IEEE
Advancing Technology for Humanity

2013 5th International Conference on Cyber Conflict

PROCEEDINGS

K. Podins, J. Stinissen, M. Maybaum (Eds.)

CCDCOE

NATO Cooperative Cyber Defence Centre of Excellence
Tallinn, Estonia

CyCon

International Conference on Cyber Conflict

TALLINN, ESTONIA

2013
5th International Conference on Cyber Conflict

PROCEEDINGS

K. Podins, J. Stinissen, M. Maybaum (Eds.)

4-7 JUNE 2013, TALLINN, ESTONIA

CCDCOE
NATO Cooperative Cyber Defence Centre of Excellence
Tallinn Estonia

IEEE
Advancing Technology for Humanity

2013 5TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT (CYCON 2013)

IEEE Catalog Number: CFP1326N-PRT
ISBN 13 (print): 978-9949-9211-4-0
ISBN 13 (pdf): 978-9949-9211-5-7
ISBN 13 (epub): 978-9949-9211-6-4

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: Marko Söönurm

CYCON 2013 SPONSORS

TECHNICAL SPONSOR

DIAMOND SPONSOR

GOLD SPONSORS

SPONSOR

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". Located in Tallinn, Estonia, the Centre is currently supported by Estonia, Germany, Hungary, Italy, Latvia, Lithuania, the Netherlands, Poland, Slovakia, Spain, and the USA as Sponsoring Nations. 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.

For information on Centres of Excellence, visit NATO's website "Centres of Excellence" at http://www.nato.int/cps/en/natolive/topics_68372.htm.

Foreword

Cyber conflicts are increasingly moving towards an operational theatre, where physical and cyber space are becoming hardly distinguishable, where the surrounding cyber ecosystem is global, dynamic and involves tens of thousands networks, and where the sheer volume of cyber security incidents data that needs to be collected, analyzed and acted upon is staggering. In this ecosystem we are talking about monitoring tens of thousands of networks that connect hundreds of thousands, if not millions of devices. Consequently, it is becoming evident that we have to move from passive perimeter-bound cyber defense, where we have spent tremendous amounts of effort, time and money in protecting every single item along the perimeter of the cyber ecosystem, towards active cyber defense and offensive actions that secure resiliency of our cyber-kinetic operations, even in the presence of adversary attacks. Not less importantly, the paradigm of perimeter cyber defense leaves our cyber-kinetic operations vulnerable to insider attacks. The models of achieving resiliency of cyber-kinetic operations are still at active research and development stage, but several of them have proven their value, including self-organization and adaptation to evolving cyber security situations, prediction of potential adversary cyber attacks and other disruptive events before their occur, and recovery and restoration of operational capacities after the attacks. Whatever the future technology of resilient cyber-kinetic operations will emerge, it is undisputable fact that the center piece of this technology is automation of cyber security data analysis and decision-making processes that underlay defensive and offensive actions in cyber conflict.

We look on automation not only as an enabling technological device to proceed with cyber conflict operations, but also extend its associated impact on the behavioral entities acting in strategic and legal space. In order to have a better chance of securing our mission critical information systems (and critical information infrastructure in general), we must make defenses smarter and more autonomous. However, while reducing the role of the human in the loop may provide benefits like quicker reaction time, better anomaly detection and a potential shift in strategic balance of power, it will also lead to new security challenges. Many of these challenges have a profound impact on the ethical, moral and legal aspects of cyber security, which can only be addressed in open-minded and multi-disciplinary discussions.

Research and development of different automated procedures of supporting defensive and offensive cyber security operations have been conducted over a number of years, but mostly as independent isolated efforts, e.g. automatic alarm correlation procedures exploited in different intrusion detection systems, automatic

vulnerability scanners, data mining techniques for generating patterns of new cyber attacks, expert systems advising IT personnel how to recover networks and systems after the impact of the cyber attack, machine learning and genetic algorithms to model the propagation of computer viruses and botnets, and others. The mission and vision of this conference is to look on automation of cyber conflict operations from synergistic multi-disciplinary perspective. The conference intention was to underscore the role of automation not just as enabling cyber security technology, but as a critical factor, which makes the future cyber defense and offense possible, and what kind of legal, social and moral implications we have to be concerned. In this context the annual Cyber Conflict (CyCon) conferences conducted yearly in Tallinn by the NATO Cooperative Cyber Defense Centre of Excellence are continuing to provide their unique perspective. This distinctiveness is marked by an innovative synergistic approach to the conceptual framework, architectures, processes and systems of cyber security and conflict. It holistically examines computer science and IT technologies, law, strategic and policy matters, military doctrine, social and economic concerns and human behavioral modeling with respect to cyber space.

The proceedings of this 5th International Conference on Cyber Conflict 2013 (CyCon 2013) are collected in this volume. The 26 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 conflict, specifically focusing on the issues of automation. They include sophisticated analyses of topics like offensive and defensive cyber activities, the concept of the cyber space, its legal and technical boundaries, and the fundamental notions of cyber attacks, cyber attackers, cyber conflict, and cyber warfare.

This volume is arranged into five chapters. The first chapter, Cyber Space – Automatic Information Sharing and Access, discusses the models of organizing cyber security information into a ubiquitous smart space, where information can be shared by cooperative cyber defense parties, can be automatically aggregated, and can be accessed depending on the operational context. The second chapter, Attack Modeling – Washing Away the Borders between Cyber and Kinetic Attack, discusses important issues of modeling cyber attacks, looks on limits of automatically generated cyber attacks, and explores new paradigms of cyber attacks, which directly impact entities and processes in the kinetic (physical) world. The third chapter, Cyber Attack Threat Assessment and Impact propagation is devoted to a wide spectrum of technical and legal issues associated with the analysis of the threat and impact of cyber attacks. The fourth chapter, Cyber Command – Towards Automatic Operations, collects number of papers that examine automatic procedures of tactical cyberspace operations, algorithms of detecting complex cyber attacks by automatic correlations of intrusion alerts from multiple sources, and novel architectures of building autonomic (automatic and autonomous) cyber

security decision-making processes. The final chapter, Cyber Conflict - Politics, Semantics, Ethics and Moral, analyses the semantics of concepts related to the cyber conflict across different languages, argues in favor of different models that relate cyber conflicts to moral, ethics and politics

We would like to thank the members of both the CyCon 2013 technical program committee and the distinguished peer reviewers for their tireless works 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 – Karlis Podins, Markus Maybaum and Jan Stinissen. Without their technical expertise, professional attitude, and personal dedication, this work would not have been possible.

5th International Conference on Cyber Conflict 2013
Programme Committee Co-Chairs

Dr Gabriel Jakobson
Chief Scientist, Altusys Corp

Dr Rain Ottis
Associate Professor
Tallinn University of Technology

Brookline, Tallinn, April 2013

Contents

Chapter 4. Cyber Command – Towards Automatic Operations

Chapter 5. Cyber Conflict – Politics, Semantics, Ethics and Moral

Introduction

For the fifth 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 *5th International Conference on Cyber Conflict* (CyCon 2013) brought together national security thinkers, strategists, political scientists, policy-makers, lawyers and technology experts interested in cyber defence, and served as a hub for knowledge and networking on an international level.

