hop (B in Figure 1), creating the so-called *virtual circuit* from the source to the destination. The client negotiates with each relay in the circuit a separate set of encryption keys in order to enhance the privacy on the circuit. In such a scenario the flow of the communication evolves as follows. Alice sends the message to the entry point using an encrypted channel, then once inside the Tor network, the message is sent from relay to relay until it reaches the exit point where the communication is sent in clear to the destination.

FIGURE 1: TOR PROTOCOL.



Because each relay sees no more than one hop in the circuit, there is no way for a malicious relay or a monitoring system to trace the communication from the source to the destination. Since it is the exit point that sends the message to the destination, Bob does not know who the real source is but he only sees the exit point.

FIGURE 2: TOR HIDDEN SERVICE PROTOCOL.



Tor also allows to publish services inside the network anonymously (Figure 2). These services are called hidden services (HSs). Mainly, a service needs to publish its presence in the network in order to be reachable. The HS selects a set of relays asking them to be its introduction points (A in Figure 2). It creates a hidden service descriptor containing its public key and the address of its introduction points, then it inserts the descriptor in a DHT using an address like *XYZ.onion* as key (B in Figure 2). The *.onion* address can be used by the client to contact the HS (C in Figure 2). The DHT is implemented in Tor using the Hidden Service Directories (HSDs). With the *.onion* address, the client downloads the descriptor and it creates a new virtual circuit to a random relay asking it to act as rendezvous point (C1 in Figure 2). The client then uses the introduction points to inform the hidden service about the rendezvous point (D in Figure 2). Finally the HS creates a virtual circuit to the rendezvous point and starts the communication with the client (E in Figure 2).

This standard design of Tor is vulnerable to the traffic confirmation attack [3] that permits an attacker to confirm that two parties are communicating by observing patterns. If the attacker controls the edges of a circuit, it can expose the hidden service. In order to avoid such a scenario, Tor introduces the concept of *guard nodes* [21]. The entry guards are special relays, flagged as "good enough" to be guard, selected at random to become the new entry point in the circuit.

Unfortunately, even using the entry guards, Tor still has some vulnerabilities [17] [11]. In fact, it is not feasible to build such a big and complex system, which is completely secure. Therefore, Tor is not claimed to be foolproof or to provide absolute anonymity but it is instead guaranteed to provide an anonymity service, which is "good enough".

3. BOTNET OVER TOR

Since 2010, the use of Tor to hide botnets infrastructures has been discussed. In particular, the famous presentation by Dannis Brown at *DefCon18* [4] has shown, for the fist time, a possible implementation of a C&C channel over Tor to provide C&C server anonymity. Even after the presentation, we have not seen real application of this idea until Guarnieri in [5] detected and analysed the first Tor-based botnet. Announced on Reddit, the botmaster published the following message: "Everything operating tru TOR hidden service so no feds will take my servers down.". The botnet is a modified version of Zeus with DDoS, bitcoin mining and credential theft capabilities. The malware contains the Zeus bot, the tor client, the GMinerbitcoin mining tool, and few libraries for GPU hash cracking. All the bots run inside hidden services and all the C&C communication happens inside the Tor network. Avoiding the use of the exit nodes, the botmaster tries to reduce the botnet traceability. It uses an old-style IRC protocol to communicate with the bots and to issue commands. To be ethically correct with the Tor philosophy, the botmaster also makes the bots act like relays enhancing the Tor network while exploiting it.

Late summer 2013, a post on the Tor mailing list [6] raised the attention on a huge increment of the network usage and the amount of users in a really short amount of time. At the beginning no one could explain such an atypical situation [7] but then researchers [8] discovered that it was caused by a very big botnet that suddenly switched to Tor. The botnet uses the HTTP protocol over Tor with a centralised structure. It uses a preconfigured old version of Tor to connect to the network.

Unfortunately, this caused problems to the Tor network due to the significant increase of Tor communication going through the relays. In fact, the computational overhead caused by the expensive encryption operations has reduced the responsiveness of the system. This made many people unhappy and raised a lot of discussion especially by the Tor users [7].

These two examples show that botnets over Tor are no longer a forum discussion but have become a reality. The main reason that motivates botmasters to move to Tor is to find a new environment to achieve stealthiness and untraceability. The Tor hidden services provide anonymous C&C servers, which are more difficult to take down. Even attacking the server with a DDoS attack is unfeasible because the whole Tor network would be under attack. However, overconfidence can be dangerous.

4. THE FAILURE OF STEALTHINESS

Seeking for resiliency, botmasters are continuously evolving their botnets in order to resist against attacks to their network. Nowadays, P2P botnets show an improved resilience [9] but still Tor represents an appealing environment for botnets. In fact, to reach such resiliency the P2P botnets apply quite sophisticated techniques, while Tor provides anonymity and resiliency also for centralised infrastructures, which require less effort.

Unfortunately, this is not completely true. We argue that the botnets over Tor are interesting solutions but not as perfect as expected. Tor-based botnets do not represent the ultimate stage of resilience and are not less affected by the same vulnerabilities. For example, P2P botnets over Tor are not yet present in real life but within the bounds of possibility and it would be interesting to analyse the impact of Tor on their resilience. Every bot runs inside Tor as an HS creating an overlay network on top of the Tor network. The bots are identified by .onion addresses and they communicate using the classic custom protocols but this time tunnelled in Tor.

Surely, these botnets are not less subject to the very same kind of attacks applied to standard botnets. Even though bots are running as HSs and so their identities cannot be revealed, the crawling attack is still applicable. It would require only to use *.onion* addresses instead of normal IPs. Crawling aims to enumerate the bots in the network and, using Tor, we can enumerate the *.onion* addresses, which are part of the botnet. In this case, the use of Tor addresses gives the crawling an important advantage. While in standard networks we risk to overestimate the size of the botnet due to dynamically assigned IPS, the *.onion* addresses are uniquely assigned to each hidden service. The addresses are linked to the keys of the hidden service and do not change over time[2]. Hence, crawling becomes much more accurate when using such addresses. As a result, the technique can almost exact estimate the botnet network size.

The sinkholing attack injects fake nodes in the peerlist of the Tor-based bots as well as it does for the normal bots. While in the standard network, a sinkhole would be a standard IP address, in the Tor network the injected address would be a hidden service *.onion* address. Tor itself does not add any extra security feature against this attack, since it is a result of flaws of the botnet protocol and not of the network used. Furthermore, in Tor we face the ethical limitation based on which we cannot inform, attack or disconnect the bots in a P2P botnet. Geographically locating or identifying the IP of the bots is not really useful in a P2P botnet, since we cannot apply any direct action on the bot itself. For this same reason, crawling inside the Tor network has the same final effect of the normal crawling.

In the centralised botnets, the use of Tor can give an immediate solution for the single point of failure (e.g. if we cannot locate the C&C server we cannot take it down). Even attacking the server can be tricky. For example, by DDoS-ing the server we attack the whole Tor network [10]. Unfortunately, Tor has vulnerabilities that can result in the deanonymisation[3] of a service. Biryukov et al. [11] describe the possibility to determine the IP address of a hidden server exploiting the use of the guard nodes. Moreover, the Tor network is vulnerable to the traffic correlation attack where an attacker controls one or significantly many relays [13]. This is not an unrealistic scenario especially considering the recent data gate scandal [14] where the NSA was monitoring almost every communication channel in countries like USA. When we have a P2P botnet over Tor, it is very difficult to deanonymise every bot in the network but when we have a single or even few C&C servers it becomes feasible. This means that even using Tor, a centralised botnet has the same source of vulnerability, namely the single point of failure.

2 The .onion address does not change over the time unless the HS explicitly reboots itself and intentionally recreates the keys. In this way, a HS appears as a new service and it has to register itself inside the Tor network. This results in a really expensive operation and highly unlikely to be performed by a bot.

3 Deanonymise a service in the Tor network means that the real IP address of the service is revealed and so Tor does not provide anonymity anymore.

Lastly, when a botnet takes part in the Tor network, it raises a lot of attention because it creates instability in an almost stable network as Tor is characterised by a slow variation in the number of nodes. A botnet is often comprised by million of bots and when so many nodes join the Tor network in a short time, it signals that something wrong is going on. Mimoso [15] points clearly that a botnet undetected for years suddenly decided to hide the C&C on Tor and at the same time exposed itself making its presence more obvious.

A botnet using Tor leaves traces even from a client point-of-view. The way Tor is used nowadays by the malware has nothing to do with stealthiness. In fact, malware currently runs the Tor client as an external process. If the client was not previously installed, the exposure of the malware activity would be trivial. In fact, by verifying the list of running processes, the malware would be detected by identifying the Tor client process. Even though this is a clear symptom of infection, bot writers have not deployed any hiding technique.

At this point, it is clear that Tor does not provide the botnets with the expected capabilities, at least in the way it is currently used.

5. RELATED WORK

While research has studied botnet identification and analysis, no focus has been put into analysing botnet over Tor.

Rossow et al. [16] describe the different techniques that the botmasters apply to create resilient P2P botnets. They present an analysis of the resiliency of different families of P2P botnets against classic attacking methods such as crawling and sinkholing. They show which level of disruption can be achieved using these kinds of attacks and which families are more subject to them. This work gives us an understanding about the increment of resilience produced by the P2P structure but also suggests how this new structure is still attackable.

Andriesse et al. [9] make an in depth analysis of the latest state of the art in resilience in P2P botnets, in particular for the Zeus botnet. In this paper, they dissect the last version of the Zeus protocol describing the algorithm used and the resilient features applied. This work shows how strong and resilient the latest P2P botnets already are without using Tor.

A lot of research has been also done with respect to the security in Tor; in particularly, it focuses on the quality of the anonymity provided.

Elahi et al. [17] address the problem of the entry guard rotation in the Tor protocol. They claim that short-term entry guard churn and explicit time-based entry guard rotation significantly degrade Tor clients anonymity. They also propose an approach to improve the entry guard selection procedure based on trust-based scheme, so that clients pick guards according to how much they trust them.

Biryukov et al. [11] present different vulnerabilities of the Tor protocol. They describe attacks to hidden services, namely denying the hidden service, harvesting hidden service descriptors, identifying entry guards and the deanonymization of hidden services. They point out serious problems in the Tor implementation.

Johnson et al. [13] tie together two important questions about Tor anonymity. What if the attacker runs a relay and what if the attacker can watch part of the Internet? They show that Tor faces great risks from traffic correlation, particularly considering an attacker that can monitor a big part of the network.

These three articles paint a clear picture of the security situation in the Tor network. They address problems in the design and in the implementation of the network that produce anonymity flaws.

6. CONCLUSIONS

Botmasters fight researchers and law enforcement everyday, trying to keep their botnets alive. They design botnets aiming to obtain resilience using any possible means. Lately, they are trying to use the Tor network in order to achieve anonymity for their services and keep their C&C channel hidden. However, we showed that P2P botnets using Tor are still vulnerable to the same kind of attacks such as crawling and sinkholing. Moreover, centralised Tor-based botnets are subject to the vulnerability of Tor itself. In fact, Tor can be affected causing the loss of anonymity if the attacker infects particular relays or a big part of the relays in the network. Seeking more resilient and stealthier properties for the botnets, the botmasters may decide to use Tor assuming that it can provide such properties for free. Instead, botnets are eventually affected by the same anonymity issues that afflict Tor. Hence, even using Tor, the security of the C&C servers is once again compromised.

This does not mean that Tor is not a good solution for botnets but botmasters have to design them taking in consideration the Tor infrastructure and in particular its vulnerabilities. They cannot just exploit the network risking to disrupt it. Instead they have to use it and to enhance it at the same time. Moreover, even at the client side, more accurate techniques can be applied to hide the Tor client. For example, botwriters can compile the Tor source code along with the malware code or apply process hollowing techniques. Tor is an appealing platform for botnets, but the risk of such a platform cannot be underestimated.

REFERENCES:

[1] C. Funk and D. Maslennikov, *IT Threat Evolution Q2 2013*, http://www.securelist.com/en/analysis/204792299/IT, Kaspesky Lab.

[2] T. Dirro, P. Greve, H. Li, F. Paget, V. Pogulievsky, C. Schmugar, J. Shah, R. Sherstobitoff, D. Sommer, B. Sun, A. Wosotowsky, and C. Xu, *McAfee Threats Report Second Quarter 2013, http //www.mcafee.com/us/resources/reports/rp-quarterly-threat-q2-2013.pdf*. McAfee Lab.

[3] J. Salo, *Recent Attacks On Tor*, Aalto University, PhD thesis, 2010.

[4] D. Brown, *Resilient Botnet Command and Control with Tor, http //www.defcon.org/images/defcon-18/dc-18-presentations/D.Brown/DEFCON-18-Brown-TorCnC.pdf*, DefCon 18, 2010

[5] C. Guarnieri, *Skynet, a Tor-powered botnet straight from Reddit, https //community.rapid7.com/community/infosec/blog/2012/12/06/skynet-a-tor-powered-botnet-straight- from-reddit*, Rapid7 2012
[6] R. Dingledine, *Many more Tor users in the past week?, https //lists.torproject.org/pipermail/tor-talk/2013-August/029582.html*, Tor Mailing List, August 2013.
[7] L. Munson, *Tor usage doubles in August. New privacy-seeking users or botnet?*, http://nakedsecurity.sophos.com/2013/08/29/tor-usage-doubles-in-august-new-privacy-seeking-users-or-botnet/29August 2013.
[8] Y. Klijnsma, *Large botnet cause of recent Tor network overload, http //blog.fox-it.com/2013/09/05/large-botnet-cause-of-recent-tor-network-overload/* Fox-It, 5 September 2013.
[9] D. Andriesse, C. Rossow, B. Stone-Gross, D. Plohmann, and H. Bos, *Highly Resilient Peer-to-Peer Botnets Are Here An Analysis of Gameover Zeus*, 8th IEEE International Conference on Malicious and Unwanted Software, MALWARE 2013, Fajardo, Puerto Rico, USA.
[10] M. V. Barbera, V. P. Kemerlis, V. Pappas and A. D. Keromytis, *CellFlood Attacking Tor Onion Routers on the Cheap*, 18th European Symposium on Research in Computer Security (ESORICS). Egham, UK, September 2013.
[11] A. Biryukov, I. Pustogarov, R.P. Weinmann, *Trawling for Tor Hidden Services Detection, Measurement, Deanonymization*, IEEE Symposium on Security and Privacy 2013.
[12] N. Borisov, G. Danezis, P. Mittal, and P. Tabriz, *Denial of Service or Denial of Security? How Attacks on Reliability can Compromise Anonymity*, In the Proceedings of CCS 2007, October 2007.
[13] A. Johnson, C. Wacek, R. Jansen, M. Sherr, and P. Syverson, *Users Get Routed Traffic Correlation on Tor by Realistic Adversaries*, In the Proceedings of the 20th ACM conference on Computer and Communications Security (CCS 2013), November 2013.
[14] *http //www.theguardian.com/world/the-nsa-files* The Guardian.
[15] M Mimoso, *http //threatpost.com/moving-to-tor-a-bad-move-for-massive-botnet/102284*, The Kaspersky Lab Security News Service.
[16] C. Rossow, D. Andriesse, T. Werner, B. Stone-Gross, D. Plohmann, C. J. Dietrich, H. Bos, *SoK P2PWNED Modeling and Evaluating the Resilience of Peer-to-Peer Botnets*, 34th IEEE Symposium on Security and Privacy, S&P 2013, San Francisco, CA.
[17] T. Elahi, K. Bauer, M. AlSabah, R. Dingledine, I. Goldberg, *Changing of the Guards A Framework for Understanding and Improving Entry Guard Selection in Tor*, In the Proceedings of the Workshop on Privacy in the Electronic Society (WPES 2012), Raleigh, NC, USA, October 2012.
[18] C. Rossow, *Using Malware Analysis to Evaluate Botnet Resilience*, Phd Thesis, 2013.
[19] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, *Your Botnet is My Botnet Analysis of a Botnet Takeover*, In Proceedings of the 16th ACM conference on Computer and communications security, CCS09, pages 635647, New York, NY, USA. ACM. 2009.
[20] R. Dingledine, N. Mathewson, P. Syverson, *Tor The Second-Generation Onion Router*, USENIX Security, 2004.
[21] L. Verlier and P. Syverson, *Locating Hidden Servers*, In the Proceedings of the 2006 IEEE Symposium on Security and Privacy, May 2006.

Chapter 5
Cyber Operational Activities

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Key Terrain in Cyberspace: Seeking the High Ground

David Raymond
Army Cyber Center
West Point, New York, USA

Gregory Conti
Army Cyber Center
West Point, New York, USA

Tom Cross
Lancope, Inc.
Alpharetta, Georgia, USA

Michael Nowatkowski
Army Cyber Center
West Point, New York, USA

Abstract: In military doctrine, key terrain refers to areas which, if seized, afford an advantage to an attacker or defender. When applied to geographic terrain, this definition is clear. Key terrain might include a hill that overlooks a valley an enemy wants to control or a crossing point over a river that must be traversed before launching an attack. By definition, dominance of key terrain is likely to decide the overall outcome of a battle. While cyber key terrain is similar to geographic key terrain in some ways, there are also significant and often counterintuitive differences. Some consider cyber terrain to be tied to a physical location and to be represented in cyberspace by routers, switches, cables, and other devices. We will argue that key terrain in cyberspace exists at all of the cyberspace planes, which include the geographic, physical, logical, cyber persona, and supervisory planes [1]. In many cases, features of cyber terrain will not be tied to a specific location, or the geographic location will be irrelevant. In this paper we deconstruct and analyze cyber key terrain, provide a generalized framework for critical analysis, and draw parallels between cyber and physical key terrain while providing examples of key terrain in cyber operations. During a cyber operation, an analysis of key terrain will aid in the strategy and tactics of both the offense and the defense. During peacetime, an understanding of cyber key terrain can be employed broadly, ranging from helping a system administrator focus scarce resources to defend his network all the way to allowing nation-state militaries to develop long-lasting and effective doctrine.

Keywords: *cyber operations, terrain analysis, cyber terrain, key terrain*

1. INTRODUCTION

Any military operation requires a thorough analysis of the situation, referred to in the U.S. military as Intelligence Preparation of the Operational Environment, or IPOE [2]. Along with

The views expressed in this article are those of the authors and do not reflect the official policy or position of West Point, the Department of the Army, or the United States Government

an analysis of the enemy's capabilities and possible courses of action, a fundamental aspect of IPOE is a detailed terrain analysis to identify key terrain. The U.S. Army defines *key terrain* as "any locality or area, the seizure or retention of which affords a marked advantage to either combatant" [3]. Identifying key terrain gives military planners, whether attacking or defending, a physical location upon which to focus their efforts.

Identifying key terrain is straightforward in kinetic conflict; key terrain in cyber operations is likewise critical, but less well understood. In some cases, a hardware device might be cyber key terrain. For example, if your goal is to temporarily deny your opponent access to a tactical network, and if they have a single router connecting them to that network, that router might be key terrain. Some cyber terrain is logical instead of physical. As an example, portions of the Domain Name System (DNS), a distributed, hierarchical, and ever changing database of domain name mappings, might be key terrain in certain situations.

Adding to the complexity is the malleable nature of some cyberspace terrain. The logical structure of a software-defined network (SDN) can change dramatically with no change to the underlying hardware, causing instantaneous shifts in terrain elements such as avenues of approach,[1] obstacles (such as packet filters and firewalls), and key terrain. Battlefield deception is inherently intertwined with key terrain, however in cyberspace deceptive terrain can be easily constructed and moved, a near impossibility on the physical battlefield. Key terrain also has a temporal aspect, a hilltop that is key to a battle might not be so once the battle is over, but in cyberspace these temporal shifts can happen much more quickly, perhaps in milliseconds. Finally, it is not always obvious who controls an element of cyber terrain. While occupation of geographic terrain is often recognized easily by the presence of troops, a cyber operator might be in full control of an adversary's device without them even knowing it.

Whether on the kinetic battlefield or in cyberspace, understanding key terrain in your situation gives you a distinct advantage over an adversary who doesn't conduct this analysis. It helps you to focus your defenses, or your attack. It may also assist in your deception effort by informing how to manipulate your network to foil an adversary attempting to penetrate it.

In this work we examine the notion of key terrain in the traditional domains of land, sea, and air, further analyze cyber terrain, and then merge these concepts to study cyber key terrain. We then provide a framework to describe how the concept of cyber key terrain can be applied in both the offense and the defense.

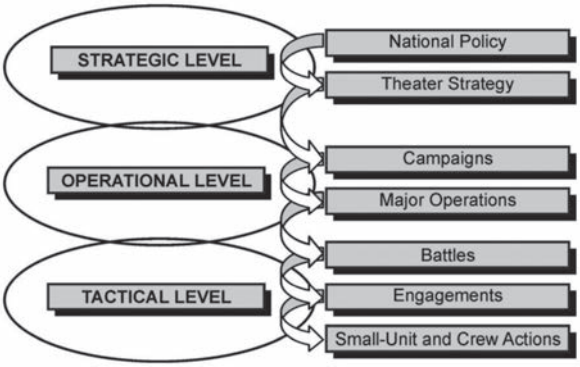
2. KEY TERRAIN IN KINETIC WARFARE

At the tactical level of war, key terrain is a straightforward concept. A hilltop that dominates an enemy's defenses or a bridge across an unfordable river might be key under the right circumstances. Key terrain provides an advantage to a combatant. Therefore, it only exists in a potentially adversarial situation – one in which a place might be attacked and should be defended.

1 An avenue of approach is defined in U.S. Joint doctrine as "[a]n air or ground route of an attacking force of a given size leading to its objective or to key terrain in its path" [8]. In section 3. B. we extend this definition to incorporate elements of cyberspace.

The concept of key terrain is most commonly applied at the tactical level of warfare, however it is relevant at the strategic and operational levels as well. Figure 1 depicts the levels of war from U.S. Army Field Manual 3-0, Operations [4]. The tactical level of war involves individuals and small units engaging in direct hostilities and the above examples of hilltops and bridges apply primarily at this level. The strategic level of war involves nation-states deciding upon national security objectives and using elements of national power (diplomatic, informational, military, and economic) to achieve them. Strategic key terrain might include a nation's capital. For example, the German occupation of Paris in June 1940 caused the French government to flee and put an end to organized resistance against the German invasion, making the city of Paris strategic key terrain. The operational level of war bridges the gap between strategic and tactical and describes a theater of war or a major campaign. An example of operational key terrain is the Khyber Pass, a key supply route between Pakistan and Afghanistan. More than 80 percent of supplies brought in by road to NATO and US forces in Afghanistan is transported through the Khyber Pass [5].

FIGURE 1: FIGURE 7-1 FROM ARMY FM 3-0: OPERATIONS. LEVELS OF WAR.



While applied most often to land-based military campaigns, the idea of key terrain is also useful in naval and aviation contexts. Midway Atoll, an American outpost and airfield 1,300 miles northwest of the Hawaiian island of Oahu, was key terrain in the Pacific theater during World War II. After Japan's attack on Pearl Harbor in December 1941 brought the United States into the war, the U.S. presence at Midway was within Japan's sphere of influence and was perceived by the Japanese as a direct threat to their homeland. This perception was reinforced in April 1942 when Lieutenant Colonel James Doolittle of the U.S. Army Air Corps led a B-25 bomber raid on the Japanese mainland. Admiral Yamamoto was determined to defeat the remainder of the U.S. Pacific Fleet by drawing it into an ambush at Midway. U.S. forces, however, had broken the Japanese naval code and were able to use intelligence gained to ambush and soundly defeat the Japanese fleet, a battle that proved to be a turning point in the Pacific theater.

The term key terrain has been used before to describe non-geographic features of an area of operations. During General David Petraeus' Senate Confirmation Hearing for Commander, International Security Assistance Force (ISAF), U.S. Forces Afghanistan, he stated that in

Afghanistan, as in Iraq, "the key terrain is the human terrain" [6]. In this context, human terrain is defined as "the human population in the [area of operations] as defined and characterized by sociocultural, anthropologic and ethnographic data and other non-geographical information" [7].