CyCon 2013 focused on automated methods in cyber conflict. Reflecting the interdisciplinary approach of NATO CCD COE, this topic was explored from strategic, conceptual, political, legal and technical perspectives within two parallel tracks. The *Strategic Track* was co-chaired by *Jan Stinissen* (NATO CCD COE) and *Dr Rain Ottis* (Tallinn University of Technology) while the *Technical Track* was co-chaired by *Markus Maybaum* (NATO CCD COE) and *Dr Gabriel Jakobson* (Altusys Corp.). Additional pre-conference workshops (organised by the International Society for Military Law and Law of War and NATO CCD COE) warmed up the venue.

The *Strategy Track* addressed policy, strategy and military doctrine on the use of automated systems in cyber conflict, including the legal and ethical aspects related to the use of such systems.

The policy-oriented presentations covered the state and industry perspective on collective and automated cyber defence, as well as the more offensive options and limitations that such automatic systems could bring along. On the military side, the concepts and risks of automated offense and defence were discussed together with the need for a commander to have an adequate cyber 'toolbox' at his disposal, possibly including a variation of automated systems.

Following the keynote introduction on the use of autonomous weapon systems and international law, legal aspects related to the use of an autonomous cyber immune system were examined, as well as the legal position of a commander using autonomous decision-making processes. Other specific issues that were addressed included cyber deception, autonomous attacks and the legal implications. The presentations on ethical aspects covered moral responsibility for the use of hybrid systems in cyber warfare, an ethical analysis of the case for robotic weapons arms control, and the risk of possible moral deskilling of the military by the increasing use of autonomous systems.

Two panel sessions were held to further encourage the debate on different policy,

strategy, legal and ethical aspects evolving around the use of automated methods. The Strategy Track was closed on the last day of the conference with a presentation reflecting on the topics raised during the event and offering an outlook for the possible future of using automated systems in cyber conflict.

The Strategy Track also included a few presentations on other cyber topics, not so closely related to the main conference theme such as the expected future use of kinetic cyber, and an analysis on the use of cyber attack as a foreign policy tool. In addition, a comparison of the use of cyber definitions in Chinese, Russian and English, and a proposal for a cyber conflict taxonomy were presented.

The *Technical Track* focused on technical aspects of automatic methods within the scope of cyber security. It provided different perspectives of automation within the scope of tactical, operational and strategic procedures in cyber conflicts, cyber conflict models and impact assessment as well as other related cyber topics related to our conference focus.

As a first highlight aspects of automatic information sharing and access control in cyber space were discussed. Besides an introduction of new information sharing models special emphasis was given to behaviour primitives and context based access control. Then the challenges of disappearing borders between cyber and kinetic attacks were addressed as well as aspects of cyber-attack threat assessment and impact propagation. Again, new frameworks were proposed and examples for a practical implementation as well as case studies were presented. Finally new approaches towards automatic operations were introduced and explained at a very detailed level.

In addition to the papers printed in this book, two panel sessions were held to discuss the technical aspects of the on-going automation-driven paradigm shift in cyber defence and the latest developments within the scope of automation in intrusion detection as well as network and malware analysis.

The Joint Sessions covered the field from highest political level down to technical attack method analysis, giving insight from government, military, law and industry point-of-views.

The editors have structured the proceedings so that chapters have common topic and correspond to conference sessions as much as possible. We believe the readers will find this approach useful as semantically close papers will be follow each other.

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 papers submitted pursuant to the call for papers, and also for the peer review of the papers submitted by invited authors, guaranteeing the academic quality of the selected papers.

Programme Committee Co-Chairs were (in alphabetic order):

- Dr *Gabriel Jakobson*, Chief Scientist, Altusys Corporation
- Cpt *Markus Maybaum*, NATO CCD COE
- Dr *Rain Ottis*, Tallinn University of Technology
- LtCol *Jan Stinissen*, NATO CCD COE

Members of the Programme Committee were (in alphabetic order):

- Dr *Iosif I. Androulidakis*, University of Ioannina
- Prof Dr *Marta Beltran*, Rey Juan Carlos University
- Dr *Steve Chan*, MIT-IBM Network Science Research Center
- Prof *Thomas Chen*, Swansea University
- Dr *Christian Czosseck*, CERT Bundeswehr
- Prof *Dipankar Dasgupta*, The University of Memphis
- Prof *Dorothy E. Denning*, Naval Postgraduate School
- Colonel Dr *Paul Ducheine*, Netherlands Defence Academy/University of Amsterdam
- Dr *Kenneth Geers*
- Prof Dr *Terry Gill*, University of Amsterdam, University of Utrecht, Netherlands Defence Academy
- Prof Dr *Michael R. Grimaila*, Air Force Institute of Technology
- Dr *Jonas Hallberg*, Swedish Defence Research Agency
- Prof *David Hutchison*, Lancaster 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
- Dr *Scott Lathrop*
- Dr *Sean Lawson*, University of Utah
- Dr *Corrado Leita*, Symantec Research Labs
- Dr *Samuel Liles*, Purdue University

- *Eric Luiijf*, MSc, TNO
- Dr *William Mahoney*, University of Nebraska at Omaha
- Prof Dr *Michael Meier*, University of Bonn
- Dr *Jose Nazario*, Invincea Inc.
- *Lars Nicander*, Center for Asymmetric Threat Studies at the Swedish National Defence College
- Prof Dr *Gabi Dreo Rodosek*, Universität der Bundeswehr München
- Prof Dr *Julie J.C.H. Ryan*, George Washington University
- Prof *Alexander Smirnov*, St.Petersburg institute for Informatics and Automation of the Russian Academy of Sciences
- Dr *Pontus Svenson*, Swedish Defence Research Agency
- *Anna-Maria Talihärm*, NATO CCD COE
- Dr *Jens Tölle*, Fraunhofer FKIE
- Dr *Risto Vaarandi*, NATO CCD COE
- Colonel Dr *Joop Voetelink*, Netherlands Defence Academy
- Dr *Jozef Vyskoc*, VaF Rovinka and Comenius University Bratislava
- Prof *Stefano Zanero*, Politecnico di Milano
- Dr *Katharina Ziolkowski*, NATO CCD COE

Special gratitude is 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's Estonia Section served as technical co-sponsor of CyCon 2013 and of these Conference Proceedings, numerous IEEE members have supported the Program Committee ensuring the academic quality of the papers and supporting their electronic publication and distribution.

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

Karlis Podins, Jan Stinissen, Markus Maybaum
NATO Cooperative Cyber Defence Centre of Excellence

Tallinn, Estonia
June 2013

Chapter 1.

Cyber Space – Automatic Information Sharing and Access

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Towards Improved Cyber Security Information Sharing

Requirements for a Cyber Security Data Exchange and Collaboration Infrastructure (CDXI)

Luc Dandurand
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

Abstract: There is a requirement for improved information sharing and automation in the cyber security domain. Current practices and supporting technologies limit the ability of organizations to take full advantage of their staff's expertise and the trust relationships they have established with each other in their efforts to secure their communication and information systems. Limitations include the lack of interoperable standards, the absence of mechanisms to govern and control the use of sensitive information, and problems validating data quality. While centralized repositories, distribution lists and web services have been adopted in an attempt to address the requirement, the underlying needs are only partly met by these approaches, which do not deliver the required efficiency and effectiveness.