3. DEFINING CYBER TERRAIN

The U.S. Department of Defense (DOD) defines cyberspace as a "global domain within the information environment consisting of the interdependent network of information technology infrastructures, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers" [8]. As with human terrain, cyber terrain will not always be directly tied to a physical location, and may include operating systems or application software, network protocols, computing devices, and even individuals or virtual personas. The DOD does not define cyber terrain, so we will define it as *the systems, devices, protocols, data, software, processes, cyber personas, and other networked entities that comprise, supervise, and control cyberspace*.

A. The Nature of Cyber Terrain

The term *terrain* is almost always used to describe physical locations that can be easily pointed to on a map. Since much of cyberspace is virtual, cyber terrain differs from physical terrain in many fundamental ways [9]. As we will see, cyber terrain spans the cyberspace planes [1], so cyber key terrain often manifests itself logically instead of physically. A router that connects a network to an Internet service provider (ISP) is an example of a cyber terrain feature. While this device resides at a specific physical location, it is not the physical location that might make it key terrain, but the logical location of the device in the network. However, physical location is not irrelevant, in that gaining physical access to take a device offline is still a valid attack vector. What it means to 'control' terrain is also different in cyberspace than in physical space. Traditionally, physical occupation of a piece of terrain is required to control it. Furthermore, it is usually obvious to both sides of a conflict who is in control of certain terrain. In cyberspace, physical proximity is not required to control a given device. System administrators routinely access devices from remote locations, and a cyber criminal might gain access to a company's network through the Internet from hundreds of miles away. A skilled attacker will try to hide his presence and remove evidence of his activities on a compromised device. The network administrator might have the illusion of being in control until the attacker needs to influence a network. In fact, an administrator may never know that one of his devices was compromised; even one that was used to penetrate his network.

The virtual nature of cyber terrain makes it possible to dynamically create, modify, and destroy cyber terrain both quickly and frequently; at machine speed. Software defined networking allows logical network architectures to be modified on the fly [10]. A defender might, therefore, be able to modify avenues of approach and move key terrain dynamically in the face of a network attack. An attacker would need to respond in a highly agile manner to overcome these changes to what is effectively the fundamental fabric of the cyber battlefield. The rate of change could far exceed human capacity and require automated responses reminiscent of

high-frequency trading, which is characterized by algorithmic techniques used to rapidly trade securities in fractions of a second [11].

The potential to practice deception operations in cyberspace is vast. Companies have long deployed deceptive 'honeynets', real-looking network segments designed to divert an attacker's attention away from valuable assets within their networks. Using software defined networking, an organization could move critical nodes from one location to another within their cloud infrastructure and instantly reconfigure the network to support the new architecture. An attacker that is pursuing a certain avenue of approach to a target might then have to abandon that pathway in favor of another, which could also be taken away at any time. This could even be done dynamically in the face of a suspected (or known) attack on, or breach of, a network.

We make a distinction between maneuver and fires in cyberspace. U.S. military doctrine defines *maneuver* as "[a] movement to place ships, aircraft, or land forces in a position of advantage over the enemy," and *fires* as "[t]he use of weapon systems to create specific lethal or nonlethal effects on a target" [8]. In cyberspace, we consider an actor to have maneuvered when he has gained access to a device or system as part of a cyber operation. Such access can be authorized or unauthorized, depending on the owner of the system and the nature of the operation. Cyber fires, such as the launching of a software exploit, or phishing email, might be used to enable cyber maneuver. Other fires, such as denial of service (DoS) attacks, are designed to achieve a specific effect without necessarily attempting to facilitate further maneuver.

B. Cyber Terrain and Cyberspace Planes

The cyber planes suggested by Fanelli [12] and refined by Raymond [1] can be used as a framework to identify terrain at various levels. Here we will introduce cyber terrain at each cyberspace plane. The planes are depicted in Figure 2.

1) Supervisory Plane. The supervisory plane provides oversight and the authority to start, stop, modify, or redirect a cyber operation [12]. Cyber terrain at the supervisory plane is comprised of elements of cyberspace that either perform a supervisory function or provide a conduit for command and control.

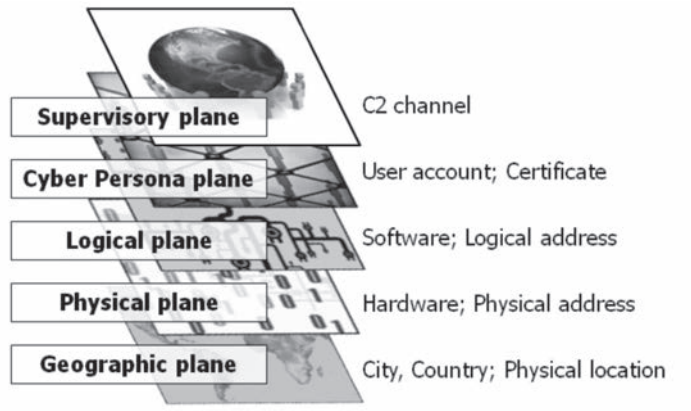
2) Cyber Persona Plane. The cyber persona plane identifies identities in the cyber domain. These identities might have a many-to-one or one-to-many relationship with physical individuals. Here cyber terrain includes such features as user accounts or credentials that provide access to information resources.

3) Logical Plane. This plane consists of the operating system, application software, and software settings on a device, and the logical links between networked devices. Terrain at this level includes a wide range of software systems, services, and protocols that keep networks running and computers doing useful work.

4) Physical Plane. The physical plane maps to the physical layer of the Open Systems Interconnect (OSI) model and includes components of a computer system and attached hardware. This plane is comprised of the devices that people often interpret as being cyber

terrain, such as the routers, switches, and other network devices that physically connect devices in a network.

FIGURE 2: CYBERSPACE PLANES AS DEFINED IN [1], WITH REPRESENTATIVE EXAMPLES.



5) Geographic Plane. The geographic plane describes the geographic area in which an information system, or portions of it, resides. It is the most static of the planes – geography changes at an extremely slow rate. While the logical location of a network device in cyberspace is often more important than its geographic position, geography can also be relevant, and failure to recognize geographic impact to operations can be costly. Geography is also important when considering the potential path of a state-sponsored cyber operation. Just like flying over one country enroute to bombing another could cause an international incident, routing attack packets through a neutral third party could have consequences. This poses a particular challenge during cyber operations when the path that data takes across the Internet can rarely be controlled or even accurately predicted.

C. Cyber Terrain Analysis Using OCOKA

Traditional military terrain analysis uses a process represented by the acronym OCOKA, which stands for Observation and Fields of Fire, Cover and Concealment, Obstacles (man-made and natural), Key Terrain, and Avenues of Approach. Hobbs applies the traditional OCOKA analysis to cyberspace [13] and we expand on his observations below.

1) Observation and Fields of Fire. *Observation* refers to the ability to see enemy forces from a particular vantage point; a *field of fire* combines this ability to observe with the ability to engage enemy targets within the maximum range of your weapon. The idea of observing cyber terrain, while different from physical terrain, is still meaningful. Reconnaissance using *whois* lookups provides IP address ranges and Domain Name Server addresses for Internet domains, along with contact information for domain administrators. Scanning a target network will tell you what hosts are accessible from your vantage point and, by scanning ports, what network

services they are running. Tools like *nmap* can be used to determine which type and version of operating system is running on a particular device and may be used to determine some of the software running on the system [14]. Observing traffic entering and leaving a network can also provide a wealth of information about that network. Examination of source and destination IP addresses can help identify individual hosts. Time-to-live (TTL) values in packet headers can tell you how many routers a packet traversed before leaving the network, which helps to help determine the network architecture. This reconnaissance will help determine which cyber weapons might be successful, giving an indication of your 'fields of fire.'

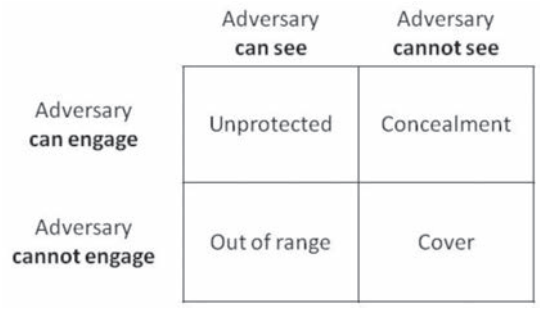
Much like physical terrain, observation is based on vantage point. Someone scanning a network from outside of a firewall will likely get an entirely different result than someone scanning the network from inside. As discussed previously, deception can be used by both attacker and defender. Attackers can hide their source IP address among a flood of false source IP addresses during network scans to hide the origin of the scans. Defenders can use honeynets to draw intruders away from their true network resources. Defenders can also use proxies or network address translation (NAT) to mask their internal network structure.

2) Cover and Concealment. In kinetic terms, *concealment* protects an individual from observation, while *cover* protects one from observation and enemy fire. *Camouflage* is sometimes used to enhance or provide concealment. In cyberspace, as in physical space, a third category exists in which a target can be seen but not engaged and is therefore out of range of an adversary's available weapons. Figure 3 depicts the categories of cover and concealment.

For the network defender, cover is often provided by firewalls that prevent traffic from reaching specific hosts while also protecting those systems from observation. An intrusion prevention system can be used to place hosts out of range of an attack by blocking malicious network traffic, but they do not provide concealment – the hosts behind an intrusion prevent system can still be observed by the attacker through authorized transactions. For an attacker, concealment is used to prevent detection. Polymorphic code and other obfuscation techniques that reduce the potential for signature-based malware detection are often used to camouflage malicious code that could otherwise easily be stopped by intrusion prevention systems. Finally, rootkits can be used by an attacker to conceal the presence of malware on a system [13].

3) Obstacles. In cyberspace, *obstacles* are those technologies or policies that limit freedom of movement within a network. These can include router-based access control lists, air gaps, firewalls, and other devices that are used to restrict the flow of network packets. In cyber terrain, the distinction between obstacles and cover is not always clean. A device installed to limit the enemy's freedom of movement can also provide cover for some network systems. Furthermore, by filtering malicious packets from traffic destined to a system visible on the network, cyberspace obstacles sometimes put target systems out of range of an attackers cyber weapons.

FIGURE 3: CYBER OCOKA CATEGORIES BASED ON ADVERSARY'S ABILITY TO SEE OR ENGAGE TARGET. CONCEALMENT MAY BE ENHANCED BY CAMOUFLAGE.



Other obstacles include user access control systems that prevent network access by all but authenticated users. Even bandwidth constraints that limit traffic flow between two network endpoints can be considered an obstacle. In a kinetic battlespace, obstacles can be either natural (like a ridgeline) or man-made (like a minefield). A similar distinction can be made in cyberspace between intentional obstacles, such as firewalls, and potentially unintentional ones. An example of an unintentional obstacle is a home wireless access point that uses port address translation to map multiple devices to a single IP assigned by an Internet service provider and in doing so, improves security of the network by masking devices inside the network.

4) Key Terrain. Earlier we defined cyber terrain, here we define cyber *key terrain* as systems, devices, protocols, data, software, processes, cyber personas, or other network entities, the control of which offers a marked advantage to an attacker or defender. Aspects of cyber key terrain will be analyzed in detail in Section 4.

5) Avenues of Approach. Avenues of approach in cyberspace are composed of the various paths that can be traversed to reach a target. The physical pathways that connect systems such as switches, routers, fiber, and Ethernet cable are often less relevant than the logical connections facilitated and limited by these devices since the devices traversed by Internet flows can change over time. An HTTP connection to a web server can be an avenue into a target network. Avenues of approach in cyber operations might also include multi-pronged attacks such as a phishing attack on an employee followed by a logical connection to resources left open by the phishing attack.

4. KEY TERRAIN IN CYBERSPACE

Cyber terrain exists across the cyberspace planes and there are many features of cyber terrain that can provide an advantage to one side or the other. By understanding this cyber key terrain, a network defender knows where to focus his energy to prevent penetration and an attacker can select a target within a network that provides maximum potential for success.

A. Examples of Cyber Key Terrain.

Here we provide examples of key terrain for each of the cyberspace planes depicted in Figure 2.

1) Supervisory Plane. Key terrain at this level might include botnet command and control servers that are used to supervise large-scale botnet-based cyber attacks. In June 2013, Microsoft and the U.S. Federal Bureau of Investigation coordinated to disable most of the Citadel botnet by cutting off communication between botnet command and control (C&C) servers and the compromised systems under their control [15]. The Citadel botnet is suspected to have compromised more than five million computers around the world and is thought to be responsible for over half a billion U.S. dollars in losses to businesses and individuals. The botnet C&C servers proved to be cyber key terrain in this operation.

2) Cyber Persona Plane. A system administrator's account might be considered cyber key terrain at the cyber persona plane if possession of that account could be used by an attacker to compromise a defender's resources. Even an unprivileged user account could be key depending on the owner of the account. In early 2011 when HBGary CEO Aaron Barr threatened to expose key members of the hacking collective Anonymous, the group attacked HBGary's network to gain access to Barr's email account login credentials, leading to publication of private emails, website defacement, and significant embarrassment to Barr and HBGary [16].

3) Logical Plane. Key at the logical plane might be the Domain Name System (DNS), which provides logical mappings between domain names (such as www.ccdcoe.org) and their Internet Protocol (IP) addresses (such as 195.222.11.253) [17]. Recent attacks by the hacker collective Syrian Electronic Army (SEA) against the New York Times and other organizations highlight the potential vulnerabilities inherent in failing to recognize a key piece of cyber terrain at the logical plane [18]. The SEA achieved its goal of defacing the New York Times website by targeting the domain name registrar rather than directly targeting the websites themselves, which may have been better defended.

4) Physical Plane. Key terrain on the physical plane might be a poorly configured wireless device that uses an obsolete security protocol. Starting in July 2005, criminals gained access to networks belonging to TJX Companies, Inc., through wireless networks operating at some of their department stores. The stores were using Wired Equivalent Privacy, or WEP, to secure their wireless networks, a protocol that was known to be insecure as early as 2001. Attackers were able to gain access to the company's database servers and steal as many as 200 million customer credit- and debit-card numbers over four years [19].

5) Geographic Plane. The geographic location of infrastructure supporting cyber operations, such as power stations and HVAC controls, could be key terrain. During Hurricane Sandy in October 2012, storm surges surpassed a two-century old record, reaching 14 feet in lower Manhattan. When saltwater rushed over the 12.5 foot seawall at a key substation near Battery Park, 3 million New Yorkers lost power for four days, including the financial district, contributing to the estimated damages of over $20 billion [20] [21].

B. Cyber Key Terrain and the Levels of War

Tactical cyber key terrain are those features that provide tactical advantage to someone attacking or defending a network. Examples might include wireless networks or physical links that allow communication at the local level, firewalls or similar devices that control traffic in a network, or local administrator privileges that could be used to compromise a network. Since tactical actions could have operational or strategic consequences, these examples could also be key terrain at higher levels depending on the context.

Operational key terrain includes features that might give an adversary an advantage in a specific campaign or major operation. A key component of Stuxnet, for example, involved software driver files signed by legitimate digital certificates from two companies that were apparently compromised as part of the development of this malware [22]. The computer systems that those companies used to store their digital certificates constitute operational key terrain. The creators of Stuxnet were able to obtain an asset from those computers that provided them an advantage when they went after their primary objective.

An example of cyber key terrain at the strategic level might be components of a supply chain that produces network devices used by a target entity. A supply chain attack that inserted vulnerable firmware in a government's network routers allowing unauthorized access, for example, could provide an adversary a significant strategic advantage.

Table 1 lists cyber key terrain across the cyberspace planes and the levels of war.

TABLE 1. REPRESENTATIVE CYBER KEY TERRAIN EXAMPLES BY CYBERSPACE PLANE AND LEVELS OF WAR

	Tactical	**Operational**	**Strategic**
Supervisory Plane	• Wireless channel used for C2 communica ions	• Security systems located in a Theater Network Operations and Security Center (TNOSC)	• Nuclear launch systems
Cyber persona Plane	• Local System administrator account	• Network credentials for theater commander	• Email account and password for presidential candidate, Supreme Court justice, or other key figure.
Logical Plane	• The operating system of desktop computer in a targeted organization	• The authoritative DNS server for a popular website	• The software running a regional cellular network
Physical Plane	• A USB key • A cellular phone • An Ethernet switch	• Regional communications cables • Air Defense Artillery Radar/early warning network	• Data center for government agency or major industry
Geographic Plane	• Physical location of network devices providing service to edge network	• Power plant providing electricity to a capital	• Building housing nation's offensive cyberspace operations capabilities

C. A Framework for Leveraging Cyber Key Terrain

Just like in a kinetic scenario, the identification of key terrain is often in the eye of the beholder and depends heavily on context. Two tacticians might look at a defensive sector and, based on experience and their approach to defending an area, identify different key terrain in the sector. Both the defender and attacker must analyze cyber terrain in the context of what he or she considers to be a 'successful' defense or attack and then identify the terrain they perceive will give them an advantage in order to focus their efforts. A general framework for identifying cyber key terrain as a **defender** is given here. This process is reminiscent of the process a tactical commander might take to identify and defend physical key terrain, but our approach is tailored to the realities of cyber terrain.

1. Identify potentially targeted assets. Defenders should start their terrain analysis by identifying the information systems or data that may motivate attackers to target the organization. It is important to keep in mind that the assets that are most valuable to an organization are not always the assets that are most valuable to attackers. Although prudent organizations always consider the risks to their "crown jewels," attackers may be interested in other assets as well, such as an administrative assistant's logon credentials. Therefore it makes sense to work from a model of different threat actors, their motivations, their capabilities, and their tactics in attempting to identify the assets that they may decide to target.

2. Enumerate avenues of approach. What are all of the different vectors that can be used to access each potentially targeted asset? It is important to consider all of the interfaces that the asset has to the outside world that the attacker could leverage on each cyberspace plane, whether they are direct network interfaces, or indirect interfaces such as removable media, or key personnel with physical access.

3. Consider observation and fields of fire. From what locations can the attacker gain access to each interface into the potentially targeted asset? At this point, the analysis may become iterative – if the attacker can reach an interface to the targeted asset from a particular system or network, it is important to enumerate the avenues of approach to that secondary system or network, and determine the locations from which those avenues of approach can be reached, and so on.

It is through this iterative analysis that a picture of key terrain begins to emerge. Are there particular vantage points that provide an attacker with a field of fire that includes many potentially targeted assets? In most networks there are infrastructure components that could provide an attacker broad access to many systems in the network, such as identity and access management systems, core firewalls, network backup systems, and end-point management systems. All of these may be considered key terrain.

It is important for defenders to avoid limiting this analysis to terrain that they control. How might an attacker target other organizations or infrastructure in order to obtain a tactical advantage? Attackers might target suppliers, business partners, service providers, or even third party websites. For example, a "watering-hole attack" is a tactic that involves compromising a website that is frequented by the intended target. Once the website has been compromised, the

attacker has an improved field of fire into their intended victim's computer network, as they can directly access the victim's web browser and provide code for it to execute. All of these vantage points should be considered.

4. Place obstacles, cover, and concealment. Once key terrain has been identified, a defender can begin to take steps to protect it. The most basic step is to limit avenues of approach. Interfaces to key terrain that are unnecessary should be deactivated. Firewalls are often used to limit the number of access vectors into a key asset in a computer network.

Of course, in order for most computer systems to work, they have to be interconnected either directly or indirectly, so it is impossible to close off every access vector. Access vectors that must remain open should be protected. Known vulnerabilities should be patched and weak passwords identified and changed. Intrusion prevention systems have been used for years to block attacks across interfaces that cannot be closed off.

The fact is that neither firewalls nor vulnerability management nor intrusion prevention systems have proven effective in practice against advanced attackers, and this is not merely because defenders have failed to perform a comprehensive terrain analysis. Attackers have proven that they can craft attacks that target vulnerabilities that defenders are unaware of, and they can conceal their attacks in such a way that they cannot be detected.

In light of the effectiveness that attackers have demonstrated at subverting traditional kinds of cover, defenders might benefit from giving more consideration to deception as a part of their defensive posture. As previously discussed, cyber key terrain can be moved, and it can be reorganized in such a way that it ceases to be valuable. A defender could lure an attacker into targeting a piece of key terrain that seems to provide access to a valuable asset, and then change the nature of that terrain once it is compromised. This approach expends attacker resources and forces him or her to reveal capabilities and techniques.

Although honeypots have been a part of defensive approaches to protecting computer networks for a long time, traditional approaches to constructing them have not always kept up with modern attackers and their tactics. It is important to design honeypots that are truly attractive to the kinds of adversaries an organization is most concerned with. A good honeypot should appear to be a key piece of terrain in order to attract an attacker's attention.

An **attacker** has a slightly different perspective as they typically operate with imperfect information about the terrain of the environment they are targeting. Often, cyber terrain cannot be observed until it is accessed, so attackers are forced to engage in a constant process of reassessment of key terrain as they progress deeper into a network. This assessment mirrors the iterative analysis that was (hopefully) performed by the defender.

A careful analysis of avenues of approach, observation points, and fields of fire can provide an attacker with a complete view of his or her options at each stage of the attack. Because attackers may be operating with imperfect information, they may have to make assumptions about the capabilities that controlling a particular asset will afford them, based on how that sort of asset

is typically used by network operators or end users. It is also important for the attacker to try to enumerate the protection technologies employed by the defender. If the attacker can reproduce the defender's complete toolset, he or she can ensure that exploits, malware, and command and control channels are not detected by that toolset.

Of course, attackers need to take care to conceal the reconnaissance used to collect their picture of the cyber terrain, as noisy reconnaissance may result in the attack being identified. Also, attackers must take care to assess whether or not the terrain is what it appears to be, as defenders may have placed honeypots or other deceptive features onto the battlefield.

5. CONCLUSION

An understanding of cyber terrain, and specifically cyber key terrain, is an important part of emerging cyber operations doctrine. It is important for operators to understand that key terrain in cyberspace can have completely different features than key terrain in the traditional sense. A much more robust technical understanding of the cyber landscape is required for a cyber operator to be able to identify and leverage key terrain in cyberspace, but developing this insight could be instrumental in allowing cyber operators to focus limited assets on the most likely path to success during offensive or defensive operations.

BIBLIOGRAPHY:

[1] D. Raymond, G. Conti, T. Cross and R. Fanelli, "A Control Measure Framework to Limit Collateral Damage and Propagation of Cyber Weapons," in *5th International Conference on Cyber Conflict*, Tallinn, Estonia, June 2013.

[2] Department of Defense, Joint Publication 2-01.3: Joint Intelligence Preparation of the Operational Environment, 2013.

[3] Headquarters, Department of the Army, Field Manual 3-90-1: Offense and Defense Volume 1, 2013.

[4] Headquarters, Department of the Army, Field Manual 3-0: Operations, 2011.

[5] S. Masood, "Bridge attack halts NATO supplies to Afghanistan," New York Times, 3 February 2009. [Online]. Available: http://www.nytimes.com. [Accessed 29 November 2013].

[6] A. Garfield, "Understanding the Human Terrain: Key to Success in Afghanistan," Small Wars Journal, 16 July 2010. [Online]. Available: http://smallwarsjournal.com. [Accessed 17 October 2013].

[7] J. Kipp, L. Grau, K. Prinslow and D. Smith, "The Human Terrain System: A CORDS for the 21st Century," *Military Review*, pp. 8 - 15, September-October 2006.

[8] Department of Defense, Joint Publication 1-02: Department of Defense Dictionary of Military and Associated Terms, 2010.

[9] M. Miller, J. Brickey and G. Conti, "Why Your Intuition About Cyber Warfare is Probably Wrong," Small Wars Journal, 29 November 2012. [Online]. Available: http://www.smallwarsjournal.com. [Accessed 15 October 2013].

[10] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks," ONF Whitepaper, April 2013.

[11] Wikipedia, "High-frequency trading," [Online]. Available: http://en.wikipedia.org. [Accessed 16 November 2013].

[12] R. Fanelli and G. Conti, "A Methodology for Cyber Operations Targeting and Control of Collateral Damage in the Context of Lawful Armed Conflict," in *4th International Conference on Cyber Conflict*, Tallinn, Estonia, June 2012.

[13] D. Hobbs, "Application of OCOKA to Cyberterrain," White Wolf Security White Paper, Lancaster, PA, June 2007.