Analysis of the specific constraints applicable in the cyber security domain led to definition of the Cyber Security Data Exchange and Collaboration Infrastructure (CDXI) capability. CDXI provides a knowledge management tool for the cyber security domain whose objectives are to facilitate information sharing, enable automation, and facilitate the generation, refinement and vetting of data through burden-sharing collaboration or outsourcing. The capability is defined through a set of high-level requirements that are both necessary and sufficient. This paper describes the high-level requirements and provides a brief description of the work performed to develop the CDXI concept to date as well as planned future work.

Keywords: *Cyber security, knowledge management, data sharing, collaboration, automation*

This work was sponsored by NATO's Allied Command Transformation under the 2012 Cyber Defence Programme of Work. This document is a working paper that may not be cited as representing formally approved NCIA, ACT or NATO opinions, conclusions or recommendations, and represents the views of only the authors.

1. INTRODUCTION

Knowledge management is commonly used as an umbrella term that covers the generation, representation, storage, transfer, transformation, application, embedding, and protecting of an organization's information ([1], [2], [3]). Knowledge management has become increasingly important to various communities as the amount of information being produced has been growing exponentially in the last decades, and timely information exchange has become essential if not critical in a broad range of domains.

In the cyber security community, there is currently a strong need for the exchange of data to support the management of vulnerabilities, threats and incidents, as well as other cyber security activities. The exchanges are necessary to achieve common goals in federated environments and to exploit collaboration opportunities. Furthermore, given the speed at which cyber-attacks unfold, there is also a need to support timely decision-making and automate responses to the greatest extent possible. These two goals can be achieved only if structured and quality-assured data is available for automated processing.

Having recognized these issues in the cyber security domain, NATO's Allied Command Transformation (ACT) sponsored the NATO Communications and Information Agency to develop the concept for a Cyber Security Data Exchange and Collaboration Infrastructure (CDXI), whose objectives are to:

- Facilitate information sharing
- Enable automation
- Facilitate the generation, refinement and vetting of data through burden-sharing collaboration or outsourcing.

As part of the development of the CDXI concept, high-level requirements that must be met to achieve the above objectives in the cyber security domain have been identified. The high-level requirements, which define the capability needed by the Alliance to manage cyber security information, are described and justified in this paper.

The remainder of the paper is structured as follows. Section 2 introduces the problem associated with information sharing and automation in cyber security, and the current state of affairs. Section 3 lists and describes the high-level requirements identified. Section 4 introduces an illustrative high-level architecture, and Section 5 presents conclusions and outlines future work that is planned or recommended.

2. BACKGROUND

The INFOSEC "Hard Problems List", under the heading "Information Provenance", identifies assuring the quality of shared data by tracking its evolution as one of the most fundamental problems in information security [4]. It can be argued that the difficulty stems from the loss of metadata that occurs when information is exchanged over systems that favour general availability and re-use over integrity, quality assurance and traceability. The problem is not exclusive to the cyber security community; areas as diverse as medicine [5], genetics [6] and law enforcement [7] are also affected by this issue.

The use of ontologies for knowledge-sharing activities has long been an important research topic ([8], [9], [10]). The importance of mapping overlapping ontologies has also been highlighted [11], and research has been conducted in the area of distributed knowledge management ([12], [13]). However, cyber security organizations have traditionally addressed information-sharing using *ad hoc* solutions such as email exchange, web-based collaboration tools such as portals and wikis, shared databases, and automated feeds of data.

In the last few years, a number of standards and initiatives that facilitate cyber security information exchange have been developed and they are gaining acceptance. ENISA (European Network and Information Security Agency) is trying to support its member states by deploying the European Information Sharing and Alert System (EISAS) [14], while the MITRE Corporation has developed a number of standardized enumeration structures and languages: Common Vulnerabilities and Exposures (CVE), Common Platform Enumeration (CPE), Common Configuration Enumeration (CCE), Common Attack Pattern Enumeration and Classification (CAPEC), and the Open Vulnerability and Assessment Language (OVAL), amongst others [15]. Industry adoption of these standards as well as other relevant standards appears to be progressing well.

The 2011 X.1500 CYBEX (Cyber Security Information Exchange Framework) Recommendation of the ITU's Study Group 17 "*describes* **techniques** *for exchanging cyber security information*" [16]. The ITU-T's X.15xx series of standards includes many of the standards and techniques developed by the U.S. National Institute of Standards and Technology under the Security Content Automation Protocol (SCAP) initiative, the MITRE enumeration structures and standards previously mentioned, and standards and techniques for the actual exchange of data, for establishing trust and policy agreement between parties, and for assuring the integrity of exchanges. Finally, a number of standards produced within the Internet Engineering Task Force (IETF) are aimed at facilitating cyber security information exchange, e.g. Real-time Inter-network Defense (RID) under RFC 6545 [17] and RFC 6546 [18] and the

Incident Object Description Exchange Format (IODEF), RFC 5070 [19].

Commercial products are just beginning to incorporate the previously mentioned standardization efforts and their supporting technologies. In a 2008 review of existing security ontologies, it was argued that *"existing ontologies are not prepared for being reused and extended and the security community still needs a complete security ontology that solves these lacks and provides reusability, communication and knowledge sharing"* [20]. While a single, complete security ontology may be an unreachable goal, it is possible to make a set of ontologies interoperable, covering all aspects of security, and this would address the requirements. More recently, subject-matter experts from the RSA organization stated that,

"Data standards for describing and transmitting threat information have advanced significantly, but much progress is needed to extend existing standards and drive wider adoption in vendor solutions. [...] Threat information-sharing and collaboration programs help organizations augment their expertise and capabilities in detecting and remediating advanced threats, but most sharing programs are hindered by a heavy reliance on manually intensive, non-scalable processes and workflows." [21].

While the development of interoperable ontologies is progressing well, a number of major challenges remain with respect to achieving effective and efficient exchange of data and automation in the cyber security domain:

- There are no mechanisms available to automate large-scale information sharing.
- Many different sources of data containing inconsistent and in some cases erroneous data exist.
- It is difficult, in some cases, to access the desired information from the large volumes of data stored on the Internet or embedded in specific products (e.g. vulnerability repositories, signatures for anti-virus products, etc.).
- Many protocols and access mechanisms are proprietary or not interoperable.
- Incompatible semantics using the same or similar words are used in different data sources covering the same topics.
- The quality of data varies and information and assurance regarding the level of quality provided is lacking.
- There is very limited support for efficient collaboration, despite the availability of subject-matter experts in a large number of organizations willing to collaborate.

- Concerns regarding the confidentiality of exchanged data in the absence of means by which redistribution can be satisfactorily controlled must be addressed.

CDXI is designed to address these challenges by providing an enterprise-level capability that facilitates information sharing, enables automation, and facilitates the generation, refinement and vetting of data through burden-sharing collaboration or outsourcing.

3. CDXI HIGH-LEVEL REQUIREMENTS

To define the capability needed to meet the objectives stated in Section 1, the problems associated with information sharing and automation in the cyber security domain were examined. As a result the challenges listed in Section 2 as well as a number of key considerations applicable to that domain were identified, which in turn led to the identification of eleven high-level requirements that the CDXI capability must meet in order to achieve its objectives. These high-level requirements are considered to be both necessary and sufficient, and are described below.