[14] G. Lyon, “nmap.org,” [Online]. Available: http://nmap.org. [Accessed 11 November 2013].
[15] J. Ribeiro, “Microsoft, FBI disrupt Citadel botnet network,” Infoworld Security Central, 6 June 2013. [Online]. Available: http://www.infoworld.com. [Accessed 16 November 2013].
[16] P. Bright, “Anonymous speaks: the inside story of the HBGary hack,” Arstechnica, 15 February 2011. [Online]. Available: http://arstechnica.com. [Accessed 16 November 2013].
[17] Wikipedia, “Domain Name System,” [Online]. Available: http://www.wikipedia.org. [Accessed 17 October 2013].
[18] K. Poulsen, “Syrian Electronic Army Takes Down the New York Times,” Wired, 27 August 2013. [Online]. Available: http://www.wired.com. [Accessed 13 September 2013].
[19] J. Pereira, “How credit-card data went out wirless door,” Wall Street Journal, 4 May 2007. [Online]. Available: http://www.wsj.com. [Accessed 17 February 2014].
[20] J. Donn, J. Fahey and D. Carpenter, “NYC Utility Prepped for Big Storm, Got Bigger One,” Associated Press, 31 October 2012. [Online]. Available: http://bigstory.ap.org. [Accessed 17 October 2013].
[21] C. Burritt and B. Sullivan, “Hurricane Sandy Threatens $20 Billion in Economic Damage,” Bloomberg, 30 October 2012. [Online]. Available: http://www.bloomberg.com. [Accessed 17 October 2013].
[22] N. Falliere, L. Murchu and E. Chien, “W32.Stuxnet Dossier, v1.4,” Feb 2011. [Online]. Available: http://www symantec.com. [Accessed September 2012].

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Fighting Power, Targeting and Cyber Operations

Paul Ducheine
Faculty of Military Sciences
Netherlands Defence Academy, Breda
University of Amsterdam
p.a.l.ducheine@uva.nl

Jelle van Haaster*
Faculty of Military Sciences
Netherlands Defence Academy, Breda
University of Amsterdam
j.vanhaaster@uva nl

Abstract: This article contributes to the operationalisation of military cyber operations in general, and for targeting purposes, either in defence or offence, in particular. The role of cyber operations in military doctrine will be clarified, its contribution to fighting power conceptualised, and the ramifications on targeting processes discussed. Cyberspace poses unique challenges and opportunities; we distinguish new elements that may be used for targeting inter alia for active defence purposes, namely cyber objects and cyber identities. Constructive or disruptive cyber operations aimed at these non-physical elements provide new ways of attaining effects. Assessing the outcome of these cyber operations is, however, challenging for planners. Intertwined network infrastructure and the global nature of cyberspace add to the complexity, but these difficulties can be overcome. In principle, the targeting cycle is suitable for cyber operations, yet, with an eye to (a) the effectiveness of offensive and defensive operations, and (b) legal obligations, special attention will be required regarding effects in general, and collateral damage assessment in particular.

Keywords: *cyberspace, fighting power, doctrine, operations, cyber operations, targeting*

1. INTRODUCTION

Cyber in its most general sense is heralded as a force-multiplier in the arsenal of both State and non-State actors.[1] Although the potential of 'cyber' is uncontested, there remain questions surrounding operationalising cyber means and methods. Since some of these questions remain

* Colonel dr. Paul Ducheine MSc, LL.M. is Associate Professor of Cyber Operations, Legal Advisor (Netherlands Army Legal Service), lecturer and senior guest researcher at the University of Amsterdam. Lieutenant Jelle van Haaster, LL.M., is a Ph.D. candidate focusing on cyber operations at the Netherlands Defence Academy and University of Amsterdam. The authors are grateful to the Board of Editors of the *Militaire Spectator*, for their kind permission to use portions of their article ęCyber-operaties en militair vermogenę (org. Dutch), in: 182 *Militaire Spectator* (2013) 9, pp. 369-387.

1 The current development of doctrine supports this notion, see for instance: U.S. DoD, *DoD Strategy for Operating in Cyberspace* (Washington DC: U.S. DoD, 2011); Netherlands MoD, *The Defence Cyber Strategy* (The Hague: Netherlands MoD, 2012); Russian MoD, Conceptual *Views on the Activities of the Armed Forces of the Russian Federation in the Information Space* [концептуальные взгляды на деятельность вооруженных сил российской федерации в информационном пространстве], available at ccdcoe.org/328 html.

unanswered, the use of cyber in military operations is frequently overlooked.[2] One of the issues leading to dismissal of 'the cyber option' is the limited understanding of the effects and implications of the use of cyber weapons in doctrinal thought and operational processes such as targeting. Understanding new means and methods is vital to adequate appreciation of, and operationalising their potential in offensive, defensive and stability operations.

Active cyber defence is generally conceived as 'entailing proactive measures that are launched to defend against malicious cyber activities or cyber attacks'.[3] States tend to entrust their armed forces with a prominent role in securing cyberspace, and hence armed forces will prove crucial in taking proactive measures both domestically and internationally. Before being able to actually conduct cyber operations within the context of active cyber defence, the armed forces have to effectively incorporate cyber capacities within their organisations. Only then can these new capabilities be used effectively for the purposes stated, including active defence, offence and supportive roles.

This article will clarify the role of cyber operations in military doctrine, conceptualise its contribution to fighting power, and discuss potential ramifications on the targeting cycle. By doing so it will contribute to the debate regarding the operationalisation of military cyber means and methods.

Contemporary military operations are not conducted stand-alone; they are a means to an end and are conducted in parallel with other (non-) military activities.[4] In order to place the military instrument in its proper context, we will first briefly expand on instruments of State power and focus on the conceptualisation of fighting power and conventional military operations (§2). Before expanding on cyber operations, it is necessary to define the unique characteristics of cyberspace (§3), and once cyberspace's landscape has been examined we will turn to cyber operations and their contribution to fighting power (§4-5). Lastly we will discuss the ramifications of conducting cyber operations for conventional targeting procedures (§6).

When describing and conceptualising the role of cyber operations, Allied doctrine will be used, primarily focusing on that published by the North-Atlantic Treaty Organisation (NATO), but supplemented with the doctrine publications of other allies. For military cyber operations we use the internationally commended definition stemming from the Tallinn Manual: 'The employment of cyber capabilities with the primary purpose of achieving [military] objectives in or by the use of cyberspace.'[5] We will discuss the subtleties and implications of this definition in this contribution.

2. THE MILITARY INSTRUMENT

In order to provide security, and for the protection of vital strategic interests, States may rely on their instruments of power: integrated or joint military power on land, sea, and in the air, as well

2 See for instance: Amber Corrin, 'The Other Syria Debate: Cyber Weapons,' fcw.com/articles/2013/09/04/cyber-weapons-syria.aspx (accessed 30 October, 2013).
3 CCDCOE, 'Latest News', ccdcoe.org/cycon/home.html (accessed 14 March, 2014).
4 NATO, AJP-1(D): Allied Joint Doctrine (Brussels: NATO Standardization Agency, 2010). Sections 107-110.
5 Michael N. Schmitt (gen. ed.), Tallinn Manual on the International Law Applicable to Cyber Warfare (Cambridge University Press, 2013). p. 258.

as diplomatic, economic, and informational means.[6] Apart from the diplomatic, informational, military, and economic instruments, the so-called DIME-instruments,[7] NATO recognises the 'wide utility [of] civil capabilities' in contemporary operations.[8] Thus, States nowadays have various instruments for achieving strategic goals to the detriment or in support of other States or non-State actors. The use of force is just one of those instruments, although it is quite different from the other instruments.[9]

Fighting Power

Armed forces apply fighting power[10] consisting of three elements: the physical, moral, and conceptual components (see Figure 1).[11] The physical component comprises first and foremost the manpower and equipment that provide the 'means to fight'.[12] Equipment consists of military platforms, systems, weapons and supplies of 'operational or non-operational and deployable or non-deployable' nature.[13] Apart from material elements, the physical component also entails sustainability and (operational) readiness.[14]

The moral component[15] involves 'the least predictable aspect of conflict', namely 'the human element'.[16] It entails 'good morale and the conviction that the mission's purpose is morally and ethically sound'.[17] The moral component is rooted in three 'priceless commodities: ethical foundations, moral cohesion and motivation'.[18] In addition, effective leadership is vital.[19]

FIGURE 1. FIGHTING POWER

6 Antulio J. Echevarria II, *Clausewitz and Contemporary War* (Oxford: Oxford University Press, 2007). p. 144.
7 NATO, *AJP-1(D)*. Sections 107-110.
8 Ibid. p. 1-3. Section 111.
9 Jachtenfuchs, *The Monopoly of Legitimate Force Denationalization, Or Business as Usual?* p. 38.
10 British Army, *ADP Operations* (Shrivenham: Development, Concepts and Doctrine Centre, 2010). p. 2-2.
11 NATO, *AJP-1(D)*. Sections 120-123.
12 British Army, *ADP Operations*. p. 2-31.
13 *Ibid*. p. 2-32.
14 Netherlands MoD, *Netherlands Defence Doctrine* (NDD) (2013). p. 69.
15 The Netherlands Defence Doctrine (NDD) refers to a 'mental component', contrary to the NATO and British 'moral component'.
16 British Army, *ADP Operations*. p. 2-10.
17 NATO, *AJP-1(D)*. Section 121.
18 British Army, *ADP Operations*. p. 2-11.
19 Netherlands MoD, *NDD*. p. 67.

The conceptual component 'provides the coherent, intellectual basis and theoretical foundation for the deployment of military units and troops'.[20] The higher levels of doctrine, the strategic and the operational, 'establish the philosophy and principles underpinning the approach to conflict and military activity'.[21] Apart from guidance, 'the conceptual component also plays a significant role in the preservation and development of the institutional memory and experience'[22] through education, innovation and lessons identified.[23]

Thus, fighting power entails the ability to effectively conduct military operations. However, fighting power is 'more than just the availability of operational means (capacities); there must also be the willingness and ability to deploy these means (capability)'.[24] When properly developed, 'capacities are elevated to capabilities' and they become fighting power.[25] Fighting power will then be employed effectively to achieve strategic goals, whether alone or in unison with other strategic instruments; this is the 'comprehensive approach'.[26]

Operation, the Manoeuvrist Approach and Comprehensiveness

Armed forces project fighting power through military operations. Operations vary in form, purpose, size, duration, and vector: land, sea, air, space, and cyberspace. This section will focus on the conceptualisation of administering fighting power through military operations. The Manoeuvrist Approach is vital to understanding the rationale for conducting military operations. This approach 'focuses on shattering the adversary's overall cohesion and will to fight, rather than his materiel […] it is an indirect approach'.[27] The emphasis is on the adversary's moral and conceptual component rather than on the physical; the purpose is to degrade cohesion in components of an adversary's fighting power.[28] The integration of various components – the Comprehensive Approach – is used not only at the strategic level, but also in actual operations at lower levels.

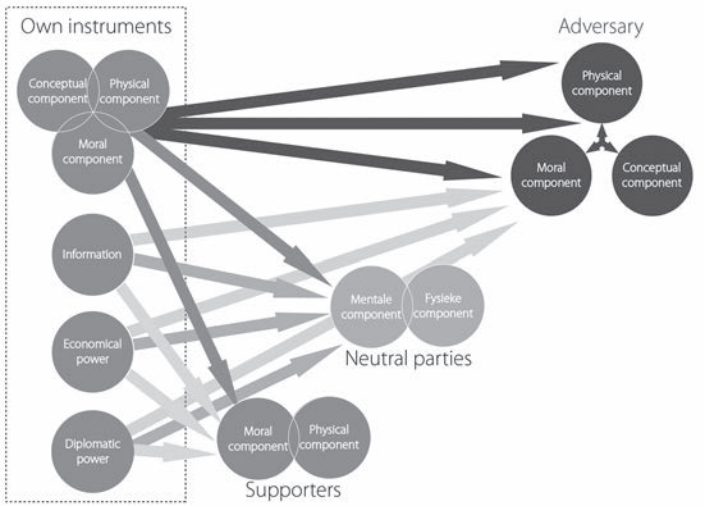
Interpreted in a broader and more modern sense, operations entail influencing actors, as opposed to the adversary, by employing different instruments in addition to the military instrument.[29] Contemporary conflict is characterised by a '[large] number of actors […] intensified by our "open" world, in which everyone can keep abreast of each military operation'.[30] Thus, operations are no longer primarily aimed at opponents, but at a wide range of actors including 'population groups, parties, countries and organisations with which there is no physical interaction'.[31]

Consequently, the military instrument is no longer the only or prime instrument in an area of operations. Activities should be tailored to increase and maintain support for operations by

20 Netherlands MoD, *NDD*. p. 71.
21 British Army, *ADP Operations*. p. 2-5.
22 Netherlands MoD, *NDD*. pp. 70-71.
23 British Army, *ADP Operations*. pp. 2-9, 2-10.
24 Netherlands MoD, *NDD*. p. 66.
25 Netherlands MoD, *NDD*. p. 66.
26 NATO, AJP-1(D). Sections 226-232.
27 *Ibid*. Section 611.
28 British Army, *ADP Operations*. p. 2-6.
29 Netherlands MoD, *NDD*. p. 108.
30 Netherlands MoD, *NDD*. p. 108.
31 *Ibid*. p. 108.

employing various DIME instruments.[32] Operations aim to decrease support to adversaries, and generate support from others.[33] Figure 2 illustrates this conceptualisation of influencing adversaries, neutral parties, and supporters.

FIGURE 2. EMPLOYING INSTRUMENTS OF STATE POWER



Activities or operations addressing adversaries are, by definition, *disruptive* in nature (Figure 2, red arrows). An attempt is made to shatter overall cohesion, which only exists by virtue of clear lines of communication, whether in terms of information or leadership or through attacking or addressing the moral and conceptual component. Without cohesion, morale, and effective leadership, opposing forces can more easily be defeated, destroyed, or outmanoeuvred.

Operations addressing neutrals and supporters are constructive in nature. Their aim is to increase support for one's own operations. By influencing neutral actors, an attempt is made to convince them to join or support the own cause (Figure 2, blue and grey arrows). The goal is to keep them neutral, but preferably to make them supportive. By reinforcing the power of supporters physically by, for example, materiel and training, the foothold within supportive groups can be increased either morally or economically (Figure 2, blue and grey arrows).

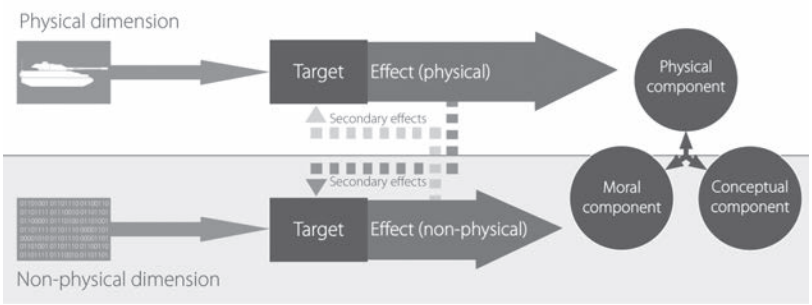
Means to an effect

Activities conducted by armed forces are a means to an end. They are intended to achieve a predefined kinetic or non-kinetic effect to the detriment or support of an actor. To that end, both lethal and non-lethal, physical and non-physical means can be applied.[34]

Lethal and non-lethal or physical and non-physical effects are complementary and intertwined. Destroying enemy materiel and personnel, part of the physical component, will primarily cause physical effects, but will also affect enemy morale, part of the moral component (see Figure 3).

32 *Ibid.*
33 *Ibid.*
34 NATO, *AJP-1(D).* p. 6-3.

FIGURE 3. MEANS, TARGETS AND EFFECTS



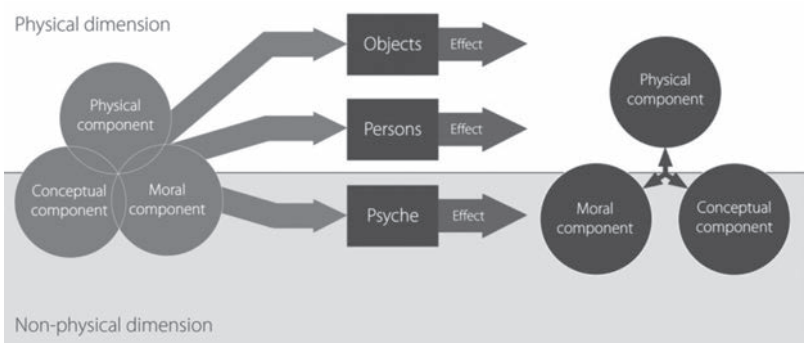
Targets

Effects, whether physical or non-physical, are addressed at a target, or addressee, the entity against which the constructive or disruptive activity is addressed. Activities or operations are conducted against, or in support of, other actors' power, including fighting power. Effects are achieved by engaging targets; these targets and addressees are selected from an actor's physical, moral, and conceptual component.

In the physical dimension objects and persons are targetable, constructively or disruptively (see Figure 4). Objects are tangible elements, for instance military systems and supplies. People vary from individuals to groups and may be hostile, neutral, or supportive.

In the non-physical dimension, the psyche of people is targetable, with the purpose of influencing the moral and conceptual components, as well as the cohesion between the components of fighting power, either constructively or disruptively. By transmitting information, an attempt is made to influence morale, mind-set, and leadership. Besides this, the cognitive perception of the situation may be altered. Effects against an actor's psyche are primarily non-physical in nature, although they can cause secondary effects (see Figure 3).

FIGURE 4. TARGET AND EFFECTS



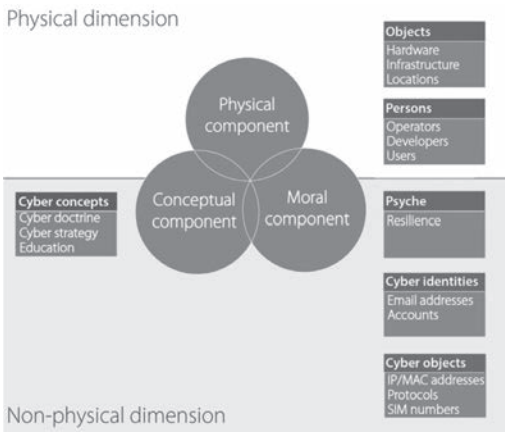
We have briefly described doctrinal viewpoints on military operations or activities. New technical developments can result in new possibilities for conducting operations, but these developments may also pose risks. In the next part we will reflect on the influence of the digital domain, or cyberspace, and cyber operations on doctrinal thinking.

3. CYBERSPACE

Cyberspace, often referred to by the popular media, is as yet poorly understood. The exact meaning of cyberspace is usually ill defined and unclear.[35] Before being able to touch on cyber operations, it is necessary to briefly delve into the meaning of cyberspace. For the purpose of this contribution, the definition offered by Chatham House is used: 'the global digital communication and information transfer infrastructure'.[36]

Cyberspace shares tangible elements with conventional domains of air, land, sea, and space,[37] but is unique as it also contains virtual, more or less ethereal, elements. Cyberspace is frequently depicted as a three layer model with five sub-layers.[38] For our purposes, and in line with the analysis above, we will scale this down to two dimensions: the physical and the non-physical. The physical dimension comprises people and objects, the physical network infrastructure such as hubs, routers, and cables, and the hardware such as computers, smartphones, and servers.[39]

FIGURE 5. FIGHTING POWER IN CYBERSPACE



35 Illustrative is the document *Securing America's Cyberspace, National Plan for Information Systems Protection An Invitation to a Dialogue* (Washington, DC: The White House, 2000). The document equips 33 notions with a cyber prefix, there are only two cyber-terms defined.

36 P. Cornish, D. Livingstone, D. Clemente & C. Yorke (2010). *On Cyber Warfare*, London: Chatham House, p. 1.

37 U.S. Army, TRADOC *Pamphlet 525-7-8 Cyberspace Operations Concept Capability Plan 2016 2028* (Fort Eustis: TRADOC, 2010). p. 9.

38 U.S. Army, *TRADOC Pamphlet 525-7-8*. p. 8, consisting of a physical, logical, and social layer comprising of the following five components: 'geographic, physical network, logical network, cyber persona and persona'. There are also other approaches to layers of cyberspace. The Open Systems Interconnection (OSI) model describes seven layers: the physical, data link, network, transport, session, presentation, and application layers. The Transmission Control Protocol/Internet Protocol (TCP/IP) recognises four layers: the link, internet, transport, and application layers. The United States Army in turn recognises three: the physical, logical, and social layers.

39 U.S. Army, *TRADOC Pamphlet 525-7-8*. p. 9.

Although based on physical elements, the distinguishing feature of cyberspace is the non-physical dimension. Virtual elements enable the transmission of data between objects in the physical network infrastructure and people.[40] Two virtual elements, the 'virtual reflection' of tangible objects and people, can be recognised: cyber objects and cyber identities.

Cyber objects are the logical elements enabling interoperability and communication between physical objects: protocols, applications, the domain name system,[41] operating systems software,[42] IP-addresses,[43] media access control (MAC) addresses,[44] encryption, and other data.[45]

Cyber identities are the digital and virtual identities of people, individuals, groups, and organisations: e-mail accounts, social-media accounts, and other virtual accounts such as phone numbers.[46] Cyber identities exist by virtue of the social and professional use of cyberspace.[47]

The non-physical dimension is the essence of cyberspace's uniqueness. Without the non-physical dimension, cyberspace would not exist. This exceptionality of cyberspace presents both opportunities and risks.

4. FIGHTING POWER IN CYBERSPACE

The question now is: how do these two 'cyber elements' relate to fighting power? This section will therefore elaborate on the components of fighting power in cyberspace by reflecting on the physical, moral, and conceptual components in cyberspace.

Physical Component

The physical dimension of cyberspace incorporates elements from the physical component of fighting power; it similarly envelops tangible objects and persons. Tangible objects relate to the network hubs, the routers, servers, and computers;[48] the physical network infrastructure, such as optic fibre or copper wire;[49] and objects facilitating non-wired transmission between hubs, such as cell sites or mobile phone masts.[50] The notion of 'persons' relates to operators of objects and users of cyberspace; for example tweeters, followers, software developers, and 'hackers'. The physical component also comprises education and training. Training and education may include conducting cyber exercises,[51] testing cyber capacities in a digital and preferably isolated test range, and supplementary education.

Cyber objects and cyber identities?

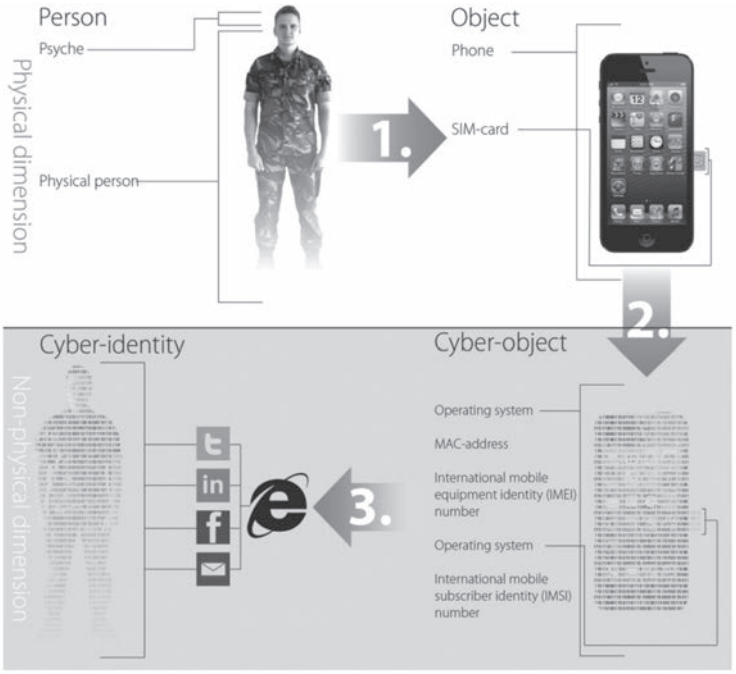
Persons and objects in cyberspace communicate using software, applications, accounts, and

40 U.S. Army, *TRADOC Pamphlet 525-7-8*. p. 9.
41 DNS system: The system used to resolve IP addresses to comprehensible website names.
42 Operating system: The software enabling the functioning of hardware.
43 IP address: The digital postal code of hardware.
44 MAC address: The identification number/code of a particular device.
45 Often referred to as the logical network layer.
46 Often referred to as the cyber persona layer.
47 David J. Betz & Tim Stevens (2011) *Cyberspace and the State*, Adelphi Series, 51:424.
48 U.S. Army, *TRADOC Pamphlet* 525-7-8, p. 9.
49 *Ibid*. p. 9.
50 Jason Andress & Steve Winterfeld, *Cyber Warfare*, 1st ed. (Waltham: Syngress, 2011). p. 120.
51 Such as NATO CCDCOE's exercise 'Locked Shields' and NATO's Cyber Exercise 'Cyber Coalition'.

protocols stemming from the non-physical dimension. These intangible entities differ from physical objects; hence their categorisation within the fighting power concept is potentially problematic.