A. PROVIDE AN ADAPTABLE, SCALABLE, SECURE AND DECENTRALIZED INFRASTRUCTURE BASED ON A FREELY AVAILABLE CORE

Collecting data from a heterogeneous set of data sources, sharing some of it with partners, and supporting automated cyber security operations while exploiting collaboration and outsourcing opportunities is a daunting challenge. While many organizations have established trust relationships with each other, few are able to agree on a single system that fits every organization's specific requirements. Adaptability is therefore required so that organizations of different sizes, different types, facing different constraints and seeking different objectives can deploy CDXI in a way that meets their specific situation. The organizations that CDXI must support range from a very small, single-site company to a large multinational federated organization. In many cases, the need to exchange information will be the only common point, and mandating a fixed configuration will lead to an ineffective and inefficient solution, if not outright failure.

CDXI must be scalable, not so much for reasons of data quantity, which remains quite modest in cyber security, but rather because an "agile data model" and correlation capabilities are necessary (see requirement B), as is the need to support dissension (see requirement I). These two requirements are expected to increase the need for storage capacity. As well, CDXI components must be scalable to meet

a wide range of hosting constraints and performance requirements in different deployment scenarios.

Because the increased need to share does not diminish the confidentiality, availability, and integrity requirements of the exchanged data, CDXI must also be secure. Therefore CDXI must provide flexible access controls to allow protection of the data as well as the possibility for custom workflows that will enable multi-step approval for actions affecting sensitive data. In order to allow greater exploitation of shared data while maintaining privacy requirements, CDXI must allow organizations to identify data elements that must be consistently replaced by privacy-protecting labels before being shared, as well as provide privacy-preserving query functionality. CDXI must allow organizations to contribute data anonymously. The CDXI architecture must also allow an organization to replace individual components in order to achieve a higher degree of assurance where it thinks it is necessary. Finally, organizations relying on CDXI must be able to review data exchanges in order to allow detection of security issues.

Organizations that need to exchange information with each other do not always recognize a single common centralized authority for establishing trusted channels for the exchange. Organizations must therefore be able to deploy and interconnect their own CDXI "instance" as they see fit. CDXI must provide for "knowledge exchanges" that allow organizations to offer their data to others as well as discover others' data offerings. As establishment of such knowledge exchanges is open to any organization, they will provide a way to mimic the current practice whereby organizations meet with each other in different, independent communities of interest (COI) that they control. In the service offerings published through the knowledge exchanges, data providers must be able to set the terms and conditions under which others can gain access to the offered data. A decentralized model allows COIs to emerge and subside without a central authority being aware of or needing to approve this.

By making the CDXI software freely available NATO will have access to data of improved quality that is contributed by the global security community. If there is convergence towards CDXI then a "critical mass" will be reached, at which point the monetary value of the data will far exceed the cost of implementing CDXI, which will be to NATO's benefit.

B. PROVIDE FOR THE CONTROLLED EVOLUTION OF THE SYNTAX AND SEMANTICS OF MULTIPLE INDEPENDENT DATA MODELS AND THEIR CORRELATION

In early work related to cyber security information exchange, one of the key difficulties encountered was obtaining agreement within a community to a standard data model. Over time, the situation has improved and there are now a number of standards that define data models and protocols that support information sharing and automated cyber security. However, there is no consistent use of these many standards, models and protocols, which makes information sharing, collaboration and automation difficult, particularly in the absence of mappings between existing data models. Furthermore, organizations are often compelled to use the data models (standardized or not) implemented in the commercial products they have acquired. These are sometimes not interoperable, which means additional effort is required to correlate the data across products. In some cases they are also inadequate, which means an organization must complement them in order to meet its specific needs. Thus despite the existence of standardized data models, organizations must still perform a substantial amount of effort to manage data models.

Therefore, to achieve the stated objectives in the cyber security domain, CDXI must allow organizations to implement standardized data models of their choosing via an "agile data model" that allows easy definition of new or existing data models without requiring a software development cycle. The proposed CDXI approach is to use "independent topic ontologies" (ITO) that capture each data model independently; this approach allows correlation of data elements across ITOs.

In this context, the term ontology is used as defined in [22]: *"a formal explicit specification of a shared conceptualization"*, and does not necessarily imply the use of ontological languages. From a software development point of view, an ITO can be seen as a logical container for a set of classes and relationships with associated attributes. An ITO is therefore a data model covering a defined domain of interest, and CDXI does not limit the size, scope, or depth of ITOs in any way. Each instance of a class or relationship must have a globally unique identifier that can be used to correlate data across available ITOs, subject to access controls.

The use of an agile data model implies that CDXI can support any data model and does not try to force a particular one on an organization or community of interest. The latter condition is necessary because defining a single, standardized ontology that covers the entire cyber security domain is not practical. Moreover, the agile data model allows CDXI users to easily implement new data models for which

no current standards exist, as is the case for enterprise security models [23][1] and network security policies [24]. Sharing ITOs while they are in the process of being defined, and collaborative refinement of them, may also facilitate standardization efforts [25]. The agile data model allows existing data sources to be brought into CDXI relatively easily, thus taking advantage of prior investments. CDXI's support for correlation across ITOs will facilitate interoperability by allowing organizations to compose data queries that exploit ITOs that are covering the same topics at the same granularity. In a large organization, this work would be done by ontologists for the benefit of end-users.

Finally, controlled evolution of ITOs must be possible. The CDXI objective of enabling automation will be achieved when organizations use data obtained through CDXI in cyber security applications. However, the agile data model allows users to modify existing ITOs as domain knowledge evolves by adding, modifying or deleting classes, relationships or attributes and by modifying the ITO syntax or semantics. Allowing ITOs to be freely changed would give rise to problems because organizations would have to revise their cyber security applications after every ITO change to accommodate the new syntax and semantics. By enforcing comprehensive version control of ITO definitions, CDXI will allow data providers to modify their data models and data consumers to adjust their automated applications independently and at their own pace.

C. SECURELY STORE BOTH SHARED AND PRIVATE DATA

CDXI must allow an organization to store cyber security data that can be either kept private or shared with other organizations. When user data is identified as being private, CDXI must ensure that the data is never made available outside of the organization. This will allow organizations to exploit the agile data model and correlation capabilities in CDXI to store organization-specific data that is never intended to be shared, link it to data obtained from external data sources, and use the correlated information to support automated applications.

D. PROVIDE FOR CUSTOMIZABLE, CONTROLLED MULTILATERAL SHARING

Since most cyber security organizations need to interact with a range of partners for different information exchanges, CDXI must provide mechanisms that allow customizable, controlled multilateral sharing. Organizations must be able to create and manage information-sharing relationships with their partners using the

[1] Although Anderson provides an enterprise security model, it is not a standardized model.

security protocols most appropriate for each individual case. All exchange of data must be through "Information Exchange Policies" (IEP) set up by the organizations themselves. It must be possible to define any number of IEPs in order to meet the various exchange requirements.

CDXI must allow for the definition of any number of "communication channels" that implement encryption, authentication and authorization mechanisms. CDXI must allow organizations to freely associate IEPs with communication channels in order to select the most appropriate means over which a particular exchange can take place. The decision to share can be applied to entire ITOs or sub-elements of ITOs, and to all of the data or to individual data records. It must be possible to define a custom workflow for activating an IEP, as well as for authorizing the sharing of individual records in an IEP when needed.

Therefore when two or more organizations agree to exchange information with each other, they must select the applicable ITOs (thus choosing a particular ontology that describes the syntax and semantics of the data to be exchanged), identify the parties to the exchange, capture the terms and conditions under which the exchange will take place, and select the communication channels that CDXI will use to execute the exchange. This approach decouples the technical details of how to create a secure tunnel for the information over possibly insecure networks from the details related to fine-grained access controls and the terms and conditions of the exchange, such as the intellectual property rights, rights to further distribute the data and uses that can be made of it. IEPs must also allow organizations to choose a suitable accounting mechanism to support commercial activities (see requirement K). Finally IEPs must also indicate whether or not recipients can edit the exchanged data; such authorization would be given to support collaboration or outsourcing.

All exchanges must be logged and made available for audit review. Furthermore, exchanged data must always remain associated with the IEP under which it was received. CDXI must enforce the terms and conditions set forth in IEPs, and specifically the condition for redistribution of the data.

E. *ENABLE THE EXCHANGE OF DATA ACROSS NON-CONNECTED DOMAINS*

CDXI is expected to be deployed in various CISs that may not be directly interconnected (e.g. highly secure networks). CDXI must provide mechanisms to facilitate exchange across these "air gaps". Such mechanisms must provide for the auditing of the transfers in a manner that would allow for the detection of sensitive information leakage or the introduction of malicious code. The exchange of data across non-connected domains must facilitate the efficient reconciliation of

conflicting changes concurrently made in all CDXI deployments participating in an exchange of data.

F. PROVIDE HUMAN AND MACHINE INTERFACES

A key requirement of CDXI is that it provide both human-specific and machine-specific interfaces. CDXI must provide a set of graphical user interfaces (GUI) that facilitate human interaction with the data, and a set of application programming interfaces (API) that facilitate machine interaction with the data. These interfaces must be well adapted to the needs of these very different types of user.

G. PROVIDE COLLABORATION TOOLS THAT ENABLE BURDEN SHARING FOR THE GENERATION, REFINEMENT, AND VETTING OF DATA

One of the objectives of CDXI is to facilitate burden-sharing collaboration and/or outsourcing for the generation, refinement and vetting of cyber security data. While a number of organizations have established a sufficient degree of trust between each other to allow for collaboration, current information systems do not provide sufficient support to make collaboration an effective and efficient approach to generating, refining, and vetting of data, and in many cases the associated level of effort for collaboration is simply too high. Where collaboration does take place, it is often inefficient due to the absence of a facilitating system. CDXI must therefore provide tools that will address this issue.

As a minimum, CDXI must provide a timely threaded discussion mechanism that can be used to annotate different data elements. As well, it must provide a chat facility that is subject to access controls and IEPs and that provides a capability to quickly establish a shared context to support discussing a particular data element.

H. PROVIDE CUSTOMIZABLE QUALITY-CONTROL PROCESSES

CDXI will be used to aggregate and transform information from many sources to feed decision-making and automated processes. Inaccurate information could cause a business process to fail, resulting in undesired effects that can vary greatly in significance. To successfully enable automation in cyber security, CDXI must provide the means to assure the quality of the data it provides.

Quality assurance (QA) within CDXI refers to the planned and systematic activities

that ensure that the data in the CDXI system meets the quality requirements specific to its intended use. QA is achieved through the application of custom quality-control processes (QCP) that are defined by users and partly managed within CDXI. Because CDXI data can be re-used for many different purposes, ITOs, QCPs and quality requirements are associated to the use that will be made of the data, based on the concept of "curation". The curation identifies the ITOs that are needed to support an automated application as well as the QCPs that will be used to filter the data to provide only that data that meets the required quality. This allows QCPs to be re-used for different ITOs where applicable, and for ITOs to be re-used for different purposes (i.e. for different curations) even if those purposes have different quality requirements. QCPs can also be included in IEPs to ensure that data exchanged with external parties meets the desired quality requirement. In addition, CDXI must allow organizations to exchange QCPs and associated information so that QCPs can be re-used, outsourced or performed in a collaborative fashion.

I. EXPOSE DISSENSION TO REACH CONSENSUS

The fact that most databases are designed to hold a single value for each attribute of a data element, in other words only "one truth", means that users cannot express disagreement about a value except by changing the value in the database (assuming they have the necessary privileges to do so), which would then change the value for all users. Since most common data repositories have no means to expose dissension about attribute values, errors and inaccuracies recognized by users remain hidden, which limits an organization's ability to improve the data upon which it relies for operations.

CDXI must therefore expose dissension by allowing multiple possible values to be shown for each field ("multiple truths") in order to allow users to see that there is disagreement and eventually either reach consensus on which value is correct or agree to disagree. Data managers in the organizations participating in an exchange of data would have the ability to see all proposed values for an attribute and to select the one they consider to be correct for their organization, or choose to have CDXI always use the most recently entered value if they do not have the expertise to decide themselves for a particular type of data. Finally, CDXI must also allow users to easily correct detected errors and inaccuracies by allowing "divergent values" to be used locally within an organization so that automated processing can proceed with the corrected data. This functionality can also help detect and address mischievous activities directed at data sources by malicious users.

J. SUPPORT CONTINUOUS AVAILABILITY OF DATA

CDXI must meet availability requirements, even in the presence of cyber-attacks. It cannot be assumed that an organization will always have external connectivity to obtain cyber security data. CDXI must therefore allow an organization to choose to hold a local copy of selected data previously exchanged so that it can continue to use that data after disconnecting all external communication links (subject to the terms and conditions set forth in IEPs).

K. ENABLE COMMERCIAL ACTIVITIES

The private sector will be more motivated to use CDXI if it provides accounting models and functionality for selling data or data-related services. This in turn will lead to better-quality data for CDXI and thus for NATO.

CDXI must therefore provide various accounting models for the usage of data, and the mechanisms must allow vendors of data and data services to control the dissemination of data exchanged under the terms of a commercial contract. Organizations must be able to sell any data element, such as content (ITO data), and the application of quality control processes, as well as professional services related to the management and refinement of CDXI data, such as assistance in defining ITOs, correlation and translation.

If commercial activities are supported, organizations that use CDXI will be able to make use of industry's extensive resources and expertise to obtain the data they require at a cost determined by market forces, and as a result NATO will have access to the best available data.

4. HIGH-LEVEL ARCHITECTURE

To illustrate an implementation approach that could address the adaptability requirement, a high-level architecture was developed. It consists of two major building blocks: the CDXI Administrative Domain (CAD) and the CDXI Security Domain (CSD). The CAD encompasses the set of CDXI components deployed by a single organization and managed through a coherent set of administrative and high-level security policies. The CSD groups the set of CDXI components deployed in a particular network that share a common set of security services and settings and that can be directly connected to each other. Any number of CSDs can be defined within a CAD, but a CSD can belong to only one CAD. In general, a CAD will correspond to an organization, but in some cases, a larger organization may wish to deploy more than one CAD to adapt the implementation of CDXI to its organizational structure and business practices.

The CAD is used to provide coherence in the management of CSDs and to define the IEPs used by CSDs for the exchange of data. Some aspects of the management of CSDs can be centralized at the CAD (e.g. management of user accounts) or performed using management interfaces in each CSD.

In addition, the high-level architecture defines the CDXI Administrative and Security Boundary Managers (CABM and CSBM respectively). These components are used to control communications between domains. The role of the CSBM is to ensure that no data is exchanged between CSDs without a valid IEP and to take care of pulling and pushing data according to the terms of the applicable IEPs using the specified communication channel. The role of the CABM is to ensure that no data is exchanged between CADs without a valid IEP, to take care of pulling and pushing data according to the terms of the applicable IEPs using the specified communication channel, and to manage the interactions that occur with the knowledge exchanges. Both types of boundary manager provide buffering of data and a reliable exchange mechanism. Multiple instances could be deployed to provide scalability and high availability via load balancing.