Cyber objects and cyber identities, being merely reflections of objects and persons, are non-physical and intangible, though intrinsically linked to their physical counterparts, although not necessarily directly. They enable the functioning of cyberspace. This is illustrated in Figure 6.

FIGURE 6. THE PHYSICAL DIMENSION HOSTS PERSONS AND PHYSICAL OBJECTS, IN THIS CASE A PERSON AND HIS SMARTPHONE. BY USING HIS SMARTPHONE (STEP 1), A PERSON CAN MANIFEST HIMSELF ON THE INTERNET (STEP 2). APART FROM THE SMARTPHONE'S PHYSICAL ELEMENTS FACILITATING DATA-EXCHANGE (E.G. ANTENNA), THERE ARE NON-TANGIBLE ELEMENTS REPRESENTING THE SMARTPHONE IN CYBERSPACE WHICH WE CALL 'CYBER OBJECTS', SUCH AS THE IP AND MAC ADDRESS, IMEI NUMBER IDENTIFYING THE SMARTPHONE, IMSI NUMBER IDENTIFYING THE USER, OPERATING SYSTEMS, AND OTHER SOFTWARE. BY MAKING USE OF THE INTERNET TO CREATE, FOR EXAMPLE, SOCIAL-MEDIA ACCOUNTS (STEP 3), A PERSON CREATES HIS CYBER IDENTITY.



Conceptual and moral component

Cyber and regular operations alike require doctrinal and operational preparation. The novel challenges and opportunities of cyber operations have to be grasped before cyber capacities can be effectively employed. These lessons have to be integrated in military training and education. Apart from being well trained and educated, armed forces require motivated personnel. Most importantly, cyber operators and developers need to have a military mind-set, which includes

for example basic knowledge of 'strategy and tactics'.[52] These elements are incorporated in the conceptual and moral component.

In order to adequately use the armed forces, military planners need to understand the inherent cohesion between the components of fighting power and be able to assess the potential contribution of cyber operations and cyber capacities to instruments of State power, fighting power and operations. To be able to do so, military planners should have sufficient knowledge of the interrelated dimensions of cyberspace. Such understanding is necessary in order to comprehend the links between social, technical, and operational processes. Once proficient, the armed forces can further tread within the non-physical realm through cyber means and methods.

Business as usual?

We have introduced distinguishing features of cyberspace, the non-physical dimension, cyber objects, and cyber identities. Some would argue that these features are not new; they fit easily within effects-based operations and information operations, and are merely an example of a soft power instrument.

Although cyber operations may conceptually share similarities with these operations, they differ in capability and targeting and are truly novel and different from other operations. The very existence of cyber objects and cyber identities results in a vast range of new possibilities; these opportunities have to be grasped, which requires awareness, acceptation, and adaptation.

Another striking difference is in the concepts of time and space. Cyber operations can be conducted at the speed of light. People and tangible objects reside within a geographically delineated State. By manifesting themselves through cyber objects and cyber identities, their reach extends globally.

Cyber object and cyber identity can, in principle, be traced back to their physical counterparts, but defending or striking back with cyber operations may prove to be politically, legally, and technically challenging.

Cyber fighting power

This section discusses the place of 'cyber' within fighting power. The concept of fighting power, as we have interpreted it, can accommodate cyber capabilities. We find cyber in the physical, conceptual, and moral components in the form of persons, be they operators, developers, or users; tangible objects such as the physical network infrastructure; and the psyche; for example, the military mind-set.

Cyber is unique with regard to the non-physical dimension of cyberspace, which includes new elements we have dubbed 'cyber objects' and 'cyber identities'. These elements can be used to access cyberspace. We will briefly discuss how to employ these elements in the following paragraph.

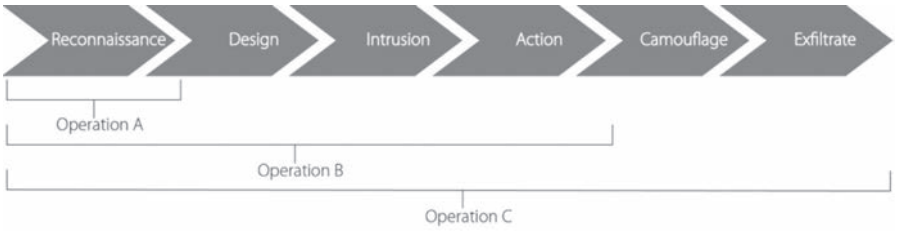
52 Andress & Winterfeld, p. 63.

5. CYBER OPERATIONS

We understand cyber operations to be 'the employment of cyber capabilities with the prime purpose of achieving [military] objectives in or by the use of cyberspace'.[53] Similar to conventional operations, the goal of cyber operations is to achieve an effect, to influence actors in or through cyberspace.

Actors can be influenced *in* or *through* cyberspace. Effects can be achieved *in* cyberspace by creating constructive or disruptive effects vis-à-vis the physical or non-physical dimension of cyberspace, using both kinetic and non-kinetic means. Conversely, constructive and disruptive effects can also be attained *through* cyberspace by, for instance, employing social-media applications to influence people or employing malware against aerial-defence systems. Cyber operations can achieve these effects stand-alone or in parallel with other operations.[54]

FIGURE 7. PHASES IN CYBER OPERATIONS



Phasing and Purposes

Cyber operations, like all military operations, have different phases, each having a different purpose. Although there are different approaches towards naming phases and sub-phases,[55] the general consensus is illustrated in Figure 7. Cyber operations do not necessarily undergo each and every phase; it varies between operations. If the goal is to gather information regarding vulnerabilities by scanning a system or network,[56] the cyber operation will stop at the reconnaissance phase (Figure 7, operation A), whereas an operation aimed at penetrating and creating a foothold in the system might undergo phase one through to phase five (Figure 7, operation B). A fully-fledged cyber operation intended to implant, retrieve, or steal a particular piece of information from a network might go through all six phases (Figure 7, operation C).

Target/addressee and effects

As with regular operations, cyber operations are addressed at a target in order to attain a desired effect. New possibilities arise since there are new elements that can be targeted: cyber identities and cyber objects. The overall goal, however, remains to influence supportive, neutral, and opposing actors.

53 Schmitt (gen. ed.), *Tallinn Manual*, p. 258.

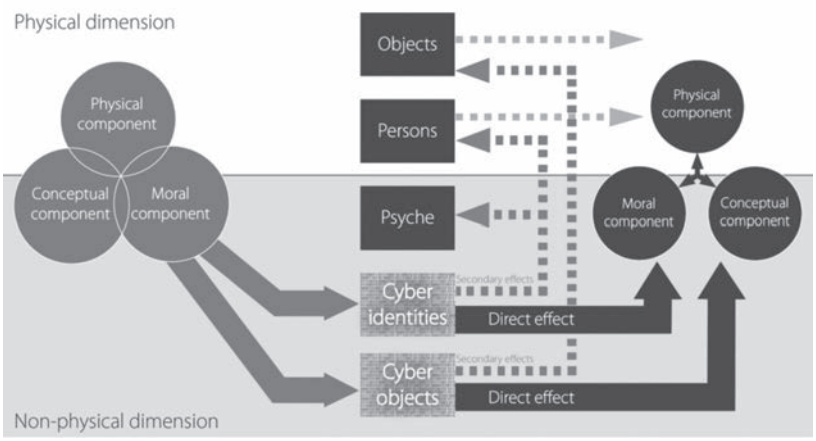
54 Terry D. Gill & Paul A. L. Ducheine, 'Anticipatory Self-Defense in the Cyber Context', 89 *US NWC International Law Studies* (2013), pp. 438-471.

55 Andress & Winterfeld, p. 171: Recon, scan, access, escalate, exfiltrate, assault, sustain; Lech J. Janczewski & Andrew M. Colarik, *Cyber Warfare and Cyber Terrorism* (Hershey: Information Science Reference, 2008). p. xv: Reconnaissance, penetration, identifying and expanding internal capabilities, damage system or confiscate data, remove evidence.

56 For instance by using Nmap (Network Mapper), which enables users to discover vulnerabilities within networks.

Cyber operations are conducted against cyber identities and cyber objects, resulting in a predefined effect vis-à-vis an actor. If successful, they result in a direct effect against these two cyber elements but, although targeting cyber objects and cyber identities, secondary effects are generated against or in support of persons, objects, and psyche (see Figure 8).

FIGURE 8. CYBER OPERATIONS AND EFFECTS



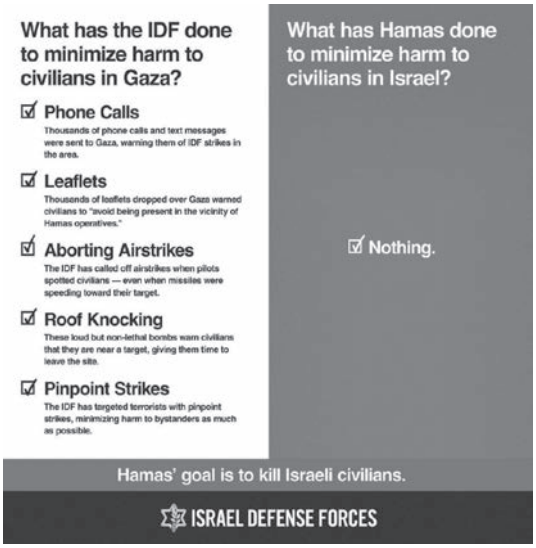
For instance, by addressing the Twitter account of a commander which forms part of his cyber identity, the direct effect is a change in that cyber identity. The secondary effect, an alteration of his state of mind, is achieved when the commander consumes the particular piece of information on his Twitter feed, which may or may not result in a psychological effect felt in his psyche. Another example is targeting the control system of an industrial machine. Initially the control system software is altered, but there are secondary results in a physical effect, for instance operating failure.

The effects achievable through cyber operations are diverse, both the constructive and the disruptive. However, even without conducting constructive or disruptive cyber operations, the mere availability of unprecedented quantities of information in cyberspace reinforces the intelligence position of every actor. We will briefly discuss how cyber identities and cyber objects can be used to generate such effects.

Constructive effects

Constructive effects can be achieved by using cyber identities and cyber objects.

FIGURE 9. USING CYBER IDENTITIES: IDF



1) Physical support

By physically supporting neutral and supportive actors, their capacity to act in cyberspace can be reinforced. Cyber capacity depends strongly on the qualitative state of networks and underlying infrastructure. By providing infrastructure, for instance computers, mobile phone masts, routers, and servers, the position of other actors in cyberspace can be reinforced and their perception or situational awareness influenced to the benefit of the sponsor. Similarly, deploying a Computer Emergency Response Team (CERT) to assist actors in securing theirs networks reinforces the position of those actors and alters their perception and situational awareness. Physical support, or the prospect thereof, could result in an increased foothold within supportive actors or an alignment shift by neutral entities.

FIGURE 10. IDF NOTIFYING HAMAS OPERATORS OF IMPEDING ACTION

FIGURE 11. KENYAN POLICE THREATENING TERRORISTS DURING THE WESTGATE SHOPPING MALL SIEGE IN NAIROBI AFTER TERRORISTS CLAIMED TO STILL OCCUPY THE MALL VIA TWITTER

2) Cyber identities

By using cyber identities, actors can be influenced. Constructive effects can consist of attempts to induce alignment-shift within neutral actors, both individuals and groups, or to reinforce the positions of supporters. In order to do so, armed forces can use social-media accounts to broadcast general information or interact with the accounts of neutral and supportive actors. Through these channels they can explain the rationale behind military operations, counter false information,[57] provide practical information regarding operations, or generate support (see Figure 9). The purpose of these activities is keeping neutral actors neutral at the least and increasing support for a mission.

3) Cyber objects

Cyber objects can be constructively used to influence neutral actors and supporters. Such effects can be generated through providing neutral and supportive actors the tools needed to protect networks such as antivirus software, virus definitions, and signatures and known exploits; tools to better use cyberspace such as data mining software, social media management software, and tools for intelligence purposes; and tools needed to exploit adversary vulnerabilities such as malware, root kits, and botnets.

Disruptive effects

Whereas constructive effects are generated to influence and support friendly actors, armed forces attempt to generate disruptive effects against an adversary.

1) Physical disruption

By physically disrupting cyber capacities belonging to neutral and supportive actors, their capability to act in cyberspace is reduced. Cyber capacity and capability strongly depend on the quality of networks and infrastructure. A network can most easily be disrupted when armed forces have access to the physical network infrastructure.[58] Actors that are able to gain access to or target network infrastructure are capable of disrupting network traffic by methods 'that predate the internet by decades', namely '[c]utting the […] lines'.[59] However, there are other benefits when operators have physical access to network infrastructure: there are no firewalls to be circumvented and they can install, uninstall, and reverse-engineer hardware and software.

2) Cyber identities

Adversary cyber identities such as spokespersons, commanders and their most influential supporters can be targeted. One of the means is decreasing their credibility, for instance by countering the validity of what they publish, highlighting false facts or claims and generally questioning their legitimacy. In order to do so, cyber identities can be used to engage and interact with the adversaries' cyber identities for the purpose of nullifying their influence.

Apart from decreasing credibility, friendly cyber identities can be used to psychologically

57 See for instance: J. Voetelink, 'Lawfare,' *Militair Rechtelijk Tijdschrift* 106, no. 3 (2013), 69-79.; Charles J. Dunlap Jr, 'Lawfare Today: A Perspective,' *Yale Journal of International Affairs* 3 (2008), 146.

58 Jason Andress and Steve Winterfeld, *Cyber Warfare Techniques, Tactics and Tools for Security Practitioners*, 2nd ed. (New York: Syngress, 2014). p. 137.

59 Carol Matlack, 'Cyberwar in Ukraine Falls Far Short of Russia's Full Powers,' Bloomberg Business Week, businessweek.com/articles/2014-03-10/cyberwar-in-ukraine-falls-far-short-of-russias-full-powers (accessed March 11, 2014).; See also: Reuters, 'Ukrainian Authorities Suffer New Cyber Attacks,' Reuters, reuters.com/article/2014/03/08/us-ukraine-cricis-cyberattack-idUSBREA270FU20140308 (accessed March 11, 2014).; Andress and Winterfeld, *Cyber Warfare Techniques, Tactics and Tools for Security Practitioners*. p. 139.

influence adversary cyber identities. Through publishing information regarding upcoming military operations, which may or may not be true, a psychological effect may be generated (see Figure 10).[60]

Adversaries' cyber identities can also be personally addressed, and a message tailored to the specific strengths and weaknesses of a target will undoubtedly affect the psyche of the person 'behind' a cyber identity (see Figure 11).[61]

Adversary cyber identities can also be blocked or hijacked. The easiest way of blocking a cyber identity is requesting the social media company to do so,[62] but there are other means that supersede the companies' authority.[63] Adversary cyber identities can also be hijacked, for instance through 'guessing' credentials[64] or employing social engineering techniques such as phishing and pharming.[65] Once hijacked, the adversary's identity can be used at the discretion of a commander. He could use it in order to deceive adversaries, publish false information to the benefit of own goals,[66] or he could just deactivate and thereby nullify the influence of the account.

3) Cyber objects

Cyber objects belonging to adversaries such as operating systems, malware and other software or data can be used and exploited.

a) Monitoring

Armed forces can gather information about an adversary's cyber objects by collecting information about their networks. Before being able to do so, the mission's cyberspace landscape has to be mapped. This 'map' would include the types of machines used, software versions, port configurations, active or live machines, interdependencies, and the general network environment. By employing software such as Nmap, such information can be gathered.[67] When armed forces have mapped the network environment in an area of operations, this information can be used to increase situational awareness of cyber activities and to earmark weak spots.

b) External manipulation

Should operational circumstances require cyber objects to be denied, denial of service attacks (DOS) can be employed. In order to be able to conduct an effective DOS attack, armed forces should possess a so-called 'botnet', which is a network of computers capable of spawning

60 Tweet @IDFSpokesperson, via <twitter.com/IDFSpokesperson/status/268780918209118208>, accessed 12 January 2014.

61 Tweet @PoliceKE, via: <twitter.com/PoliceKE/status/382161864106737664>, accessed 12 January 2014.

62 See for instance: Bill Gertz, 'User Suspended: Twitter Blocks Multiple Accounts of Somali Al-Qaeda Group during Kenya Attack,' freebeacon.com/user-suspended/ (accessed January 8, 2014).

63 For instance reporting a user '*en masse*' will result in account suspension.

64 For example by making use of 'brute force' attacks employing tools such as THC Hydra ('Hydra') and John the Ripper ('John') to automatically guess credentials.

65 Andress & Winterfeld, p. 141.

66 Cnaan Liphshiz, 'Israeli Vice Prime Minister's Facebook, Twitter Accounts Hacked,' jta.org/2012/11/21/news-opinion/israel-middle-east/israeli-vice-prime-ministers-facebook-twitter-accounts-hacked (accessed January 8, 2014); Grace Wyler, 'AP Twitter Hacked, Claims Barack Obama Injured in White House Explosions' businessinsider.com/ap-hacked-obama-injured-white-house-explosions-2013-4 (accessed January 8, 2014).

67 Nmap (Network Mapper) enables users to scan networks to collect information regarding port configuration, vulnerabilities, operating systems and active machines. Source: Nmap, 'About,' nmap.org (accessed March 11, 2014).

large amounts of data on command.[68] Creating a botnet would require some preparation, since malware has to be written or bought, distributed, and executed.[69] Alternatively, a botnet can also be taken over,[70] leased or bought from a botnet owner.[71] Besides that, armed forces can persuade supporters to partake in a Distributed DOS (DDOS) attack against an adversary by providing the tools, for instance software called Low- or High-Orbit Ion Cannon,[72] and the target's IP-addresses.[73] No matter the method, when successful these attacks render a cyber object inoperable and inaccessible.[74] That may consequently result in decreased operability of the connected physical object.[75] Effects are achieved by targeting adversary cyber objects with a DOS attack. Targets could include official websites, command and control systems, logistical support systems, third-party suppliers' systems, financial services for military personnel, and connected tactical operating systems. It is important to comprehend the potential effects of a DOS attack in advance, otherwise these cyber operations may have unintended side effects of a regional, national, or international nature.

c) Intrusion and internal manipulation

Apart from denying access to cyber objects externally, a wider range of actions can be conducted from the inside. Internal manipulation requires access to a cyber object's 'back-end', hence an operator has to force entry. In order to do so, an operator can crack easy passwords using brute force techniques.[76] If unsuccessful he can also resort to social engineering techniques such as phishing.[77]

Apart from these methods, access can be forced by exploiting software vulnerabilities if an exploit is available for a specific vulnerability.[78] Well-known exploitable vulnerabilities, or

68 Andress and Winterfeld, *Cyber Warfare Techniques, Tactics and Tools for Security Practitioners*. pp. 216-217.

69 Ramneek Puri, 'Bots & Botnet: An Overview,' *SANS Institute 2003* (2003). pp. 1-2.; Nicholas Ianelli and Aaron Hackworth, 'Botnets as a Vehicle for Online Crime,' *CERT Coordination Center* 1 (2005), 15-31. pp. 16-17.

70 Ryan Vogt, John Aycock and Michael J. Jacobson Jr, 'Army of Botnets,' *Network and Distributed System Security Symposium*, no. February (2007). p. 2.

71 See for instance: Yuri Namestnikov, 'The Economics of Botnets,' *Kapersky Lab* (2009).

72 'The original LOIC Tool was built by Praetox Technologies as a stress testing application. The tool performs a simple DoS attack, by sending a sequence of TCP (Transmission Control Protocol), UDP (User Datagram Protocol) or HTTP (Hyper-Text Transfer Protocol) requests to a target host.' Source: Aiko Pras et al , *Technical Report 10.41 Attacks by Anonymous' WikiLeaks Proponents Not Anonymous* (Enschede: University of Twente, Centre for Telematics and Information Technology, [2010]).

73 Steve Mansfield-Devine, 'Anonymous: Serious Threat Or Mere Annoyance?' *Network Security January* (2011), 4-10. p. 7.

74 Pfleeger & Pfleeger, *Security in Computing*. pp. 427-433; See e.g.: Eduard Kovacs, 'DDOS Attack on DigiD Impacts 10 Million Dutch Users,' news.softpedia.com/news/DDOS-Attack-on-DigiD-Impacts-10-Million-Dutch-Users-348791.shtml (accessed October 30, 2013).

75 Such as financial traffic services and online payment services, see also: Don Eijndhoven, 'On Dutch Banking Woes and DDoS Attacks,' argentconsulting.nl/2013/04/on-dutch-banking-woes-and-ddos-attacks/ (accessed January 8, 2014).

76 Such as (THC-)Hydra and John (the Ripper). 'Hydra' and 'John' are tools enabling an attacker or pentester to automatically and systematically guess passwords (brute force) and automatically try a list of potential credentials (dictionary attack).

77 Jason Andress and Steve Winterfeld, *Cyber Warfare Techniques and Tools for Security Practitioners*, 1st ed. (Waltham: Syngress, 2011). pp. 103-105.

78 Matthijs R. Koot, Personal communication entailing comments on Dutch Article 'Militair Vermogen en Cyberoperaties' (Fighting Power and Cyber Operations), November, 2013.

'exploits', are available online either in databases[79] or enclosed in specific software.[80] Apart from applications and databases, specialised companies sell less- or unknown exploits to the highest bidder.[81] By employing brute-forcing tools, social engineering techniques, and exploits an operator can gain access to an adversary's cyber object.

Once an attacker has access to a cyber object, he can gather information inside the system and use this information to gain control over the cyber object. If the attacker successfully takes control over the cyber object, for instance a control system of an air defence turret, he can manipulate the object and subsequently operate it at his commander's bidding. Through gaining control over cyber objects, commanders can generate a variety of effects. The cyber objects could be used for future operations in the form of botnets, or used to control physical objects such as the operating systems of military platforms, or create other physical effects such as denying an area by opening a floodgate.

d) Destruction

Manipulation of cyber objects affects functions and functionality. Destroying a cyber object would result in function failure. Yet, destruction in the physical domain seems easier than in the non-physical domain. Would it, for instance, be possible to destroy or erase cyber objects? Often there are back-ups and redundant applications; erasure of cyber objects would only be complete once they are entirely removed. In most cases, it would be hard to completely erase applications and thus it would only lead to temporary failure, i.e. until back-ups are used to restore the system.

e) Human manipulation

As made clear in recent publications, content can also be used to manipulate and deceive, or in a more accepted terminology, to influence people.[82] As Greenwald demonstrates, information, true or false, may be provided as content on social media, blogs, and websites, all of which are cyber objects. Not only human perception and situational awareness may thus be affected, in addition their reputation could be challenged and, ultimately, destroyed.[83]

So?

Military and other goals can be achieved by using cyber identities and cyber objects to exert effect on other actors' cyber objects and identities. There are many other ways of using these unique features of cyberspace; we have merely scratched the surface of possible uses of cyber

79 See for example: Exploit Database, 'Windows Exploits,' exploit-db.com/platform/?p=windows (accessed March 14, 2014).; Shodan Exploits, 'Windows XP Exploits,' Shodan HQ, exploits.shodan.io/?q=windows+xp (accessed March 14, 2014).

80 See for example Metasploit, an application used for scanning, selecting exploits for the scanned system, equipping an exploit with a payload and executing it on a target system. Source: Rapid 7, 'The Attacker's Playbook: Test Your Network to Uncover Exploitable Security Gaps with Metasploit.' rapid7.com/products/metasploit/ (accessed March 14, 2014).