5. CONCLUSIONS AND FUTURE WORK

The cyber security community requires tools to facilitate information sharing and automation, and the tools must allow for burden-sharing collaboration and outsourcing in the management of cyber security data. To address these needs, a knowledge management capability called the Cyber Security Data Exchange and Collaboration Infrastructure (CDXI) was defined. In the light of characteristics specific to the cyber security domain, the high-level requirements that must be met for the capability to achieve its objectives were identified.

As well, limited-depth investigative prototyping activities were conducted to determine which technologies are most suitable for implementing the agile data model. Possible options identified to date for implementing the agile data model include:

- Special constructs using relational database management systems (RDBMS):
  - Allowing the CDXI application to use the SQL Data Description Language (DDL) (e.g. CREATE, ALTER, DROP statements)
  - Use of an Entity, Attribute, Value (EAV) schema, which allows definition of the data model using only SQL Data Manipulation Language (DML).
  - Anchor modeling, which describes the data at high normalization levels using a graph notation based on anchors, attributes, ties, and knots (an approach similar to EAV).

- Triplestores for RDF (Resource Description Framework) or OWL (Web Ontology Language) as used for the development of semantic webs.
- Non-SQL solutions or schema-less databases such as MongoDB.

The prototyping activities conducted to date for the agile data model were based on the first two special constructs above and the use of a conventional RDBMS. The first prototype activity was based on the use of DDL, while the second was similar to the one introduced in [26] for the definition of genome ontologies. While the findings of these limited-depth trials suggest that the dynamic creation of ITOs using DDL would be a better approach than the use of an EAV schema, further work is required to confirm this and to assess the other approaches as well. At the moment the expectation is that the final implementation of an agile data model will likely not rest on a single solution but on a combination of the technologies mentioned above.

An initial proof-of-concept design was also developed. This work helped identify lower-level requirements and technical approaches for the implementation of CDXI, and is documented in NATO technical reports.

ACT has sponsored validation of the CDXI capability defined in this paper through an engagement with NATO stakeholders and subject-matter experts in NATO nations, industry, and academia, as well as a review of existing prototypes and capabilities that provide similar functionality. If the initial feedback indicates that it is necessary, the validation activity may be extended to include the development of a proof-of-concept. Once the CDXI capability is validated, options available for the procurement of an operational, production-grade CDXI will be considered. In parallel, further work will likely be conducted to refine specifications, identify minimum performance requirements, and investigate the suitability of existing technologies and standards in order to support the procurement process.

REFERENCES

[1] Hedlund, G. (1994). A Model of Knowledge Management and the N-Form Corporation. *Strategic Management Journal, 15*, 73-90.

[2] Alavi, M., & Leidner, D. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, *25:1*, 107-136.

[3] Tyugu, E. (1993). Large Engineering Knowledge Bases. *Artificial Intelligence in Engineering*, 8(4), 265-270.

[4] INFOSEC Research Council. (2006). Hard Problems List. Washington DC, Cyber Security and Information Assurance Interagency Working Group (CSIA IWG).

[5] Nardon, F., & Moura, L. (2004). Knowledge sharing and information integration in healthcare using ontologies and deductive databases. *Medinfo*, 62-66.

[6] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., & Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research, 32*(suppl 1), D262-D266.

[7] Ferrara, L., Mårtenson, C., Svenson, P., Svensson, P., Hidalgo, J., Molano, A., & Madsen, A. (2008). Integrating data sources and network analysis tools to support the fight against organized crime. *Intelligence and Security Informatics*, 171-182.

[8] Fulton, J. (1992). Technical report on the semantic unification meta-model – Standards working document ISO TC184/SC4/WG3 N103. Seattle, IGES/PDES Organization, Dictionary/Methodology Committee.

[9] Allen, J., & Lehrer, N. (1992). DARPA/Rome Laboratory Planning and Scheduling Initiative Knowledge Representation Specification Language (KRSL), Version 2.0.1 Reference Manual. ISX Corporation.

[10] Gruber, T. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies, 44*(5-6), 907-928.

[11] Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: the state of the art. *The Knowledge Engineering Review, 18*(1), 1-31.

[12] Bonifacio, P., Bouquet, P., Mameli, G., & Nori, M. (2004). Peer-Mediated Distributed Knowledge Management. *Lecture Notes in Computer Science, 2926*, 31-47.

[13] Ehrig, M., Tempich, C., Broekstra, J., van Harmelen, F., Sabou, M., Siebes, R., Staab, S., & Stuckenschmidt, H. (2003). SWAP: Ontology-based Knowledge Management with Peer-to-Peer. *Workshop ontologiebasiertes Wissensmanagement* (pp. 17-20). Lucern: Gesellschaft für Informatik, Lecture Notes in Informatics (LNI), P-28, Bonn.

[14] ENISA. (2011). EISAS (enhanced) report on implementation.

[15] Martin, R. (2008). Making Security Measurable and Manageable. *Proceedings of the IEEE Military Communications Conference*, 19. San Diego.

[16] ITU-T. (2011). Overview of cybersecurity information exchange. Geneva, ITU-T.

[17] Internet Engineering Task Force Request for Comments 6545, "Real-time Inter-network Defense (RID)", K. Moriarty, IETF, April 2012.

[18] Internet Engineering Task Force Request for Comments 6546, "Transport of Real-time Inter-network Defense (RID) Messages over HTTP/TLS", B. Trammell, IETF, April 2012.

[19] Internet Engineering Task Force Request for Comments 5070, "The Incident Object Description Exchange Format", R. Danyliw, J. Meijer, & Y. Demchenko, IETF, December 2007.

[20] Blanco, C., Lasheras, J., Valencia-García, R., Fernández-Medina, E., Toval, A., & Piattini, M. (2008). A Systematic Review and Comparison of Security Ontologies. *The Third International Conference on Availability, Reliability and Security*, 813-820.

[21] Hartman, B. M. (2012). RSA Security Brief February 2012 – Breaking Down Barriers to Collaboration in the Fight Against Advanced Threats.

[22] Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199-220.

[23] Anderson, E., Choobineh, J., & Grimaila, M. (2005). An Enterprise Level Security Requirements Specification Model. *38th Hawaii International Conference on System Sciences.* IEEE Computer Society.

[24] Cuppens, F., Cuppens-Boulahia, N., Sans, T., & Miège, A. (2005). A Formal Approach to Specify and Deploy a Network Security Policy. *International Federation for Information Processing*, *173*, 203-218.

[25] Sofia Pinto, H., Staab, S., & Tempich, C. (2004). DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely controlled and evolvInG Engineering of oNTologies. *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, 393397. Valencia, Springer.