81 Mathew J. Schwartz, 'Blackhole Botnet Creator Buys Up Zero Day Exploits,' Information Week, informationweek.com/security/vulnerabilities-and-threats/blackhole-botnet-creator-buys-up-zero-day-exploits/d/d-id/1108075? (accessed March 14, 2014).; Andy Greenberg, 'Shopping for Zero-Days: A Price List for Hackers' Secret Software Exploits,' Forbes, forbes com/sites/andygreenberg/2012/03/23/shopping-for-zero-days-an-price-list-for-hackers-secret-software-exploits/ (accessed March 14, 2014).

82 Glenn Greenwald, 'How Covert Agents Infiltrate the Internet to Manipulate, Deceive, and Destroy Reputations', The Intercept (24 February 2014), https://firstlook.org/theintercept/2014/02/24/jtrig-manipulation/ (accessed 15 March 2014).

83 Although described in the context of disruptive effects, this method is also available for constructive purposes.

identities and objects. The wide range of possibilities and opportunities opens up cyberspace as an operating or 'warfighting'[84] domain for armed forces, States, belligerent groups, individuals, and other actors.

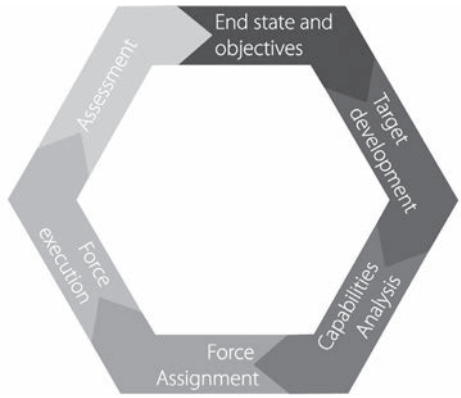
Targeting procedures have crystallised over the years and are firmly rooted in most modern armed forces. New means and methods, such as those involving cyber, pose challenges to the targeting procedures armed forces employ. In the next section we will discuss ramifications for contemporary targeting procedures as a result of the emergence of cyber operations.

6. TARGETING

Targeting in general

Military operations are executed in order to produce an effect on other actors with a view to higher strategic objectives. Actors can be influenced by applying fighting power and other instruments against an addressee or target during operations – in short, through targeting. Targeting is 'the process of selecting and prioritizing targets and matching the appropriate response to them'[85] with the purpose of determining the 'effects necessary to accomplish operational objectives; [selecting] targets that achieve those effects; and [selecting] or [tasking] the means, lethal or non-lethal, with which to take action upon those targets'.[86] A target can be 'an area, structure, object, person, organisation, mind-set, thought process, attitude or behavioural pattern'.[87] Before touching on the ramifications of cyber operations for targeting, it is necessary to briefly describe the targeting process. The targeting process is a cyclic process and consists of distinct phases (See Figure 12).[88]

FIGURE 12. TARGETING CYCLE



84 The Joint Chiefs of Staff [JCS], *The National Military Strategy of the United States of America A Strategy for Today; A Vision for Tomorrow* p. 18; The Chairman of the Joint Chiefs of Staff, *The National Military Strategy for Cyberspace Operations* p. 3.

85 British Army, *ADP Operations*. p. 5-13; JCS, *Joint Publication 3-60 Joint Targeting* (Washington, DC: JCS, 2007). p. viii.

86 Giulio Di Marzio, 'The Targeting Process: This Unknown Process (Part 1),' *NATO Rapid Deployable Corps Italy Magazine*, no. 13 (2009), 11-13. p. 13.

87 British Army, *ADP Operations*. p. 5-13.; JCS, JP3-60. p. viii.

88 Most often, six phases are recognised; See also: USAF, 'Air Force Pamphlet 14-210' fas.org/irp/doddir/usaf/afpam14-210/part01.htm (accessed January 8, 2014). Section 1.5.1.

Desired end-states and objectives provide initial input. Together with guidelines issued such as Rules of Engagement, they comprise the *first phase* of the process that is initiated in order to achieve an effect leading to the achievement of an object or end-state.

In the *second phase* targets are selected, developed and prioritised by systematically examining potential targets,[89] resulting in a target list with various potential targets that may contribute to achieving an end-state or objective.

The *third phase* entails evaluating available capabilities in order to determine options,[90] and matching the potential targets from phase two 'with [available] weapons or other capabilities to create the desired effects on the target(s)'.[91] Critically important throughout the whole targeting process, primarily in this phase, is the collateral damage estimate and assessment.[92] Weapons or capabilities may not cause collateral damage disproportionate to the military advantage anticipated.

From phase one to three, the commander may decide to execute an operation against a target, and tasking orders can be 'prepared and released to the executing components and forces',[93] weapons or capabilities can be allocated, and forces assigned to the operation in *phase four*.

Phase five, execution, follows after further mission planning and taking precautionary measures to verify information, minimise collateral damage, and issue warnings when appropriate and feasible. Phase five results in the actual operation against the target.[94]

Phase six is aimed at collecting information 'about the results of the engagement [in order] to determine whether the desired effects have been created'.[95] The output from phase six can serve as input for phase one, since after assessing effects it might prove necessary to adjust guidelines or conduct a follow-up action against the target.

The targeting process, being an operations instrument, is complemented by legal considerations derived from the law of armed conflict (LOAC). Without going into details, the questions and issues involved are: is the target a military objective, is collateral damage expected, is the collateral damage assessed to be excessive to the military advantage anticipated, is mitigation of collateral damage by 'tweaking' means and methods possible, and are precautionary measures feasible.

Targeting in cyberspace

Faced with unique cyber identities and cyber objects in the virtual or non-physical domain, the ramifications of targeting in or through cyberspace will now be addressed. Since targeting of the physical dimensions of cyberspace is well known and covered by the process just presented, we will focus on discussing targeting cyber identities and objects during cyber operations.

89 JCS, *JP3-60*. p. II-4.
90 JCS, *JP3-60*. p. II-10.
91 *Ibid*. p. II-11.
92 See Art. 52(2) AP I.
93 JCS, *JP3-60*. p. II-11.
94 *Ibid*.
95 *Ibid*. p. II-18.

1) Phase one: Effects and guidelines
Phase one of targeting cyber elements does not differ from regular targeting; cyber operations are a means to an end, just like other military operations and activities. Cyber operations are merely an addition to the commander's arsenal for generating effects, although it is evident that proper concepts, personnel, equipment, mind-set, and training are required.

Guidelines relevant to the context and conduct of cyber operations will accompany stated purposes. With an eye to the legitimacy of cyber operations they will, like other operations, be restricted for operational, political and legal reasons. It is to be expected that States, unilaterally or in coalition, will somehow express their position on the applicability and application of LOAC and human rights law to these operations. Whether or not using manuals as a point of departure, before employing cyber capabilities States will issue guidance to their troops. In addition to LOAC interpretations and positions, as in conventional operations it is commonplace to issue ROE relevant to these weapons and operations. For instance, by the use of a 'weapon release matrix' for cyber capacities, by restricting the use of cyber operations to designated digital domains or networks, or by authorising specific cyber weapons.

2) Phase two: Target development
Cyber objects and cyber identities are non-physical elements available as capabilities as well as targets or addressees. As the targeting process is designed for both lethal and non-lethal targeting, and recognises the application of soft power against the psyche of actors, it can in principle incorporate both physical and non-physical targets.

Questions arise regarding the feasibility of targeting cyber identities and cyber objects in operations and the rationale for so doing. For instance, it is fairly obvious that an adversary's cyber objects and cyber identities may be targeted subject to LOAC and ROE,[96] but can we similarly target cyber objects and cyber identities of supportive or neutral groups and individuals?

Parallels can be drawn from contemporary conflict; operations not only address adversaries, but a wide range of other actors. Apart from combating opponents through force, operations are aimed at diminishing support for adversaries by targeting the hearts and minds of the local population.[97] By supporting the local population through humanitarian aid (e.g. water, food, medical care), security (e.g. training local police, patrolling the area, combatting lawlessness) and economic aid (e.g. microcredits), an attempt is made to influence them to the benefit of the deployed force. Nowadays, the local population is increasingly online and thus would present a logical target for constructive cyber operations, as adversaries do for disruptive cyber operations.

3) Phase three: Capabilities Analysis
Phase three aims to find the right 'tools for the job'. Since cyber identities and cyber objects are connected to the physical dimension (people and objects), direct and secondary effects are achievable. Direct effects, either constructive or disruptive, are feasible through cyber

96 Noam Lubell, 'Lawful Targets in Cyber Operations - Does the Principle of Distinction Apply?', in: 89 *US Naval War College International* g (USNWC ILS) (2013), pp. 252 ff.

97 U.S. Army and U.S. Marine Corps, *Army Field Manual 3-24/ Marine Corps Warfighting Publication 3-33.5 Counterinsurgency* (Washington, DC: United States Army, 2006). p. A-5; British Army, ADP: Operations. p. 5-2; Netherlands MoD, NDD. p. 68.

operations against cyber objects and cyber identities, potentially followed by secondary effects against people and physical objects. This differs from kinetic targeting, where lethal force may destroy people or objects as the direct physical effect, and a secondary non-physical effect may occur.

Collateral damage estimation and assessment is crucial in targeting decisions. Apart from LOAC obligations, collateral damage or 'unintended effects'[98] is crucial with an eye to strategic objectives and long-term effects; for instance the perceived legitimacy of, and popular support for, operations and the military. Due to the globalised character of (social) media and increasing possibilities for 'citizen journalism',[99] and 'lawfare' to be used to discredit operations and reputation,[100] planners seek to effectively assign capabilities to targets, whilst minimising collateral damage.[101]

Thus, the collateral damage assessment of direct non-physical and secondary physical effects when targeting cyber identities and cyber objects will become increasingly important.[102] First of all, the anticipated military advantage should be assessed, and secondly the collateral damage expected should be qualified and quantified. Finally these two should be weighed, and the collateral damage must not be excessive. This three-tiered collateral damage assessment, complicated as it is in kinetic operations, will require research and training in cyberspace before it is usable at all.

4) Phases four-six

Of special interest during cyber operations is the issue of precautionary measures.[103] Care has to be taken to avoid unintended effects throughout the operation. Afterwards the effects can be assessed, and unlike regular operations, the effects of some cyber operations may be easier to quantify through other cyber operations. For example, the effects of conducting a constructive cyber operation such as influencing the perception of the local population can be assessed through monitoring the increase in positive sentiment on social media.[104]

7. CONCLUSION

We set out to operationalise military cyber operations, conceptualise their contribution, and discuss their ramifications for the targeting cycle. Having discussed the instruments of State

98 JCS, *JP3-60*. p. I-11.

99 Stuart Allen & Einar Thorsen, *Citizen Journalism Global Perspectives* (New York: Peter Lang Publishing, 2009). p. ix-xi; See e.g. compromising 'Operation Neptune Spear' (or the raid on Bin Laden) on Twitter: Melissa Bell, 'Sohaib Athar's Tweets from the Attack on Osama Bin Laden,' <washingtonpost.com/blogs/blogpost/post/sohaib-athar-tweeted-the-attack-on-osama-bin-laden--without-knowing-it/2011/05/02/AF4c9xXF_blog.html> (accessed January 9, 2014).

100 John F. Murphy, 'Cyber War and International Law: Does the International Legal Process Constitute a Threat to U.S. Vital Interests?', in: 89 *USNWC ILS* (2013), pp. 309ff.

101 Netherlands MoD, *NDD*. p. 99; NATO, AJP-1(D). p. 2-10. Section 221; British Army, *ADP Operations*. p. 3-7.

102 Schmitt, Michael N., The Law of Cyber Warfare: Quo Vadis? (September 4, 2013). 25 *Stanford Law & Policy Review*, (2014- Forthcoming), at SSRN: <http://ssrn.com/abstract=2320755>, p. 22.

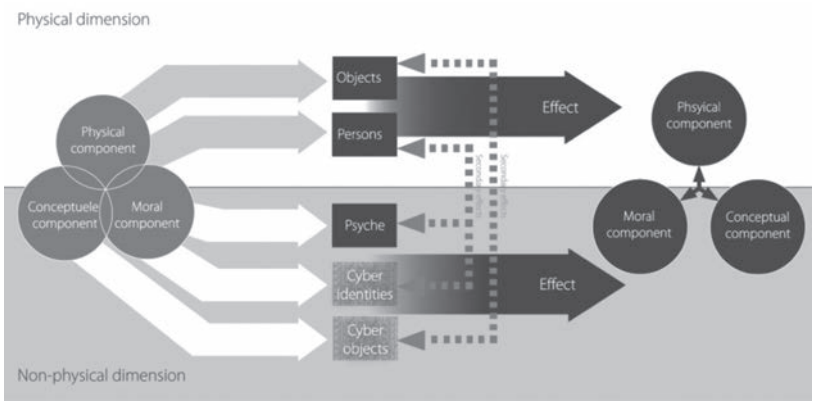
103 Schmitt, *Tallinn Manual*, p. 159ff; Eric Talbot Jensen, 'Cyber Attacks: Proportionality and Precautions in Attack', in: 89 *USNWC ILS* (2013), pp. 198 ff; Paul Walker, 'Organizing for Cyberspace Operations: Selected Issues', in: 89 *USNWC ILS* (2013), pp. 341 ff.

104 In order to do so data mining tools can be employed to collect, verify, cluster, and display the sentiment within a specific population.

power, the military instrument of fighting power is composed of various activities both military and non-military, forceful and non-forceful, and kinetic and non-kinetic. Cyber operations fit within today's concepts of fighting power, including the Manoeuvrist and Comprehensive Approaches; they are an addition to contemporary instruments. As such, cyber operations enhance capabilities for offensive and defensive purposes, including so called active defence.

Operationalisation of cyber means and methods still requires considerable effort. Whilst fighting power in cyberspace requires ordinary elements like manpower, materiel, motivation, training, concepts, and doctrine, the unique characteristics of cyberspace may pose challenges as unique non-physical elements, cyber objects and cyber identities, are present. These virtual elements not only offer new means and methods of (constructively or disruptively) influencing supportive, neutral and adversary actors, but require research and conceptualisation as well.

FIGURE 13. FIGHTING POWER AND CYBER OPERATIONS



Targeting procedures can incorporate new ways of influencing actors, since they recognise kinetic and non-kinetic targeting through physical and non-physical means, resulting in physical and non-physical effects. Assessing distinctiveness, effects and effectiveness both primary and follow-on, and collateral damage, may still prove difficult. This will require proper research, tooling and training. We conclude with an overview of the position of cyber operations in 'regular' operations (see Figure 13).

BIBLIOGRAPHY:

Allen, Stuart and Einar Thorsen. *Citizen Journalism Global Perspectives*. New York: Peter Lang Publishing, 2009.

Andress, Jason and Steve Winterfeld. *Cyber Warfare Techniques and Tools for Security Practitioners*. 1st ed. Waltham: Syngress, 2011.

Andress, Jason and Steve Winterfeld. *Cyber Warfare Techniques, Tactics and Tools for Security Practitioners*. 2nd ed. New York: Syngress, 2014.

Antoci, Angelo, Fabio Sabatini, and Mauro Sodini. 'See You on Facebook: The Effect of Social Networking on Human Interaction.' *European Research Institute on Cooperative and Social Enterprises* (2010).

Bell, Melissa. 'Sohaib Athar's Tweets from the Attack on Osama Bin Laden.', accessed January 9, 2014, washingtonpost.com/blogs/blogpost/post/sohaib-athar-tweeted-the-attack-on-osama-bin-laden--without-knowing-it/2011/05/02/AF4c9xXF_blog.html.

Betz, David J. and Tim Stevens (2011) *Cyberspace and the State*, Adelphi Series, 51:424.

British Army. *Army Doctrine Publication Operations*. Shrivenham: Development, Concepts and Doctrine Centre, 2010.

Clausewitz, Carl von. *On War, Translated and Edited by Michael Howard and Peter Paret*. Princeton: Princeton University Press, 1976.

Cornish, P., D. Livingstone, D. Clemente and C. Yorke (2010). *On Cyber Warfare*, London: Chatham House.

Corrin, Amber. 'The Other Syria Debate: Cyber Weapons.', accessed 30 October, 2013, fcw.com/articles/2013/09/04/cyber-weapons-syria.aspx.

Coyle, Cheryl L. and Heather Vaughn. 'Social Networking: Communication Revolution Or Evolution?' *Bell Labs Technical Journal* 13, no. 2 (2008): 13-17.

Di Marzio, Giulio. 'The Targeting Process: This Unknown Process (Part 1).' *NATO Rapid Deployable Corps Italy Magazine* no. 13 (2009): 11-13.

Dunlap Jr, Charles J. 'Lawfare Today: A Perspective,' *Yale Journal of International Affairs* 3 (2008), 146.

Echevarria II, Antulio J. *Clausewitz and Contemporary War*. Oxford: Oxford University Press, 2007.

Eijndhoven, Don. 'On Dutch Banking Woes and DDoS Attacks.', accessed January 8, 2014, argentconsulting.nl/2013/04/on-dutch-banking-woes-and-ddos-attacks/.

Exploit Database. 'Windows Exploits', accessed March 14, 2014, exploit-db com/platform/?p=windows.

Gertz, Bill. 'User Suspended: Twitter Blocks Multiple Accounts of Somali Al-Qaeda Group during Kenya Attack.', accessed January 8, 2014, freebeacon.com/user-suspended/.

Gibson, William. 'Burning Chrome.' *Omni* (July, 1982): 72-107.

Gibson, William. *Neuromancer*. New York: Berkley Publishing Group, 1984.

Gill, Terry D. and Paul A. L. Ducheine. 'Anticipatory Self-Defense in the Cyber Context.' *United States Naval War College International Law Studies* 89, (2013): 438-471.

Greenberg, Andy. 'Shopping for Zero-Days: A Price List for Hackers' Secret Software Exploits.' Forbes, accessed March 14, 2014, forbes.com/sites/andygreenberg/2012/03/23/shopping-for-zero-days-an-price-list-for-hackers-secret-software-exploits/.

Ianelli, Nicholas and Aaron Hackworth. 'Botnets as a Vehicle for Online Crime.' *CERT Coordination Center* 1, (2005): 15-31.

Jachtenfuchs, Markus. 'The Monopoly of Legitimate Force: Denationalization, Or Business as Usual?' *European Review* 13, no. 1 (2005): 37-52.

Janczewski, Lech J. and Andrew M. Colarik. *Cyber Warfare and Cyber Terrorism*. Hershey: Information Science Reference, 2008.

Jensen, Eric Talbot. 'Cyber Attacks: Proportionality and Precautions in Attack', in: 89 *United States Naval War College International Law Studies* (2013), p. 198.

Kovacs, Eduard. 'DDOS Attack on DigiD Impacts 10 Million Dutch Users.', accessed October 30, 2013, news.softpedia.com/news/DDOS-Attack-on-DigiD-Impacts-10-Million-Dutch-Users-348791 shtml.

Liphshiz, Cnaan. 'Israeli Vice Prime Minister's Facebook, Twitter Accounts Hacked', accessed January 8, 2014, jta.org/2012/11/21/news-opinion/israel-middle-east/israeli-vice-prime-ministers-facebook-twitter-accounts-hacked.

Namestnikov, Yuri. 'The Economics of Botnets.' *Kapersky Lab* (2009).

Nmap. 'About.', accessed March 11, 2014, nmap.org.

Noam Lubell, 'Lawful Targets in Cyber Operations - Does the Principle of Distinction Apply?', in: 89 *United States Naval War College International Law Studies* (2013), p. 252.

Mansfield-Devine, Steve. 'Anonymous: Serious Threat Or Mere Annoyance?' *Network Security* January, (2011): 4-10.

Matlack, Carol. 'Cyberwar in Ukraine Falls Far Short of Russia's Full Powers.' Bloomberg Business Week, accessed March 11, 2014, businessweek.com/articles/2014-03-10/cyberwar-in-ukraine-falls-far-short-of-russias-full-powers.

Miller, Daniel and Don Slater. *The Internet An Ethnographic Approach*. Oxford: Berg, 2000.

Ministry of Defence of the Russian Federation. 'Conceptual Views on the Activities of the Armed Forces of the Russian Federation in the Information Space,' accessed March 20, 2014, ccdcoe.org/328.html.

Murphy, John F., 'Cyber War and International Law: Does the International Legal Process Constitute a Threat to U.S. Vital Interests?', in: 89 *United States Naval War College International Law Studies* (2013), p. 309.

William Gibson No Maps for these Territories. Directed by Neale, Mark. New York: Docurama Films, 2000.

Netherlands Ministry of Defence. *The Defence Cyber Strategy*. The Hague: Netherlands Ministry of Defence, 2012.

Netherlands Ministry of Defence. *Netherlands Defence Doctrine*. Den Haag: Ministerie van Defensie, 2013.

North Atlantic Treaty Organisation. *Allied Joint Publication 1(D) Allied Joint Doctrine*. Brussels: Nato Standardization Agency, 2010.

Pfleeger, Charles and Pfleeger, Shari. *Security in Computing*. 4th ed. Boston: Pearson Education, 2006.

Plato. 'Ἀλκιβιάδης.' In *Plato with an English Translation VII*, edited by Lamb, W. London: William Heinemann Ltd., 390-342 B.C.

Pras, Aiko, Anna Sperotto, Giovane Moura, Idilio Drago, Rafael Barbosa, Ramin Sadre, Ricardo Schmidt, and Rick Hofstede. *Technical Report* 10.41: *Attacks by Anonymous' WikiLeaks Proponents Not Anonymous*. Enschede: University of Twente, Centre for Telematics and Information Technology, 2010.

Puri, Ramneek. 'Bots & Botnet: An Overview.' *SANS Institute 2003* (2003).

Rapid 7. 'The Attacker's Playbook: Test Your Network to Uncover Exploitable Security Gaps with Metasploit.', accessed March 14, 2014, rapid7 com/products/metasploit/.

Reuters. 'Ukrainian Authorities Suffer New Cyber Attacks.' Reuters, accessed March 11, 2014, reuters.com/article/2014/03/08/us-ukraine-cricis-cyberattack-idUSBREA270FU20140308.

Rheingold, Howard. *The Virtual Community Homesteading on the Electronic Frontier*. Reading: Addison-Wesley Publishing Company, 1993.

Schmitt, Michael N. (gen. ed.) *Tallinn Manual on the International Law Applicable to Cyber Warfare*. Cambridge: Cambridge University Press, 2013.

Schmitt, Michael N. 'The Law of Cyber Warfare: Quo Vadis?' (September 4, 2013). *Stanford Law & Policy Review*, Vol. 25, 2014, Forthcoming. Available at SSRN: http://ssrn.com/abstract=2320755.

Schwartz, Mathew J. 'Blackhole Botnet Creator Buys Up Zero Day Exploits.' Information Week, accessed March 14, 2014, informationweek com/security/vulnerabilities-and-threats/blackhole-botnet-creator-buys-up-zero-day-exploits/d/d-id/1108075?.

Shodan Exploits. 'Windows XP Exploits.' Shodan HQ, accessed March 14, 2014, exploits.shodan.io/?q=windows+xp.

The Chairman of the Joint Chiefs of Staff. *The National Military Strategy for Cyberspace Operations*. Washington, DC: Office of the Chairman, 2006.

The Joint Chiefs of Staff. *Joint Publication 3-60 Joint Targeting*. Washington, DC: The Joint Chiefs of Staff, 2007.

The Joint Chiefs of Staff. The National Military Strategy of the United States of America: A Strategy for Today; A Vision for Tomorrow. Washington, DC: Office of the Chairman, 2004.

The White House. Securing America's Cyberspace, National Plan for Information Systems Protection: An Invitation to a Dialogue. Washington, DC: The White House, 2000.

United States Air Force. 'Air Force Pamphlet 14-210: Intelligence Targeting Guide.', accessed January 8, 2014, fas.org/irp/doddir/usaf/afpam14-210/part01.htm.

United States Army. *Cyberspace Operations Concept Capability Plan 2016 2028*. Fort Eustis: The United States Training and Doctrine Command, 2010.

United States Army and United States Marine Corps. *Army Field Manual 3-24/ Marine Corps Warfighting Publication 3-33.5 Counterinsurgency*. Washington, DC: United States Army, 2006.

United States Department of Defense. *Department of Defense Strategy for Operating in Cyberspace*. Washington DC: United States Department of Defense, 2011.

Voetelink, J. 'Lawfare,' *Militair Rechtelijk Tijdschrift* 106, no. 3 (2013), 69-79.