[26] Nadkarni, P., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., & Miller, P. (1999). Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. *Journal of the American Medical Informatics Association*, 6(6), 478–493.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Deriving Behavior Primitives from Aggregate Network Features using Support Vector Machines

Owen McCusker
Guardian Services
Sonalysts, Inc
Waterford, CT, USA
mccusker@sonalysts.com

Scott Brunza
Guardian Services
Sonalysts, Inc
Waterford, CT, USA
scottso@sonalysts.com

Dipankar Dasgupta
Professor of Computer Science
University of Memphis
Memphis, TN, USA
dasgupta@memphis.edu

Abstract: Establishing long-view situation awareness of threat agents requires an operational capability that scales to large volumes of network data, leveraging the past to make-sense of the present and to anticipate the future. Yet, today we are dominated by short-view capabilities driven by misuse based strategies; triggered by the structural qualities of attack vectors. The structural aspects of cyber threats are in a constant flux, rendering most defensive technologies reactive to previously unknown attack vectors. Unlike structural signature based approaches, both the real-time and aggregate behaviors exhibited by cyber threats over a network provide insight into making-sense of anomalies found on our networks. In this work, we explore the challenges posed in identifying and developing a set of behavior primitives that facilitate the creation of threat narratives use to describe cyber threats anomalies. Thus, we investigate the use aggregate behaviors derived from network flow data establishing initial behavior models used to detect complex cyber threats such as Advanced Persistent Threats (APTs). Our cyber data fusion prototype employs a unique layered methodology that extracts features from network flow data aggregating it by time. This approach is more scalable and flexible in its application in large network data volumes. The preliminary evaluation of the proposed methodology and supporting models shows some promising results.

Keywords: *Behavior analysis, aggregate behaviors, network flow analysis, anomaly detection, machine learning*

1. INTRODUCTION

The North Atlantic Treaty Organization (NATO) is faced with the increasing need to support international operations that leverage the use of complex end-to-end architectures. NATO Network Enabled Capability (NNEC) is an integral program focused on meeting these needs [1]. The ubiquity of these net-centric information systems is realized by the connectivity of hand-held technologies, operated and managed by users in the field, to backend mission support systems managed by tens of thousands of administrators. The attack surfaces associated with such systems allows cyber threat agents (e.g. nation states, hackivists) to employ the use many types of coupled attack vectors, such as phishing and key-logging; gaining access and persisting in these environments for years unnoticed. The complexity of these cyber threat agents has grown steadily in recent years, and is exhibited in the employment of distributed cyber missions that operate over various time scales within our Information and Communication Technology (ICT).

In 2013, Kaspersky Lab uncovered the actions of ‘’Red October” which they feel has been harvesting intelligence from high profile organizations since 2007 [2]. This espionage group incorporated a set of simple attack vectors that allowed them to penetrate and persist in both public and private organizations for prolonged period of time. According to Verizon in 2012, threat agents incorporate multiple threat actions during an attack, and these attacks can go on form months well within our supply chains and distributed throughout our networks [3]. Yet our detection models and cyber defence capabilities are still tuned for single ingress points, and mostly employ rule-based defensive strategies.

There is an array of defence-in-depth capabilities that can be employed in concert to deter the sophisticated attacks from threat agents including: firewalls, Multi-factor authentication, role and attribute based access control end-point security, Network Intrusion Detection/Prevention Systems (IDS/IPS) training and policy creation and enforcement. Each capability provides deterrence to attack vectors in a slightly different way. Most monitoring and response capabilities can be categorized into misuse detection and anomaly detection. While the misuse detection can only detect known attacks, anomaly detection on the other hand can detect unknown and zero-day attacks. However, anomaly-based detection methods suffer from false positives, as all anomalies may not relate to attacks. This work is on leveraging behaviour-based anomaly detection with focus on hierarchical aggregated features/ attributes of monitored hosts.

One of the reasons why threat agents pose such a significant risk to national infrastructure is that cyber defence capabilities are dominated by misuse-based capabilities that provide short-view situation awareness. Most of these capabilities

correlate volumes of real-time events making sense of what is happening at any given instant in time but do not scale well over longer time periods. The anomaly detection paradigm offers the ability to adapt to emergent threats based on past events.

In of 2009, BBN addressed this issue by proposing a notional architecture that can scale at increasing network speeds using event aggregation [4]. Key to the success in their approach is the use of Scyllarus, an event correlation system [5]. This correlation system clusters events by measuring the similarity of their attributes. Our position is to take a host-centric posture, instead of event-centric, focusing on the aggregate behaviours of hosts as extracted from using network flow traffic.

Over the past few years behaviour-based models have emerged to bridge the gap in capability focused on anomaly defection of emergent threats [6], [7], [8]. These systems are mostly event-centric, where behaviours are extracted from event features and aggregated over time in terms of a source and a destination. For example, in [7], aggregate event graphs are used make sense of behaviours obtained from sensors. The event takes into account both the source and destination providing a connection, or edge in the graph. In another example, Rehak uses classifiers agents to score events as legitimate or malicious [6]. Lastly, LNLL created a system SETAC that uses a distributed model to detect both local and global anomalistic behaviours within their networks [8]. Unlike the previous systems, we position our host-centric work to develop layers of classifiers, with the first intermediate step toward establishing a set of primitives used measure overall behaviours of hosts.

In our previous work [9], we developed a host-centric cyber data fusion capability based on a layered methodology (Figure 1), which transforms network flow data into aggregate features of hosts over various time windows. We collected network flow data using SiLK over a period of six months to begin our exploration of the data [10]. One of our findings suggests that when a group of host is observed over a period of time, they behave in very consistent ways.

In our current work, we are looking to develop an adaptive methodology that leverages the past aggregate behaviours of normal operation of systems to build a predictive model. We then compare the predicted behaviour of a host or set of hosts with the actual behaviours to determine the classification of the abnormality of a host. The overall classification is measured in terms of a set of behaviour primitives. In order to minimize false positives in attack detection, this approach incorporates some signalling mechanism similar to the biological immune system.

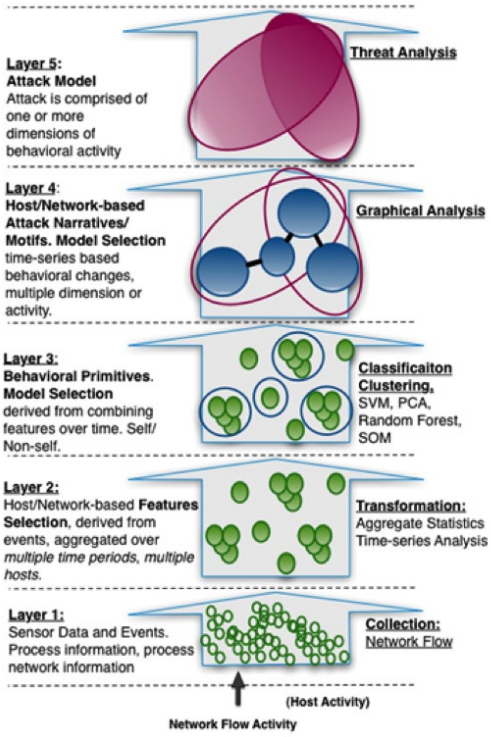

Figure 1. A Layered Methdology. Most CND technology is developed to operate in Layer 1, processing volumes of raw data and events from sensor technology. Our unique approach transforms the data first, Layer 2, and then applies classification models in Layer 3 and above

Figure 1 illustrates a layered detection methodology approach, which allows the Team to work independently at different abstraction layers of the overall problem. In this approach, we can focus on developing algorithms in Layer 1-3 and in the future we can focus on social-based algorithms in Layer 4-5. We envision multiple algorithms to leverage in the layered model.

The approach we take in this research is to establish a rich set of behaviour primitives that facilitates long-view situation awareness. The behaviour primitives represent the basis for a behavioural language through which we can someday create threat narratives that are shared as actionable intelligence in real-time throughout a trusted community of cyber defenders. Narratives are viewed as graphs of behaviour primitives that capture aggregate description of threat agents. These threat narratives can represent social relationships and/or characteristics that are shared between a group of hosts, geographic region, and/or autonomous system.