Vogt, Ryan, John Aycock, and Michael J. Jacobson Jr. 'Army of Botnets.' *Network and Distributed System Security Symposium* no. February (2007).

Walker, Paul. 'Organizing for Cyberspace Operations: Selected Issues', in: 89 *United States Naval War College International Law Studies* (2013), p. 341.

Weber, Max. 'Politics as a Vocation.' Chap. Hans H. C. Wright Mills, In *From Max Weber Essays in Sociology*, edited by Gerth, Hans H. and Charles Wright Mills. London: Routledge, 1918.

Wyler, Grace. 'AP Twitter Hacked, Claims Barack Obama Injured in White House Explosions', accessed January 8, 2014, businessinsider com/ap-hacked-obama-injured-white-house-explosions-2013-4.

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Cyber Fratricide

Dr. Samuel Liles
Purdue Cyber Forensics Laboratory
Purdue University
West Lafayette, USA
sliles@purdue.edu

Jacob Kambic
Purdue Cyber Forensics Laboratory
Purdue University
West Lafayette, USA

Abstract: The United States military is currently one of the most powerful forces on the face of the planet. It achieves this in part through a high level of organization and interoperability borne through the use of the continental staffing system by the U.S. and many of its NATO allies. This system is meant to separate functions and facilitate efficient flow of information to those who need to make command decisions. While it has proven effective in prior conflicts, it has become outmoded in the information age, instead stifling necessary coordination and collaboration through isolation and insulation between roles. This paper contends that the constructs used by the continental staffing system, like that of area of operation, and rigid segregation of duty through tradition, expose a seam in the system which leads to unanticipated and negative consequences on friendly forces referred to as "cyber fratricide." Cyber Fratricide may be considered the unintentional impedance or interference between operational/tactical elements of friendly forces in the cyber realm involving the compromise or liquidation of assets, information, or capabilities of those forces. This is especially important when considering active or transactional hostilities by multiple actors. This is especially true in the case of shooting back in cyber space or active defence. By observing the most common possible forms of cyber fratricide and their enabling factors, conclusions may be drawn on possible mitigations through technical controls and reengineering of the continental staffing system to reduce cyber fratricide in active defence. This paper is a discussion of one issue in active defence and is not meant to be a complete treatise on the topic.

Keywords: *active defense, cyber fratricide, risk tolerance*

1. INTRODUCTION

The United States Military has proven itself to be one of the most capable forces on the face of the planet. It maintains this capability, in part, through a high degree of organization and specialization. One driving component of this organization is the use of the continental staff system, which enumerates functional areas of expertise. The continental staff system, used by NATO countries, assigns numbers to these areas of expertise. For instance, the intelligence officer is identified by the number 2, the operations officer by 3, and the communications officer by the number 6. The continental staff system is meant to separate functions and facilitate the

efficient flow of information to those who need it to make command decisions. Historically, the continental staff system has provided an effective method of structuring this information flow for maximum benefit.

We have known for quite some time that this organizational scheme is proving to be less effective as military operations both expand into and rely more heavily upon the cyber domain (Arquilla, 1993). Closer observation of the continental staff system reveals that its rigidity and compartmentalization, formerly benefits of that system, can, in the current information age, lead to unanticipated and negative consequences. This paper considers these consequences, and proposes that "cyber fratricide" is a real threat that needs to be addressed. Cyber fratricide is the unintentional impedance or interference between operational/tactical elements of friendly forces in the cyber realm, and can involve compromise or liquidation of assets, information, or capabilities. In what follows, the causes of cyber fratricide are discussed, examples of how cyber fratricide might occur are examined, and finally, strategies to avoid cyber fratricide are explored.

Currently, staff roles are assigned to specializations that are in likely conflict with their original purpose, which can cause strains on the aforementioned organizational structures. Additionally, achieving situational awareness requires the intelligence, operations, and communications officers to function together when dealing with cyber assets, yet, by design, several of the roles are mutually exclusive and constrained with respect to their visibility and interaction with cyber assets. In order to fully qualify these statements and explore the issues in further depth, some context is required for both the original functional roles and their typical purview (in terms of area of operation).

Each unit or military command has an area of operation. This area can be as small as a few hundred square meters at the squad level or multiple continents at the combatant commander level (the highest division of responsibility/mission in the US armed forces). The discussion that follows will focus on the battalion through combatant commander spectrum, and does so mostly interchangeably. These generalizations are crude but intentional, and the patterns being addressed here should hold up fairly well across this spectrum. Area of operation will play a significant role in one aspect of the later discussion on cyber fratricide through active defence.

Three officer positions in the continental staff system are most pertinent to analyse the issue of cyber fratricide and will now be discussed in greater detail: the intelligence officer (2), the operations officer (3), and the information communication technology (ICT) officer (6) (Joint Chiefs of Staff 1993, II-4).

Traditionally, the intelligence officer (2) has been tasked with collection and stewardship of knowledge about enemy assets (Joint Chiefs of Staff 2007b, III-14). In the cyber domain, these assets come in forms like critical infrastructures, communication nodes, components of current intelligence collection methods, and accesses created to the other (cyber) assets mentioned so far. The intelligence officer is supposed to keep the knowledge of tools, techniques, and procedures used in the intelligence collection process secret, divulging only the intelligence products dictated by the mission and circumstances that arise during its execution; however,

depending on the operational level of an intelligence officer, he or she may not actually be directly creating or implementing accesses for collection, but rather is only a consumer of intelligence themselves, engaging a party external to the mission to create/activate an access to collect/observe from at their behest. In this case, the intelligence officer may have obtained relevant operational intelligence to filter and disseminate, but have no knowledge of its provenance nor the mechanism by which it was obtained. This causes problems because the intelligence officer has caused the creation, through external mechanisms, of an access to an enemy asset for observation. Unlike other forms of access or observation, the cyber domain is transactional. This means that accesses created at the behest of the intelligence officer for his observation and action in cyberspace may allow an adversary observation and action back into the intelligence officer's organization. It is worth noting that the intelligence officer (2) will authorize or be the user of accesses created through active exploitation of information assets.

The operations officer (3) is the person who will act upon this intelligence, and operationalize the plans of the commander, ensuring the resources are ready as planned by the strategies and plans officer (5, not previously discussed)(Joint Chiefs of Staff 2011, II-1). This officer is motivated to achieve mission objectives and overcome any obstacle to the success of the mission. The operations officer will confer with his or her other staff officers when moving a plan forward to ascertain that there are no issues or concerns prior to moving past the line of departure, where the line of departure is the point at which the possibility of contact with the adversary will become material. It is imperative that this staff officer has as much information as possible upon which to build/implement the organizational strategy—coordination and collaboration are specific concerns in this capacity.

The information and communication technology officer (6) keeps communications available and manages the infrastructures required to provide the commander with command and control. This officer will coordinate which frequencies are used in a battle and how much bandwidth is available or provisioned to entities in the area of operation (Joint Chiefs of Staff 2011, D-3). In the past, the ICT officer was considered to be primarily a support role along with the logistics officer (4), but this officer is rapidly transitioning to a role as a cyber operator. Here-in lies the problem. This transition is the fulcrum upon which a series of past policy decisions start to bend towards a breaking point: asking an ICT officer, primarily trained in facilitating communication, to project power into enemy held positions exposes a fundamental flaw and observable cascading policy failure in the current implementation of the continental staff system.

When the commander wants to proceed with an operation in cyberspace he or she may want to achieve a myriad of possible goals: blind the enemy for a few moments, deny them access to an asset in a combined arms fire, create a point of societal disruption, or deny safe haven to a command and control system, as a few examples. Regardless of the request, the current process of information flow would necessitate obtaining a doctrine or planning document from the strategies and plans officer, passing it to the operations officer, who in turn would make changes and or additions to the plan, obtaining any required information that he or she can from the intelligence officer, and finally, coordinating command and control communications through the ICT officer (Joint Chiefs of Staff 2006, I-14). Despite the apparent utility and simplicity of

this information flow, which is dictated by the continental staff system and its processes, the reality is that this is not how the flow actually occurs for operations in cyberspace.

2. CYBER FRATRICIDE

Instead, in this realm, the traditional flow of communications breaks down and this breakdown can in turn lead to cyber fratricide. The cyber fratricide occurs when agents in one friendly domain negatively impact the actions of agents in another friendly domain because of the blurry boundaries inherent to cyber conflict. Several forms of cyber fratricide are possible, depending on the configuration of agents involved and their associations with one another.

These associations are more readily explained by dividing assets into different groupings. When discussing any conflict domain, assets are conventionally color-coded, with red indicating enemy assets, green indicating neutral assets, and blue indicating friendly assets. For our purposes, blue can be further separated into intelligence, operational, and domestic assets.

This division allows the identification of three forms of cyber fratricide. The first is blue operational entity on blue intelligence entity because these two entities are specifically not in close, bidirectional communication. The second is blue operational entity on green due to close association with a red information asset. Finally, a third form of cyber fratricide occurs due to ineffectual use of the area of operation paradigm and involves blue military operations acting on blue domestic assets in contravention of national laws and norms, possibly in violation of the Posse Comitatus Act (a limitation on the use of military personnel against US civilian population). These three forms of cyber fratricide are further explored in what follows.

The cyber domain is currently held to be within the purview of the communications officer (Joint Chiefs of Staff 2011, II-1). This officer's mission is primarily defensive in the context of cyber situational awareness. In order to carry out a cyber-fires mission, however, communications officers may be called upon to execute/conduct offensive activities (Computer Network Attack) that transit a "blue" network. Such a situation involves the first form of cyber fratricide— it is possible for any munitions, regardless of domain, to injure friendly troops thus creating blue on blue fratricide. In the case of the communications officer this could degrade, disrupt, or even destroy his ability to provide his primary (defensive) functional capacity. If asked to facilitate or attempt a cyber-fires mission from a blue network, the communications officer is being metaphorically asked to shoot at his foot and hopes he misses. In addition to possibly infecting, attacking, or degrading service to friendly nodes within the blue network during execution of the attack, he or she may incidentally grant a red entity access to the network or destroy blue assets in the course of his or her original defensive duties. As an example, if an officer asked to secure the network found an access that he or she did not have prior knowledge of (but was created or requested by an intelligence officer), they might reflexively apply security controls to the connection and destroy or disclose the access. In such a scenario, the ICT officer was not in direct, bidirectional communication with the intelligence officer who, following protocol, did not disclose the means used to collect the operational intelligence, or possibly was not aware of exactly how the access was created/initiated. These examples highlight the first form of cyber

fratricide by a blue operational entity on a blue intelligence entity due to the breakdown in communications and information flow spurred by the compartmentalization of the continental staff system in its current implementation.

The operations officer has yet another problem: the concept of area of operation itself is inherently flawed and outmoded in terms of a "cyber" fires mission. For example, an information asset may be accessed and leveraged by a terrorist cell in Afghanistan that is proxied through Russia by way of a Chinese Internet Service Provider with the operational asset physically located somewhere in Atlanta, Georgia. In such a case, the functional area of operation might realistically span all of the combatant commands combined. Acting on the asset would realistically be a blue operational entity acting on a blue civilian asset, currently controlled or accessed by a red operational entity transiting a green network. Further exacerbating the matter is that should the targeted red asset instead be within the locale of the red entity, Afghanistan in this case, it may still simultaneously be a subset of a green asset. That is to say that the red asset might be purchased from and managed by a third, green party that is unaware of its use for nefarious purposes or it may exist within an allied or neutral sovereignty. Considering an operation against the red entity illuminates a second form of cyber fratricide – the incidental targeting of a green entity due to its close association with a red asset.

Another consideration is that, in the current United States military paradigm, the cyber mission is inextricably linked to the intelligence function. A testament to this is the close association and collaboration between United States Cyber Command and the National Security Agency. However, the intelligence officer may only have a vantage point over (or able to develop intelligence products for) missions that are in his or her area of operation. Then consider that if an organization outside the scope of such an operation, like the U.S Cyber Command, is creating the accesses or is facilitating intelligence collection they may not, and likely should not, be communicating that activity. Additionally, if another intelligence organization is involved in the creation of access to a red asset, said organization may not even be in the target approval process of the asset for the mission's area of operation and thus unaware of intentions of the designated combatant commander. Finally, consider again the compromising position of the communications officer who, in the course of his primary (defensive) duties in these situations, is thus placed at odds with the operations officer, the intelligence officer, and his own commander when setting up a "cyber" fires mission.

3. EXAMPLES

To help illustrate these scenarios of cyber fratricide in a more concrete manner, a vignette of a mock operation utilizing cyber capabilities coupled with real world examples will now be examined. Envision that a commander wants to create a specific effect. Perhaps the commander has a mission to arrest or detain a high value red adversary within his or her area of operation. It is determined that, for a combat team to enter the area without using extensive force, a disruption of the traffic control system of a city is needed. The mission summary, then, is that blue cyber forces will disrupt, degrade, or destroy a city traffic control system. The expected effect is traffic congestion slowing response of red forces to the incursion of blue ground forces.

The planning and operations officers have evaluated several possible scenarios and outcomes of each scenario, and green-light the operation.

A kinetic attack on the traffic control system might alert red forces to a pending offensive, but a technical disruption might be interpreted by red command as incidental, and slow the realization of the true nature of the outage. In this case, since blue knows the traffic snarl will occur, blue air assets will provide reconnaissance of egress points. Blue ground assets will acquire and detain the red leader while making egress from red territory. It is expected that a small team of blue ground assets will not be detected until contact with the red leader, and that red response after realizing the nature of the attack will be constrained by the outage. Thus, a small operation will have larger strategic consequences.

A traffic control system is a real time system that uses sensory input to create a specific set of behaviours at the light-signal end. In many cities these kinds of signal computers are centrally controlled. The red asset of the traffic control lights are fully in the area of operation. Reconnaissance of these cyber assets by intelligence entities of blue confirms that the control systems themselves are fully in the area of operation. Unfortunately, the intelligence officers have not been apprised of the nature of the mission due to its classification. The intelligence officers therefore did not consider that a green entity has been outsourced to monitor traffic control systems in this area of operation. Furthermore, that green commercial entity is operating out of a control center positioned in the U.S. The outsourcing of such tasks, even between hostile adversaries, is commonplace. This is an example of the principle of globalization at work, and is the first unforeseen complication in the operation.

The next command decision is which blue cyber operators will engage in the mission within the area of responsibility. This is actually a tenuous point that should be considered carefully. In current conceptions, the entirety of cyberspace is often (mistakenly) considered to be a valid and available attack source. The question of whether the blue cyber operators should be located in the continental United States or in the area of responsibility of the commander does not have a simple answer. If the attack is launched from the United States itself, then there is no legal construct to keep the adversary from returning fire. On the other hand, if it is launched from the current area of responsibility of the commander, and then fires are directed at the United States, inadvertently/incidentally in the case of targeting the green control center, it could easily be construed as “targeting blue civilian infrastructures” and therefore be classified as a war crime. This is a thorny and convoluted legal problem.

Coordinating fires in cyber can also be a problem. Since situational awareness can be degraded by the compartmentalized command staff structure, it should come as no surprise that the operational capacity in cyber can also degraded. If the fires mission is put on the communications officer then a host of legal and policy implications ensue. In the narrative followed thus far, the concept of injecting the Department of Defense network (DISN) with a virus or cyber weapon for delivery to a civilian system is tantamount to treason—even in combat. So if the blue operator uses the cyber weapon across their own network, there are grave policy consequences looming.

The communication officer may also be providing services to the intelligence group through a coordination point or staff member. This becomes relevant when you think about the intelligence information assets that the communications officer may not even know exist. Yet it may be the intelligence officer who prepares the red information asset for exploitation and provisions separate networks for just this occasion. As such, it will likely be the intelligence officer who actually disables the red information asset. However, this is contrary to that staff officer's role and the person "pushing the button," metaphorically, should be the operations officer. Such routine deviations also point to a systemic issue in the application of tradition organizational constructs (especially the current continental staff system) to the cyber domain. This rather involved and murky example is just what creates the danger of cyber fratricide under the current concept of operations and staff structure.

In addition to the fictional example of a U.S. operation that was just presented, we can also observe documented situations abroad that underscore key elements of cyber fratricide discussed. In 2008, Pakistan engaged in what was described as an act of "information provincialism" when it decided to censor youtube.com ostensibly due to the potential of certain content to foment civil unrest (Stone, 2008). This operation however went awry and in the implementation process, Pakistan configured the externally facing BGP (Border Gateway Protocol) interface to black-hole traffic destined for youtube.com, this configuration then being propagated to the Internet at large (Stone, 2008). The result was youtube.com being "black-holed" across the world, producing an effect which accomplished their mission set but also created an international diplomatic incident.

This last portion is crucial to the incident's significance in the vein of cyber fratricide: a technical control was implemented, to effect, which also had far reaching, negative consequences throughout the organization and incidentally its allies. It also emphasizes the difficulty in controlling aspects of the area of operation within cyber from a technical perspective. Had the operation enjoyed a more tempered success and been effective only within its intended area of operation, the Pakistani nationalized network infrastructure, there were still possibly unforeseen issues. Completely screening an entire source of information and information distribution, particularly social media, sincerely degrades situational awareness. If they wanted to allow select elements within the governmental institution to monitor Youtube® at that point, they would have to create an access, which could then undermine the control put in place and complicate the operation. This control also fails largely because of the technical countermeasures not taken into account during the planning phase or evaluated during the implementation (or "Action") portion of the operation (for instance, the use of proxy hosts).

Another more direct example of cyber fratricide in the context of military operations can be found in the alleged Chinese cyber espionage campaigns described in Mandiant's "APT1" report. The premise of the report is that Chinese operatives under direct supervision of the People's Liberation Army (PLA) have been infiltrating private sector entities of other nations, notably the US, and extracting voluminous amounts of secrets/classified information. One of the reasons that these activities were detectable and directly attributable to the PLA was the separate provisioning of attack networks. While generally this is a standard practice in offensive operations, in this particular instance it was incredibly anomalous due to China's otherwise

strict control of information flow in and out of the country, sometimes colloquially referred to as "the Great Firewall of China." Because of the tight controls implemented by this censoring group, the attack infrastructure for the APT group became very apparent, and Mandiant was able to identify that "of the 614 distinct IP addresses used […] 613 (99.8%) were registered to one of four Shanghai net blocks" (Mandiant, 2013, p 4). This is an excellent example of cyber fratricide, where the activities of one blue operational unit degrades or destroys the assets or operational capacity of another blue group.

The vignette and events highlighted only scratch the surface of what is possible—as they demonstrate, the construct used for area of operation, and the information flow of the continental staff system, can have serious impacts that may lead to cyber fratricide. Additionally, other scenarios can be envisioned in which cyber fratricide could lead to a host of issues such as unintended red access to blue networks, or information exposure to red about blue assets, logistics, relationships, or personnel. Gravely, these situations could lead to the degradation or complete failure of the operation after leaving the line of departure, possibly at the cost of life to teams on the ground. This can also reverberate at scopes well beyond of the operation, affecting the entire organization.

4. CONCLUDING REMARKS AND POSSIBLE SOLUTIONS

In order to address the issue of cyber fratricide, changes to both the processes and organizational structures of the continental staff system are necessary and they concept of Area of Operations are neccesary. This can possibly be accomplished through the introduction of injection points, the use of additional technical controls, and the fixing of expectation gaps with respect to mutually exclusive objectives of specific staff positions within the continental staff system. Having a high level overview of the cyber targeting team, while knowing the specific staff issues, will allow us to engage in good situational awareness and decrease cyber fratricide.

For better information convergence during the operational planning phases, injection points can be created that are similar to those assessment points currently in place for the targeting and planning phases. In this way, a feedback loop can be established as the operation commences to improve agility and situational awareness thus reducing the possibility of cyber fratricide, particularly when also feeding in assessments from prior information operations. This can be complimented by redefining of duties for the established staff positions to meet the current need as we continue to expand operations in the cyber realm.

An ICT officer is charged by law and necessity to maintain the communications' fidelity, sanity, and resilience at all times. This officer is not a fires officer, but true to its intended purpose, is supporting an infrastructure necessary to the operations. As such, he or she simply cannot be used to project power into enemy held positions without the threat of degradation or compromise to that internal infrastructure. Therefore, a separate entity needs to carry out the offensive role in cyber. Equally, however, the operations officer cannot use their own network connections to conduct attacks. This compromises the position of the ICT officer, because, as

noted previously, the transactional nature of cyber means that doing so can create an access back into the attacking network by which the adversary may respond. Uniquely, this means that the overall effects can be detrimental beyond the scope of the fires mission: if you fly an airplane into combat from an aircraft carrier, it rarely has a significant impact on the carrier, yet in cyber, you can have issues and impact across the entire organization. There are several architectural and doctrinal changes that can help mitigate this risk of cyber fratricide and can be facilitated by technical solutions.

Architecturally, developing a command and control (C2) apparatus that is capable of taking into account the health of the C2 apparatus itself and the segregation of this apparatus from supporting technical infrastructure (with the understanding that full segregation is not truly possible) would be a vast improvement defensively. In addition to this seperation, there should also be a convergence in the support infrastructure through the implementation of coordination and decision support systems that will allow increased communication earlier on in the process between strategies and plans officers and ICT officers, and the introduction of both targeting and threat reduction tools such as the Theater Battle Management Core Systems (TBMCS) used by the US Air Force (Department of the Air Force, 2010).

Doctrinally, the current method of defining area of operation is derelict in the cyber domain. Without a mapping function that allows for holistic situational awareness and targeting of cyber assets both physically and logically, the current construct for area of operation is not only incomplete and ineffectual, it also produces a stragegic blindspot that greatly increases the risk of cyber fratricide. Cyber assets should instead be mapped dynamically in the logical space using the Internet Protocol (IP) addresses associated with Media Access Control (MAC) addresses and the physical using traditional latitude and longitude. This mapping would help prevent the engagement of blue civilian assets and improve the awareness of red assets that are actually a subset of a green entity.

If the issues with the current continental staff system and its processes are not addressed, attrition of forces, assets, and capabilities due to cyber fratricide will continue to rise in the future proportionally or possibly exponentially to the increase in cyber operations. The consequences of a single incident at the fire team level could have an impact up through the combatant command level, meaning that even a linear increase in incidents could be exponentially catastrophic to operational and tactical functions.

REFERENCES:

Arquilla, J , & David Ronfeldt. (1993). “Cyberwar is coming!” Comparative Strategy 12.2: 141-165.

Department of the Air Force. (2010). “Theater Battle Management Core System – Force level and unit level.” Retrieved From https://www.fbo.gov/index?s=opportunity&mode=form&id=f348b7a7d05fafe494f079bf519c947f&tab=core&_cview=1

Joint Chiefs of Staff. (2011). Joint Publication 3-13.1: Electronic warfare. Government Printing Office, Washington DC.

Joint Chiefs of Staff. (2007a). Joint Publication 2-0: Joint intelligence. Government Printing Office, Washington DC.

Joint Chiefs of Staff. (1993). Joint Publication 3-05.5: Joint special targeting and mission planning procedures. Government Printing Office, Washington DC.

Joint Chiefs of Staff. (2007b). Joint Publication 3-60: Joint targeting. Government Printing Office, Washington DC.

Joint Chiefs of Staff. (2006). Joint Publication 5-0: Joint operation planning. Government Printing Office, Washington DC.

Mandiant. (2013). APT1: Exposing One of China's Cyber Espionage Units. Retrieved from http://www.mandiant.com/apt1

Stone, B. (2008). Pakistan Cuts Access to YouTube Worldwide. Retrieved from http://www.nytimes.com/2008/02/26/technology/26tube.html

BIOGRAPHIES

Editors

Cpt **Pascal Brangetto** is a supply officer in the French Army. He graduated from the Military Administration Academy in 2006 and served as a 1st lieutenant at the 4th French Foreign Legion Batallion as a deputy administrative and supply officer. Then he went on to serve as an administrative and supply officer at the 1st Medical Battalion in Metz and was deployed for a tour of duty in Afghanistan during the ISAF operation in 2010. Before being posted as a legal researcher at the CCD COE in 2013, he was a staff officer in the French Joint Supply Service Directorate in Paris. Captain Brangetto is a graduate from the Institut d'Etudes Politiques in Aix-en-Provence.

Cpt **Markus Maybaum** is a German Air Force officer with more than 20 years of professional experience in the field of IT and IT security. Before his current assignment as a scientist at the NATO CCD COE's Research and Development Branch, he worked in several different national and international management, leadership and expert positions focussing on information technology, software engineering, cyber security and arms control. Besides a diploma in business administration from the German Air Force College, Markus holds a masters degree in informatics from the German Open University of Hagen specializing in IT security and he is currently pursuing a PhD in information technology with a focus on technical aspects of arms control in cyber space at Fraunhofer FKIE, Germany. Markus lives in Estonia together with his wife Simone and their four children.

Lt Col **Jan Stinissen** is a military lawyer in the Netherlands Army in the rank of Lieutenant-Colonel. He graduated from the Royal Military Academy in 1987. Lieutenant-Colonel Stinissen served as a military lawyer in different positions in The Netherlands and in Germany, one of them being a legal advisor at the Headquarters 1(German/Netherlands)Corps, but also as a policy advisor at the Directory of Personnel of the Ministry of Defence. Lieutenant-Colonel Stinissen was deployed as Legal Advisor to the Commander of the Netherlands Contingent of the Implementation Force (IFOR), Bosnia Herzegovina, and as Legal Advisor to the Commander Regional Command South, International Security Assistance Force (ISAF), Afghanistan. His current position is Researcher at the NATO CCD COE. Jan Stinissen holds a Master in Law from the University of Utrecht, The Netherlands.

Authors

Matteo Casenove is an MSc student at the Amsterdam Vrije Universiteit in Parallel and Distributed Computer Systems. He is working on his dissertation in collaboration with the NATO Cooperative Cyber Defence Centre of Excellence on "Undetectable Exfiltration Using Polymorphic Blending Techniques". He worked as intern at the NATO CCDCOE on Malware Analysis and Active Cyber Defence. Moreover, he worked as research assistant at Kent University in UK in Federated Access for Cloud. He has a strong passion and interest in everything related to Cyber Defence and Cyber Warfare.

Stephen Cobb has been researching information assurance and data privacy for more than 20 years, advising government agencies and some of the world's largest companies on information security strategy. Cobb also co-founded two successful IT security firms that were acquired by publicly-traded companies. The author of several books and hundreds of articles on information assurance, Cobb has been a Certified Information System Security Professional since 1996. He now lives and works in San Diego as part of a global research team at ESET, the Slovakia-based Internet security software company.

Robert S. Dewar is a researcher and tutor in the Politics Department at the University of Glasgow, where he is also studying for a PhD examining European cyber security from an institutionalist perspective. He holds an MA (Hons.) in Modern History from the University of St. Andrews and an MSc in Global Security (Politics, Information and Security) from the University of Glasgow. Robert has recently completed a historiography of European cyber security. This is due for publication in summer 2014 in "Challenges and Critiques of the EU Internal Security Strategy: Rights, Power and Security" edited by Dr Maria O'Neill and Mr Ken Swinton and published by Cambridge Scholars.

Judson Dressler is a PhD Candidate at Rice University with current research topics including computer and network security, cyber situational awareness, and social media's effect on military operations. He has 9 years of U.S. government experience in computer and network security operations, risk assessment and analysis of critical network infrastructures, and strategic-level security policy including positions within the National Security Agency and U.S. Cyber Command. Mr. Dressler holds an MS in computer science from the Air Force Institute of Technology and has taught as an Assistant Professor of Computer Science at the U.S. Air Force Academy.

Col **Paul Ducheine** is currently an Associate Professor for Cyber Operations at the Netherlands Defence Academy and a Legal Advisor (Netherlands Army Legal Service). From 2008-2012, he served as an Associate Professor of Military Law. He is a researcher at the Amsterdam Center for International Law. He started his military career in 1983, and joined the Engineer Regiment (as a combat engineer) in 1987. Dr. Ducheine holds a degree in Political Sciences (MSc, Amsterdam Free University) and Law (LL.M., University of Utrecht). In 2008 he defended his PhD-thesis Armed Forces, Use of Force and Counter-Terrorism. A study of the legal aspects of the role of the armed forces in combating terror (org. Dutch) at the University of Amsterdam.

Keir Giles serves as Director of Conflict Studies Research Centre (CSRC), a group of experts in Eurasian security which until 2010 formed part of the UK Defence Academy. Keir brought the CSRC team into the private sector to establish an independent consultancy, which continues to specialise in providing deep subject matter expertise to private and government customers on a broad range of security issues affecting Russia and its European neighbours and partners. Keir's specialist research areas are Russia's military transformation and Russian approaches to information and cyber security. Keir Giles is an Associate Fellow of the Royal Institute of International Affairs (Chatham House) for the Russia-Eurasia and International Security programmes.

Mario Golling is a PhD student at the Universität der Bundeswehr München (UniBwM), where he graduated in business informatics in 2007. His key aspects of research activity are network security, cyber defence, intrusion detection and next generation internet. He has many years of experience in running operational networks as well as teaching and training network administration/security. Among other things, he is a member of the Working Group IT Security of the UniBwM.

Fatih Haltas is security specialist in Cyber Security Institute, under TUBITAK (The Scientific and Technological Research Council of Turkey). He has experience of numerous penetration tests for prominent private and public organisations of Turkey. He has contributed to creation and implementation of technical scenarios of national and international cyber security exercises as TR-CERT member. Since the date, Nov 2012, He has been working as visitor research engineer in the CenteR for Interdisciplinary Studies in Security and Privacy (CRISSP), New York University in Abu Dhabi (NYUAD). His research areas include malware analysis, attacker profiling and pro-active solutions in cyber defence. Mr. Haltas's studies have been presented in noted international cyber security conferences.

Oona A. Hathaway is the Gerard C. and Bernice Latrobe Smith Professor of International Law at the Yale Law School. Her current research focuses on international law, the intersection of U.S. constitutional law and international law, the enforcement of domestic and international law, and national security law. Professor Hathaway is the founder and director of the Yale Law School Center for Global Legal Challenges, is a Professor (by courtesy) of the Yale University Department of Political Science, Professor of International Law and Area Studies at the Yale University MacMillan Center, and serves on the Executive Committee of the MacMillan Center at Yale University. She is a member of the Executive Committee of the American Society of International Law, serves as a member of the Advisory Committee on International Law for the Legal Adviser at the United States Department of State, has testified before Congress several times, and has consulted regularly with the Senate Foreign Relations Committee on current issues of constitutional and international law.

Caitríona H. Heinl is a Research Fellow responsible for research on cybersecurity under the Homeland Defence Programme at the Centre of Excellence for National Security (CENS) at the S. Rajaratnam School of International Studies (RSIS), Singapore. She is a Solicitor (non-practising) and admitted as a New York Attorney-at-Law. Prior to CENS, she was responsible for Justice and Home Affairs policy at the Institute for International and European Affairs (IIEA), Ireland. She holds a Masters (MPhil) in International Relations from the University of Cambridge.

Enrique de la Hoz received his Telecommunication Engineering degree from the University of Alcala in 2001 and Works as an Associate Professor at the University of Alcala since 2003. He has just been appointed as the University of Alcala president's delegate for information technologies. His research activities are mainly related to authentication and authorization architectures and cyber defence training for military forces. He has been involved with several taskforces promoted by TERENA on the fields of authorization and authentication in academic networks, which include a research visit to UNINETT (Norway). He is a member of the Spanish

Research Institute for Police Sciences, where he has participated, in several projects related to cyber security. In 2012 and 2013, he has participated as an external consultant for one of the Spanish Navy teams in the 3rd and 4th Cyber security exercises organised by the Spanish Army. Regarding cyber defence, he has an ongoing cooperation with the Spanish military research institute "Instituto Tecnológico La Marañosa", on the fields of cyber defence exercise design and deployment. Participation in research projects in the last 5 years: 5 competitive research projects, 8 contracts with companies.

Andreas Kornmaier, born in 1973, he joined the armed forces in 1992 and joined the University in Munich in 1995. After graduating with the Computer Science diploma, he worked in several - not exclusively technical - security related fields. He has accompanied developments and testing of command and control systems as well as looking into technical details that threaten computer systems. He has broadened his experience working in an international organization for several years in the security environment and participating in cyber exercises.

Dr. **Samuel Liles** is an associate professor in the Cyberforensics Laboratory at Purdue University in West Lafayette, Indiana. He specializes in research of transnational cyber threats and cyber forensics. His research interests are conflict, technical intelligence, forensic attribution, and systems forensics. He has held academic appointments at Purdue University in Indiana, and the National Defense University of the United States. He has served in the United States military and in municipal and county law enforcement. In the early 1990s he escaped to industry working on information technology and information assurance projects around the United States for several corporations. He worked for various major and regional telecommunications companies and consulting companies until 2003 when he joined the ranks of academia.

Eur Ing **Kevin Mepham** served with the British Royal Air Force (RAF) as an Engineer Officer; during this time he built up significant experience in safety-critical systems development and Cyber Security. After leaving the RAF he continued to build upon this experience within the commercial consultancy and international telecommunications environments, supporting this strong foundation with academic study to keep abreast of the latest developments in the field. As part of his continuing professional development in this area, he is currently undertaking a PhD in Cyber Security, looking in particular at Cyber Incident Response. Kevin is currently working as Head of an Information Assurance department within NATO.

Lt Col **David Raymond** is an Armor Officer in the U.S. Army and is currently serving as an Associate Professor in the Army Cyber Institute at West Point. He holds a Ph.D. in Computer Engineering from Virginia Tech, a Master's Degrees in Computer Science from Duke University, and a Bachelor's Degree in CS from the United States Military Academy. LTC Raymond teaches senior-level computer networking and cyber security courses at West Point and conducts research on information assurance, cyber security, and online privacy.

Cpt **Jason Rivera** is an officer in the United States Army whose primary focus is in the realms of operations and planning. He is currently pursuing a Master's Degree in Security Studies from Georgetown University's School of Foreign Service. His previous educational experience

includes a Master's Degree in Economics from the University of Oklahoma and two Bachelor's Degrees in Political Science and Economics from the University of Nevada - Las Vegas.

Jussi Timonen is a Ph.D. Student (The Finnish National Defence University) and currently working at the Finnish Defence forces C4 agency. Timonen has been a researcher with the Finnish National Defence University, Department of Military Technology from 2010 onwards. His main research areas are information fusion, common operational picture and situational awareness in critical infrastructure.

Nikos Virvilis is an information assurance scientist in the Cyber Security Division of the NATO Communications and Information Agency, in Netherlands. His research focuses on advanced persistent threat detection and mitigation. In the past, N. Virvilis has worked as an information assurance consultant/security expert for Encode S.A. and the Hellenic Armed Forces.

Bruce W. Watson is co-founder and director of the FASTAR Research Group and the Centre for Decision Making and Knowledge Dynamics, both at Stellenbosch University. In addition, he holds a visiting professorship in Computer Science at King's College London, and is chief scientist at IP Blox. In 1995, Watson received his first Ph.D in computing science from Eindhoven University of Technology, after studying discrete mathematics and computer science at the University of Waterloo. He later returned to Eindhoven as chair of Software Construction. Watson's second Ph.D, in computer science, is from the University of Pretoria in 2012. His recent book is The Correctness by Construction Approach to Programming, 2012.Parallel to his academic career, he worked as a compiler engineer at several companies (e.g. Microsoft), followed by engineering and architecture work on virtual machines and pattern matching algorithms (e.g. for Cisco).Watson's first dissertation contained taxonomies and new algorithms for pattern matching and regular expressions. With more than a dozen researchers, his FASTAR Research Group performs fundamental research in new algorithms and implementations for high-performance pattern matching. IP Blox develops leading edge technologies for deep-packet inspection, with a focus on performance and virtualization.

Matthias Wübbeling studied computer science at TU Dortmund University until 2011 and started his career in the field of electromobility at the Faculty of Mechanical Engineering. In 2012 he joined the Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE in Bonn and the University of Bonn. As an IT security researcher and doctoral student, he focuses on computer network security. Most recently, he is working on the classification of Internet routing anomalies and their consequences for the security of organizations and end users. Moreover, his teaching activities cover low level programming, applied IT security and microcontrollers such as Arduino.

CCDCOE

NATO Cooperative Cyber Defence
Centre of Excellence
Tallinn, Estonia

Contact & Feedback
publications@ccdcoe.org
CFP1426N-PRT

ISBN: 978-9949-9544-0-7

9 789949 954407

2014 6th International Conference on Cyber Conflict
P.Brangetto, M.Maybaum, J.Stinissen (Eds.)
2014 © NATO CCD COE Publications, Tallinn

Low-cost active cyber defence

Karlis Podins
University of Latvia
Riga, Latvia
karlis.podins@gmail.com

Iveta Skujina
National CERT
Riga, Latvia

Varis Teivans
National CERT
Riga, Latvia

Abstract: The authors of this paper investigated relatively simple active strategies against selected popular cyber threat vectors. When cyber attacks are analysed for their severity and occurrence, many incidents are usually classified as minor, e.g. spam or phishing. We are interested in the various types of low-end cyber incidents (as opposed to high-end state-sponsored incidents and advanced persistent threats) for two reasons:

- being the least complicated incidents, we expect to find simple active response strategies;
- being the most common incidents, fighting them will most effectively make cyberspace more secure.

We present a literature review encompassing results from academia and practitioners, and describe a previously unpublished hands-on effort to actively hinder phishing incidents. Before that, we take a look at several published definitions of active cyber defence, and identify some contradictions between them.
So far we have identified active strategies for the following cyber threats:

- Nigerian letters – keep up conversation by an artificial intelligence (AI) text analyser and generator;
- spam – traffic generation for advertised domains;
- phishing – upload of fake credentials and/or special monitored sandboxed accounts;
- information collection botnets – fake data (credit card, credentials etc.) upload.

The authors analysed the proposed strategies from the security economics point of view to determine why and how these strategies might be effective. We also discuss the legal aspects of the proposed strategies.

Keywords: *security economics, cyber crime, active cyber defence, Nigerian letters, spam, phishing, botnet*

The views and opinions presented in this paper are solely those of the authors, and should not be considered as official policy of the affiliated institutions

1. INTRODUCTION

The term ‘active cyber defence’ has been around for at least a decade (a paper by Wood *et al*. (2000) is devoted to problems around active cyber defence). The range of cyber threats can be viewed as starting from low-end (less complicated methods like spam, phishing etc.) and going to high-end (most complicated attacks e.g. Stuxnet and other state-sponsored malware). There is no clear line of division between low-end and high-end incidents; rather it is a continuous spectrum. But if we look at both ends of the spectrum there is a clear distinction between a single spam campaign and state-sponsored APT with multiple 0-day vulnerabilities. It is fairly obvious when looking at the cost estimates; an instance of spam campaign as reported by Kreibich *et al*. (2008) of 400 million emails in a three-week period, according to Goncharov (2012) would cost approx. $4000, while the Stuxnet development costs has been estimated to be around $10 million (Langner, 2010).

The authors would like to explore active cyber defence methods that are easy to implement with widely available technologies, therefore being low-cost, as the title of the paper suggests. We would like to investigate if and how such active cyber defence strategies could be applied to the occurrences of low-end cyber incidents that every netizen experiences daily.

Even though the incidents might not be technologically advanced, the sheer abundance of them causes significant losses, e.g. Rao and Reiley (2012) estimate annual spam-related costs for US companies and individuals at $20 billion, and the annual cost of advance-fee fraud to the UK economy is estimated to £150 million (Peel, 2006).

Before studying active cyber defence strategies, we look at several published definitions of active cyber defence. When comparing definitions of active cyber defence from different sources, we find some similarities and, surprisingly, some contradictions in definitions by several bodies within the US government. This paper offers both a review of published active cyber defence strategies and some novel active strategies. We have identified several active cyber defence strategies that are easy to implement and seem promising in countering some popular low-end cyber crime. As a proof of concept, we have implemented an active strategy against phishing sites and successfully tested it against two such sites.

In section 2, we list several published definitions of active cyber defence, section 3 is devoted to advance fee fraud, section 4 takes a closer look at email spam, section 5 inspects the phenomenon of phishing, section 6 deals with information stealing botnets, section 7 covers legal aspects, section 8 describes practical considerations and authors’ experiments.

2. CONCEPT OF ACTIVE CYBER DEFENCE

Before proceeding any further, we should briefly review the concept of ‘active cyber defence’ as defined by other authors. A brief literature study revealed some contradictions in the definitions published so far.

The 2011 US Department of Defense (DoD) Strategy for Operations in Cyberspace (DoD, 2011) stresses the real-time property of active cyber defence:

> 'Active cyber defense is DoD's synchronized, real-time capability to discover, detect, analyze, and mitigate threats and vulnerabilities. It builds on traditional approaches to defending DoD networks and systems, supplementing best practices with new operating concepts. It operates at network speed by using sensors, software, and intelligence to detect and stop malicious activity before it can affect DoD networks and systems. As intrusions may not always be stopped at the network boundary, DoD will continue to operate and improve upon its advanced sensors to detect, discover, map, and mitigate malicious activity on DoD networks.'

The US Defense Advanced Research Projects Agency stresses the absence of cyber offensive capabilities when describing their Active Cyber Defense program (DARPA, 2012):

> 'These new proactive capabilities would enable cyber defenders to more readily disrupt and neutralize cyber attacks as they happen. These capabilities would be solely defensive in nature; the ACD program specifically excludes research into cyber offense capabilities.'

The US Department of Defense Dictionary of Military and Associated Terms (DoD, 2010) unfortunately does not provide an explicit definition for active cyber defence, but it provides separate definitions for "active defense" and "cyberspace operations".

It defines 'active defense' as:

> 'The employment of limited offensive action and counterattacks to deny a contested area or position to the enemy.' (1)

It defines 'cyberspace operations' as

> 'The employment of cyberspace capabilities where the primary purpose is to achieve objectives in or through cyberspace.' (2)

For this paper we adopt the following definition, which can be derived by combining definitions (1) and (2):

"Employment of limited offensive **cyberspace capabilities** to deny a contested area or position to the enemy, **in or through cyberspace**."

3. ADVANCE FEE FRAUD

One of the most popular cyber crimes (or attempts at) that almost every netizen has experienced first-hand is advanced fee fraud. The victim is tricked into trusting a cyber criminal and sends some advance fee in prospect of receiving huge rewards. A very popular variation of this is

the so called Nigerian scam, when a message (usually email) is sent stating the victim has won a lottery or got an inheritance, or is asked for help to transfer money. It is also known as a '419 scam', referring to the article 419 of Nigerian Criminal Code dealing with fraud. Advance fee fraud is not a cyber-only phenomenon; in-depth research on advance fee fraud has been published, and for more thorough background and historic information see Smith *et al.* (1999). While advance fee fraud is not restricted to cyberspace, cyber offers the cheapest way of communicating with potential victims, so it is a crime that uses and abuses cyberspace.

Let's look at the mechanics of this scam:

1. Criminals use spam distribution channels to send out emails containing the 'hook' and asking victims to respond via email;
2. Some of the recipients respond;
3. In communication back and forth between criminals and victim, the victim is asked to pay an advance fee to enable the reception of a large monetary reward. The victim is asked to transfer the advance fee via an untraceable money transfer service, e.g. Western Union; and
4. The victim transfers advance fee, and the money is cashed out.

There are several possible passive strategies to fight advance fee fraud:

1. Stop spam distribution channels and improve spam filtering - this type of fraud requires cheap bulk distribution of scam messages, because only a few people are light-minded enough to fall for such scams. This strategy breaks the scam in stage 1, and is elaborated on further in the section dealing with email spam;
2. Stop anonymous and untraceable money transfer services. This would attack the scam scheme at stage 4 described above.

Both stopping anonymous or untraceable money transfer services and stopping spam distribution seem like difficult problems. There might be some other passive strategies the authors are not aware of, but to the best of our knowledge, passive strategies do not present an acceptable solution to the problem. That is why we move our attention to active strategies, attacking the scam as it progresses through stages 2 and 3.

We assume the following qualitative cost-estimate for operating advance-fee fraud scheme:

- sending out spam in huge quantities - cheap
- email discussion with potential victims - medium
- money transfer and cash-out - expensive

Sending spam is cheap; one could even say it is virtually free, as we will discuss in the email spam section below. Carrying out a discussion with potential victims over email and phone is more expensive, requiring manual labour and some proficiency in the target language. Cashing out is probably even more expensive because of the limited number of cash-outs a person can

do in a given amount of time. It seems reasonable to focus active strategies against the more expensive stages of the scam to maximize the damage inflicted on the scam operators.

We envision the following active strategies:

- Attack the email discussions between scammers and victim by identifying scam mail and using natural language processing algorithms to carry out conversations with the scammers. If done on a large scale, this dramatically increases the costs to scammers, as before they had 100% genuine human response that allowed for manual email conversation with victims. This ratio can easily be reduced almost to zero, forcing scammers to develop advanced mechanisms to identify genuine humans from computer generated responses, basically forcing them to solve the spam problem, which seems to be hard. This idea has been discussed before in some web forums (Halfbakery, 2004), but no references in an academic discussion could be found. A possible solution would be to have a 'scam button' in the email client or webmail similar to the spam button that would forward the email to a fully automated system for carrying on a conversation with the scammers. This does not require an AI algorithm to pass the Turing test. Withstanding a few rounds of conversations would be sufficient to substantially increase the costs for scam operators. Existing natural language processing algorithms would be sufficient for this task, an interesting research would be to use the famous ELIZA algorithm from 1966 (Weizenbaum, 1966) for such a purpose.
- Attack the cash-out. This needs active collaboration with the money transfer provider and would only work if cash-out is carried out by means of centralised money transfer provider such as Western Union, not with decentralised means like Bitcoins or similar. As a prerequisite, the money transfer service provider must be willing to cooperate with law enforcement. When requests for money transfers are received, those could be forwarded to the money transfer provider, which could generate marked transfer numbers. This number needs to be forwarded to the scammers. When cashing-out, the person walks up to the counter and presents the transfer number, the payment system can display a warning and he or she could be arrested by the police force if legislation supports it.

Attacking cash-out would work only for a limited number of cash-out schemes. Recently a popular scareware in Latvia asked victims to pay by purchasing PaySafeCard prepaid vouchers and sending the codes printed on the vouchers to the scammers (CERTLV, 2013). Scammers could use the code received to purchase easily resalable goods such as iTunes gift cards or electronics in an online shop. In such cases there is no physical cash-out vulnerable to attack.

4. EMAIL SPAM

Abuse of electronic messaging systems to send unsolicited bulk messages is a daily and annoying occurrence for any netizen. Stopping spam is a hard task, and industrial-grade spamming has

been a problem for almost 20 years (Cranor, 1998) and 69% of email traffic in 2012 was spam (Symantec, 2013). Although the spam ratio in global email has slightly decreased over the last few years, we cannot consider spam fighting a success story. We expect that for the foreseeable future the costs of sending spam will remain negligible, a report by Trend Micro estimates the black market price for spam distribution at $10 for a million spam emails (Goncharov, 2012).

Although spam filtering is constantly improving, some spam is always likely to get to recipients' mailboxes. The regular email user is not aware of the amount of spam filtered out by the ISPs and email providers, and the costs associated with developing and running the spam filtering software.

We would like to focus on unsolicited commercial email, i.e. email with commercial content that is sent to a recipient who has not requested it (Hoffmann, 1997) which, for example, may advertise an online shop selling fake pharmaceuticals. When taking a look at the economics of unsolicited commercial email operation, we could identify several stages:

1. Bulk email sent out;
2. Spam delivered to inbox;
3. A few users fall for the advertised product or service and purchase through the advertised website.

To maximize sales in stage 3, spammers must send out as much as possible in stage 1 (not taking into account advanced dynamic spam filtering). Taking into account the small percentage of users that fall for the advertised goods or services (less than 0.00001% according to Kanich (2008)), the amount of emails sent to become a reasonably profitable operation must be huge.

To influence stage 1, the number of infected hosts on the internet needs to be reduced – that seems to be a hard task. A lot of effort is spent in spam filtering to influence step 2 and have less spam delivered to the email inboxes. It is also possible to attack the spamming operation at step 3 with at least the following methods:

- Blacklist advertised websites;
- Community DDoS the advertised websites - if instead of deleting the spam emails at ISP/email provider/user level, requests could be generated to the advertised websites; this would create some costs that would grow proportionately to the amount of spam sent out. Currently, when email is categorized as spam, spammers do not receive any penalty, this could change just by modifying the functionality behind the spam button in your email environment, be it email client or web mail. This approach does not involve automatic detection of spam – there are lots of research towards this goal, and it seems a difficult problem. We rely on users as the final spam filter, and merely suggest that the 'spam' button in the email client would have more advanced functionality than simply deleting the email.

Spammers could probably evade blacklisting by frequently changing domain names for their web shops or even using disposable domain names uniquely created for each spam message.

Similar ideas have been implemented in production by a commercial company named Blue Security (PCWorld, 2005) but abandoned for various reasons. Such actions were quickly labelled as internet vigilantism. Taking into account that the current lack of law enforcement on the internet resembles that of the Wild West, such a term might be used without negative connotations.

5. PHISHING

Phishing is aiming at getting users to disclose sensitive information such as passwords, financial account information, or social security numbers (Ramzan, 2010). Let us look at phishing via email as a prominent example. A user might be asked to disclose some sensitive information in reply to an email, and enter it in a malicious website, or some other way. An interesting phenomenon are malicious websites which are not advertised in phishing emails but use typo squatting, i.e. use domain names which are very similar to the legitimate site, and attract users who make a typing mistake when entering the address of the legitimate website.

When phishers gather credentials and other sensitive information these can be, among others, sold on the black market, or used to access some services to steal valuables (money, stocks, in-game items etc.). The phishers must provide a way for phished users to submit the sensitive information, and this interface looks like an attractive target to attack.

We see two distinctive scenarios for such an attack:

- Flood phishing interface with fake data;
- Submit credentials corresponding to monitored/sandboxed accounts.

Flood the interface with fake data

We assume that under typical modus operandi, the 'phished' data is of very high quality. Only genuine users enter their login credentials, social security numbers, credit card details etc., so the only source of bad data is users – either typos or memory error (e.g. not remembering the correct password). Since phishing interfaces are publicly accessible, it is easy to attack them by submitting lots of fake data. The phishers are now faced with a large volume of low-quality data; this data needs to be checked, which might involve some costs depending on the type of data collected: for example, a username/password might be automatically checked for free just by logging in the legitimate service (it could be checked during the phishing phase performing some sort of man-in-the-middle scenario), while checking passport numbers might not be free.

If checking phished data involves some cost, it is possible to submit fake data to reach a threshold when phishing stops being profitable. Irrespective of checking cost, generation of fake data could be brought to such high level that either network, CPU or storage resources could be overwhelmed, and the phishing site would become inaccessible. Such idea is also proposed by Shah *et al.* (2009). As a likely response to the proposed strategy, phishing operators could introduce captcha mechanisms, just like the way operators of legitimate sites are fighting bots nowadays.

Submit credentials corresponding to monitored/sandboxed accounts

Credentials corresponding to monitored/sandboxed accounts could be submitted to the phishing interface. Once criminals use those accounts, they can be automatically tracked and some further information might be extracted, depending on the case. For example. when dealing with banking credential theft, monitored bank accounts could be used to find out ways to transfer the stolen funds. The monitored/sandboxed account would have the same look and feel as a genuine internet bank account to trick phishers in trying to transfer the funds to their associates. A popular way to transfer money from a compromised account is to use a money mule, a person who wittingly or unwittingly forwards the received money to the phisher's account in a way that is not transparent to law enforcement. We assume money mules are an expensive resource for cyber crime to acquire (possible mules need to be attracted, recruited, some cover story for the company employing the mules needs to be maintained etc.), and effectively disclosing the identities of money mules and reporting to law enforcement would be a major setback for phishers. The monitored/sandboxed accounts should be designed to withstand scrutiny by the attackers, the account should have a legitimate-looking transaction history, test transactions should be at least simulated (e.g. attackers could send or receive small amounts of money in the account as a test, before proceeding to cash-out).

Serial numbers, IMEI, MAC and other unique numbers for goods purchased could also be recorded by sellers or forwarded to banks. Once banks detect fraudulent payments, lists of unique numbers identifying stolen goods could be produced. Such lists could be used by law enforcement when inspecting grey markets. The internet could also be searched to locate the devices to better understand the geography of cyber crime.

6. INFORMATION STEALING BOTNETS

When facing botnets that search infected machines for information such as login credentials, credit card numbers etc., again a fake information submission strategy could be used. Usually drones gather the information and upload it to a dropzone either automatically or when instructed via a command & control (C&C) channel. Depending on the specifics of the C&C protocol the botnet uses, bot herders could find it impossible to distinguish between genuine information and fake information uploaded using the same bot id (this assumes the bot C&C/upload to the dropzone protocol is reverse-engineered).

If uploads are digitally signed, private keys need to be extracted from infected machines, making it much more complicated. So far the authors have not found sources stating that botnets use public key cryptography to sign uploaded data, but it is safe to assume that bot herders would implement it if such methods gained popularity.

If encountered with an unknown C&C protocol or digitally signed uploads, one voluntarily gets infected in a controlled environment. It is possible to run the malware in a sandbox/virtualised environment, supply the bot with fake data and use the original bot code to upload the fake data to the dropzone. If a large pool of diverse IP addresses is available, enough fake bots could join

a botnet and spoil the gathered data. Bot herders would need to check the data to distinguish genuine data from fake, inflicting costs on their operations.

Botnets in general have been a popular area to apply active strategies. The Conficker botnet was such a challenge that security professionals organised a Conficker Working Group to coordinate the takedown attempts (CWG, 2010), but it can be disputed whether it was a low-cost endeavour, taking into account the number of expert man-hours spent on this task. Some estimate of the economic dimension of this is the $250,000 bounty (not awarded so far) that Microsoft announced for information on persons behind the Conficker botnet (Microsoft, 2009). Some successful botnet takedown operations have had a major impact on global amount of spam, like those of Waledac and Rustock takedowns: this topic is elaborated by, among others, Dittrich (2012) and Czosseck et al. (2011).

7. LEGAL CONSIDERATIONS

This paper identifies several main issues from a legal perspective worth considering. The authors analyse the Latvian law and regulations that could be applied in proposed active response strategies. The objective is to offer a framework in which the discussions on proposed preventive measures could be evolved further.

Active response strategies can be initiated and performed by public authorities, industry, or private individuals. All of them are subject to national law and should act in the framework of national regulations. Public authorities will act only if the attack meets certain criteria set forth in law and only in accordance to procedures defined by the law. These procedures in some cases make the process slower than it is needed to prevent or even investigate the cyber threats. Public authorities can act and apply either the public and administrative or criminal law if they are notified by the victim.

Legal capacity of public authorities

The scope of administrative law in Latvia covers administrative violations which are acknowledged as unlawful actions or inactions which must be committed with intent or negligence and must endanger state or public order, property, the rights and freedoms of citizens, or management procedures specified in the law.[1] However the law does not issue regulation and does not provide liability regarding violations of computer systems or computer data. Although there is liability for unfair commercial practice, unsolicited distribution of an advertisement or commercial information, which in certain situations could be applied to spam, the limits of jurisdiction restrict the regulation to territory of Latvia and international agreements. Such cyber threats in most cases will be multijurisdictional in their nature which would hinder cooperation and effect of the law in the field of administrative violations. E.g., a citizen of State A sends spam letters. State A does not qualify spamming as violation. Recipients in six other states receive the spam letters and all of these states have different regulations regarding spam. State B regulates unsolicited e-mails, but to qualify it as violation, a certain threshold of damage must be met. The victims from State B, that have suffered damage, can make a claim to the public authorities, but in order to meet the threshold of damage the authorities need more

1 The Administrative Violations Code 1984 (Latvia), s 1

claims. As states do not have obligations to cooperate in these cases the state can not gather enough cases to act and even if it does, there is still State A that does not qualify spamming as violation and does not have internationally binding obligation to act.

Another part of public law, criminal law, does have the regulation aimed at the protection of society against cybercrime, but the criteria which must be met in order to qualify an act as the criminal offence demand the establishment of all the constituent elements of an offence set out in the law.[2] In most cases the harm done to a single individual may be comparatively small and will not qualify as criminal offence, although the nature and harm of the threat to the interests of a person or to society is substantial and, if gathered, can be subject to criminal law. Cyber threats are aimed at victims without considering the factor of territory of the state, so the limits of jurisdiction apply to harm as well. This substantially hampers the ability to identify enough victims in order to apply the criminal law.

In some cases public authorities may use our proposed measures in the investigative process to identify and to stop the source from going after other victims and causing greater damage. These actions could be used as preventive measures in order to face potentially harmful behaviour.

What can private individuals and legal persons do

The lack of regulation of cyber threats which by their nature are less harmful then criminal offences and cannot be dealt within the scope of administrative law, provide a favourable setting for the victims of those acts to seek out different defence techniques which they can employ themselves.

We should note that an active response could not only cause positive results and lessen the crime, but can also cause undesirable effects. The need to defend the network may occur when something worse comes back as revenge against the actions taken. Collateral damage may occur in the process of active response (e.g. if the spam letters are sent by competitor in the name of a company which actually is not an initiator of these unsolicited e-mails) or provoked attacks (e.g. if we use active response strategies to stop unsolicited messages we can expect that systems of our ISP can suffer from DDoS attack as well). Orin S. Kerr (2005) draws attention to several such undesirable outcomes.

There is no doubt that person can and even must reduce all possible risks in order to avoid the threats and attacks. But when the threat or the attack occurs and the harm is done, the victim has a choice either to notify state authorities or act on his own. In both cases law and regulations apply.

A former official at the National Security Agency and the Department of Homeland Security, Stewart Baker (2012), suggested that within the national legal system applicable law can have ambiguous wording and as a result the victim can argue his rights not only to protect himself, but actively engage in defence against the threat '… to conduct at least limited surveillance of a machine that is, after all, directly involved in a violation of the victim's rights'. Fred C. Stevenson Research Professor of Law Orin S. Kerr points out several undesirable effects this conclusion can cause 'As long as someone believes that they were a victim of a computer

2 The Criminal Law 1998 (Latvia), s 1 (1)

intrusion and has a good-faith belief that they can help figure out who did this or minimize the loss of the intrusion by hacking back, the hacking back is authorized' (Kerr, 2012a). This discussion points out the grey area of regulation where the victim takes risk to violate the rights of attacker. The same arguments would apply to proposed active defence methods if the use of false data would cause violation of attacker's or third party rights.

Proposed measures involve use of false information. Latvian law and regulations establishes certain conducts where provision of false information is deemed to be illegal. Firstly, if person has an obligation to provide certain information to public authorities, then provision of false information is a breach of the law. As the attacker does not represent the public authority and there is no regulation under which there is obligation to provide data to attacker, this regulation cannot be applied to the proposed situation.

The legal provisions of most criminal offences that involve the use of false information are specific to the obligation under the law to provide information. Regulation of computer systems related criminal offences on the other hand defines computer fraud as an action taken knowingly by entering false data into a computer system for the acquisition of the property of another person or the rights to such property, or the acquisition of other material benefits, in order to influence the operation of the resources.[3] Active response by which false information is provided to source of threat falls under the scope of this section. However the active response strategy does not acquire any property or the rights to such property, or other material benefits. The motive should be taken into account as well – the active response eliminates the threat. As a result active response cannot be deemed a criminal offence according to this section unless the missing elements occur.

The second well-represented view which confers on the victims a right of active defence refers to affirmative defence. It is important to note, that in this legal concept the self-defence does not deny the fact of offence - the criminal act is done, but it asserts a defence of the offender that would negate the legal effect of the offence. In the authors' view the reason why the self-defence is the last and least preferred course of reasoning by legal practitioners is the difference in the consequences of criminal procedures; in the self-defence case the offender has to admit that he committed the offence and after that has burden of proof of circumstances which exclude his criminal liability, but if the self-defence argument is not used then the offender just denies all allegations and the fact of the offence, so the burden of proof solely lies on the law enforcement.

Eugene Volokh, Gary T. Schwartz Professor of Law at UCLA, points out some common reasons why digital self-defence should be viewed without negative connotations. He writes that generally speaking the use of force is allowed '… the law has never treated defense of property as improper "vigilantism"', he continues '… the right to defend yourself and your property (subject to certain limits). By using this right, you aren't taking the law into your own hands. You're using the law that has always been in your hands' (Volokh, 2012). The opposing view is represented by Orin S. Kerr (2012b) and draws attention to the lack of precedent regarding 'cyber self-defence' as the law, at least in the US, does not have clear wording or case law that interpret the rights to defend property to be applicable to cyber defence.

3 Criminal law 1998 (Latvia), s 177.1 (1)

The criminal law of Latvia establishes several circumstances which exclude criminal liability. Self-defence is one of those admissible conditions. The criminal law provides:

> Necessary self-defence is an act which is committed in defence of the interests of the State or the public, or the rights of oneself or another person, as well as in defence of a person against assault, or threats of assault, in such a manner that harm is caused to the assailant. Criminal liability for this act applies if the limits of necessary self-defence have been exceeded.[4]

Private individuals and legal persons may use necessary means to defend their interests and rights, and in some instances the interests and rights of others, but they must act so under specific circumstances allowed by law. Necessary self-defence can be used as a defence providing that:

- The threat itself is illegal;
- The threat is actual and has already occurred;
- Actions to protect the property can be taken only to protect lawful rights and interests;
- Actions taken are proportional to the nature and the danger of threat (using reasonable force); and
- Only the source of threat suffers damage.[5]

It is important to note, that a threat must be on going and this requirement rules out any claim to use active response to prevent a potential threat. When the threat has been averted, there are no further grounds for self-defence. This shows the limits of the active response. The active response can be taken only for the time period while the threat is occurring.

Thus, it is useful to know the extent to which active response is reasonable. The Supreme Court of the Republic of Latvia explains 'defense disproportionate to the nature and the danger of the threat must be recognized as evident, if objectively there was no need for use of such means and methods to avert the threat'.[6] The question is if such cyber threats as Nigerian letters, spam, phishing, and information collection botnets is ignored by their potential victims, then is there still a need for active self-defence?

The last criterion establishes the amount of damage that can be incurred. If interpreted literally, the criterion implies that necessary self-defence is not a passive defence limited to deletion of undesirable content, but rather actions taken to cause damage to the attacker. So if a potential victim ignores the cyber threat the actions of the victim cannot be considered as necessary self-defence for several reasons: the victim's actions are legal, the actions taken are passive and the attacker has not suffered any damage. If the victim uses active defence measures, then it is important not to exceed the limits of necessary self-defence.

The proposed active defence mechanisms as a well-weighed instrument can be used by public authorities, CERTs or technicians within the scope of law. The authors have not found any court decisions on this issue, but by evaluating legal regulations it is possible to conclude that in

4 Criminal law 1998 (Latvia), s 29 (1)

5 Uldis Krastiņš, Valentīna Liholaja, Aivars Niedre, Kriminallikuma komentāri 1.grāmata Vispārīgā daļa, (1999 Riga AFS) 125

6 Case-law in application of necessary self-defence [1995] Plenary Session of the Supreme Court of Republic of Latvia 3, 1996 Latvijas Republikas Augstākās tiesas plēnuma lēmumu krājums 1990-1995

certain situations the activities may become illegal and there is a risk that the limits of necessary self-defence could be exceeded.

8. PRACTICAL CONSIDERATIONS AND PROOF OF THE CONCEPT

The authors have chosen the strategy of feeding phishing pages with fake information to implement as a proof-of-concept to test the proposed strategy in real life and demonstrate feasibility of such an approach.

The amount of fake information should be substantial in comparison with genuine phished traffic. The volume should exceed the genuine data by several times. If the aim is to overwhelm the resources of phishing site, there is no upper limit for fake data. In the case when a phishing site is suspected to be hosted on a compromised server also providing legitimate services, the fake data stream needs to be limited in order to prevent the server from crashing. A further in-depth study should be done to determine the trends of malware/phishing site hosting: whether the bad guys still use hacked servers for hosting their services, or move to use more reliable dedicated Virtual Private Servers, possibly in bulletproof hosting companies.

We have implemented the strategy as a set of python scripts and tested it on two phishing websites, both of which closed down in a reasonably short timeframe after the feed started. This is by no way a scientific proof of universal effectiveness of the proposed strategy though.

There are several points that need to be considered in order to make the generated data difficult to filter from the genuine phished data:

Content

The generated content needs to look authentic and legitimate, and that depends on the target of the phishing operation. Usernames and passwords should follow the same pattern as those of the target – usually not fully random but contain the names and surnames of the targeted country (if applicable). Passwords could be taken from popular password lists or generated from dictionaries of the target language. Credential data leaked to the public in some previous incidents could be reused in modified or unmodified form, because it does not add harm, or some algorithm to generate credible usernames and passwords could be devised. We gathered usernames, emails and passwords from public websites hosting leaked credential data. When generating 'fake' data, usernames and passwords were picked randomly from leaked lists of different incidents so the original leaked username and password pair was not reused. Domain names for email addresses might be changed to adjust to the targets of the phishing campaign (e.g. phishing campaign with Italian targets would not expect too many mail ru email accounts). Re-use of leaked credential data offers a way to generate fake data that is difficult to filter out by the phishing site operators, contrary to randomly generated strings.

Metadata

Metadata such as user-agent strings in http/https connection headers, time, time zone, and

counters in various protocol headers/footers should not be static or enable effective filtering in any other way. Our implementation randomly picks user-agent fields from pre-defined list.

Infrastructure and other considerations

Feeding the phishing sites should be done from a large pool of source IP addresses from various autonomous systems and randomized in time. Use of IP addresses should make sense; if a Latvian bank website is being phished for, most source IPs should be in the Latvian IP space but some should come from abroad, otherwise attackers can filter away Latvian source IPs to obtain genuine data, although much less in volume. In our implementation we used proxy services to randomize the source IPs. Submission of fake data should be carried out for long periods of time rather than have peaks of activity; it should be randomized in time rather than predictable. If fake data is sent in peaks or scheduled at regular intervals (e.g. first minute of every hour), phishers will have an easy time filtering out the generated submissions.

9. CONCLUSIONS

The authors have reviewed a range of active strategies that deal with some popular sorts of low-end cyber crime. All the proposed strategies can be promptly implemented with the available technical know-how and infrastructure, without need of any R&D investments. We have succeeded in providing several low-cost active cyber defence strategies, contrary to the popular belief that active cyber defence is limited to huge budget projects. As proof of concept, we implemented one of the described active strategies and most likely made the internet just a little bit better place by closing two phishing sites. We can assume that spammers, scammers and other evildoers will adapt their modus operandi once active cyber defence measures will start to exert noticeable pressure. This means the security community must constantly innovate to keep in touch in such arms-race like circumstances.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their scrutiny and valuable comments, enabling us to substantially improve the paper.

REFERENCES

Baker, Stewart , 2012, ‘RATs and Poison Part II: The Legal Case for Counterhacking’ (The Hackback Debate, November 2, 2012) http://www.steptoecyberblog.com/2012/11/02/the-hackback-debate (accessed 14 February 2014)

CERT.LV, 2013, CERT.LV brīdina par ‘Policijas’ izspiedējvīrusa izplatību., https://cert.lv/resource/show/251 (accessed 12 February 2014)

Cranor, Lorrie Faith, and Brian A. LaMacchia, 1998, ‘Spam!.’ *Communications of the ACM* 41.8 (1998): 74-83.

Conficker Working Group, 2010, ‘Conficker Working Group: Lessons Learned’, http://www.confickerworkinggroup.org/wiki/uploads/Conficker_Working_Group_Lessons_Learned_17_June_2010_final.pdf, (accessed 4 September 2014)

Christian Czosseck, Gabriel Klein, Felix Leder, 2011, On the Arms Race Around Botnets – Setting Up and Taking Down Botnets, in Proceedings of the 3rd International Conference on Cyber Conflict, Tallinn, Estonia 7-10 June 2011

DARPA, 2012, Active Cyber Defense (ACD), http://www.darpa.mil/Our_Work/I2O/Programs/Active_Cyber_Defense_%28ACD%29.aspx, (accessed 13 December 2013)

Dittrich, David, 2012, 'So you want to take over a botnet.' *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*. USENIX Association, 2012.

US Department of Defense, 2010, US Department of Defense Dictionary of Military and Associated Terms, http://www.dtic.mil/doctrine/new_pubs/jp1_02.pdf, (accessed 12 February 2014)

US Department of Defense, 2011, US Department of Defense, Strategy for Operations in Cyberspace *www.defense.gov/news/d20110714cyber.pdf (accessed 12 February 2014)*

Goncharov, Max, 2012, 'Russian underground 101.' *Trend Micro Incorporated Research Paper* (2012).

Halfbakery, 2004, Advance fee fraud (419) reply-bot, http://www halfbakery.com/idea/Advance_20fee_20fraud_20(419)_20reply-bot (accessed 13 December 2013)

Paul Hoffmann, 1997, Unsolicited Commercial Email: Definitions and Problems, Internet Mail Consortium report, http://www.imc.org/imcr-002 html (accessed 13 December 2013)

Kanich, Chris, et al., 2008, 'Spamalytics: An empirical analysis of spam marketing conversion.' *Proceedings of the 15th ACM conference on Computer and communications security*. ACM, 2008.

Kerr, Orin S., 2005, 'Virtual Crime, Virtual Deterrence: A Skeptical View of Self-Help, Architecture, and Civil Liability'. *Journal of Law, Economics & Policy*, Vol. 1, January http://papers.ssrn.com/sol3/papers.cfm?abstract_id=605964 (accessed 24 February 2014).

Kerr, Orin S., 2012a, 'The Legal Case Against Hack-Back: A Response to Stewart Baker' (The Hackback Debate, November 2, 2012) http://www.steptoecyberblog.com/2012/11/02/the-hackback-debate (accessed 14 February 2014).

Kerr, Orin S. 2012b 'A Response to Eugene Volokh' (The Hackback Debate, November 2, 2012) http://www.steptoecyberblog.com/2012/11/02/the-hackback-debate (accessed 14 February 2014).

Kreibich, Christian, *et al*., 2008 'On the Spam Campaign Trail.' LEET 8 (2008): 1-9.

Langner, Ralph, 2010, 'The short path from cyber missiles to dirty digital bombs', http://www.langner.com/en/2010/12/26/the-short-path-from-cyber-missiles-to-dirty-digital-bombs/ (accessed on 12 February 2014).

Microsoft, 2009, Microsoft Collaborates With Industry to Disrupt Conficker Worm, http://www.microsoft com/en-us/news/press/2009/feb09/02-12confickerpr.aspx, (accessed 12 February 2014).

PCWorld, 2005, Spam Slayer: Bringing Spammers to Their Knees, PCWorld, http://www.pcworld.com/article/121841/article.html, (accessed 13 December 2013).

Peel, Michael, 2006, *Nigeria-related financial crime and its links with Britain*. London: Chatham House

Ramzan, Zulfikar, 2010, 'Phishing Attacks and Countermeasures.' *Handbook of Information and Communication Security*. Springer Berlin Heidelberg, 433-448.

Rao, Justin M., and David H. Reiley, 2012, 'The economics of spam.' *The Journal of Economic Perspectives*, 87-110.

Shah, Ripan, et al., 2009, 'A proactive approach to preventing phishing attacks using Pshark.' *Information Technology New Generations, 2009. ITNG'09. Sixth International Conference on*. IEEE

Smith, Russell G., Michael N. Holmes, and Philip Kaufmann, 1999, '*Nigerian advance fee fraud.*', in Trends & Issues in Crime and Criminal Justice, July 1999.

Symantec, 2013, Symantec, Internet Security Threat Report 2013, http://www.symantec com/content/en/us/enterprise/other_resources/b-istr_main_report_v18_2012_21291018.en-us.pdf, (accessed 13 December 2013).

Volokh,, Eugene, 2012, 'The Rhetoric of Opposition to Self-Help' (The Hackback Debate, November 2, 2012), http://www steptoecyberblog.com/2012/11/02/the-hackback-debate (accessed 14 February 2014).

Weizenbaum, Joseph, 1966, 'ELIZA—a computer program for the study of natural language communication between man and machine.' *Communications of the ACM* 9.1 (1966): 36-45.

Wood, Bradley J., Saydjari, O. Sami and Stavridou Victoria, 2000, 'A proactive holistic approach to strategic cyber defense.' *SRI International. http //www. cyberdefenseagency. com/publications/Cyberwar_Strategy_and_Tacti cs. pdf* (accessed 12 February 2014).