This paper is organized as follows. In Section 2 we review work related our proposed approach. Section 3 we discuss our methodology [11], which includes ground truth development, feature selection, model development and model evaluation. In Section 4 we discuss the conclusions of our results.

2. RELATED WORK

In this section, we review related intrusion detection research leading to behavior analysis. We then present important works on distributed collaboration and correlation, intrusion detection leveraging network flow, cyber situation awareness, and review pertinent work on knowledge discovery. This section contrasts the evolving threat with the models that were used in establishing existing detection technologies.

In 1987, an intrusion detection model proposed by Denning focused on the identification of network attacks directed toward a single host [12]. The threat, at that time, was comprised mostly of attackers attempting to gain remote access to a host. Soon after this model was proposed, the introduction of worms was officially acknowledged by the release of the Morris Worm in 1988 by Robert T. Morris [13]. Since this time, there has been a constant tug-of-war between the introduction of new threat types and the development of new techniques to meet the evolving detection requirements. Ghosh et al. [11] developed an application level behavior model for intrusion detection.

A. *MULTI-EVENT CORRELATION AND DISTRIBUTED COLLABORATION*

In Section 1, as discussed by [4], event correlation can facilitate aggregation and scaling to network speed. BotHunter [14] is a system built specifically for the correlation of events occurring within specific network locales. This system focuses on detecting network dialog communications between various bots within a botnet and is driven by alerts from SNORT [15]. These dialogs represent different communication behaviors exhibited by a bot during its lifecycle. An event trail is created that triggers an alert based on specific bot behaviors that occur.

The Worminator project leverages the distributed collaboration of events generated from an IDS in order to establish attack patterns [16]. The system leverages alert aggregation and reduction to reduce the cost of the exchanging raw data. A correlation scheduler is used to set up peers to exchange alerts. The Worminator paper highlights the need to reduce and manage the large volumes of alerts that are exchanged between detection peers. Worminator uses Bloom filters to manage privacy by setting up private watch lists.

Our proposed host-centric model is driven by network flow captured using SiLK instead of an event-centric IDS. We derive profiles consisting of features extracted from the network communication between various hosts. These behavior profiles are fed into a classification and correlation engine.

B. KNOWLEDGE DISCOVERY AND ADAPTABILITY

In Section 1 we discussed the knowledge discovery needs for a system to adapt by leveraging the past to the present in [6]. Knowledge engineering has been applied to intrusion detection in MADAM ID [17] where association rules mining was used offline to construct new rules to detect threats in a misuse detection system. Knowledge discovery has been applied in another way for misuse detection in the Intelligent Intrusion Detection System (IIDS) [18]. Misuse signatures are viewed as rules through which a genetic algorithm creates a set of rules by combining behaviors based on network connection information. In both cases, rules are directly related to threat signatures. We propose a more abstract view dealing with knowledge discovery, where threats are represented in a set of behavior primitives and extracted features.

C. DMNET – A CYBER DATA FUSION PROTOTYPE

The overall system [9] focuses on the notion of tracking various network objects, O, e.g. hosts, hostgroups, and networks, and determining if they are threats. Tracking these objects involves collecting events and data from a number of different network sensors, e.g., network flow, NIDS, honeypots, and creating a sample space.

In our current data fusion system, network flow data and alerts generated by network sensors reflect the totality of information and model's sample space, S, available to the detection system regarding the objects to be analyzed.

To utilize this data, it is first normalized and transformed into a representation that is conducive to algorithmic processing. The fusion engine operates over a sample space denoted as S representing sensor data. This fusion operation is represented by an object behavioral analysis function, B.

The aggregated behavioral analysis of the sample for a specific object O, $B(S_O)$, produces a feature characteristic, or behavior, for that object denoted by F_0 accumulated within a set time window $F_{tw,O}$. The sample space, S, is then transformed into an aggregated feature space F_S. The Time window, tw, consists of periods such as hour, day, month, year. F_0 is represented by a n-tuple, or n-gram, of individual time-based features, for example $F_{month,O} = \langle f_1, f_2, \ldots f_n \rangle$, describes O over a period of a month. These features consist of structural, behavioral, and/or application specific properties of O over a given time period.

Information from the deployed sensors is fed into the fusion engine. Sensors could include a variety of network, appliance, or host-based software or hardware. The sensor information could be in the form of netflow [10] or pcap records, network intrusion detection/prevention system feeds, alerts from honeypots, or anti-virus reports. This information is parsed and then normalized by a perception module. Normalization refers to the process of converting the parsed information into a form that is standard and readily understood and manipulated by modules further down in the processing chain.

The data fusion component maps normalized data to vectors of high dimensionality. This is achieved by a profiling function that parses the raw normalized events produced by the vectors and aggregates them to form a basic network object and embeds them in a vector space. After the profiling is completed, each fusion element is associated with a feature characteristic that describes it according to the profiling function that was applied. Note that the features that can be extracted depend upon the type of sensor provided to the system as a source of network data. They range from summary data such as netflow, to fine-grained information such as pcap header dumps produced by tcpdump.

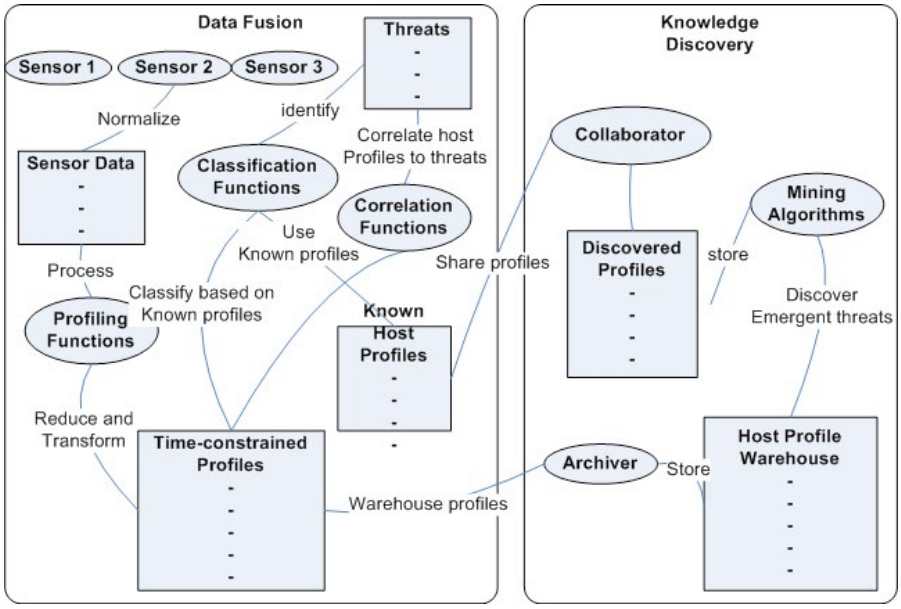

Figure 2. Dmnet Cyber Data Fusion Prototype. This architecture represents a combined fusion and data mining methdology.

D. AGGREGATE BEHAVIOR ANALYSIS

Most current technology operates at Layer 1 (Figure 1) in our methodology applying classification models to raw data and sensor events. We need technologies that scales to the volumes of data and events being created by our cyber sensors.

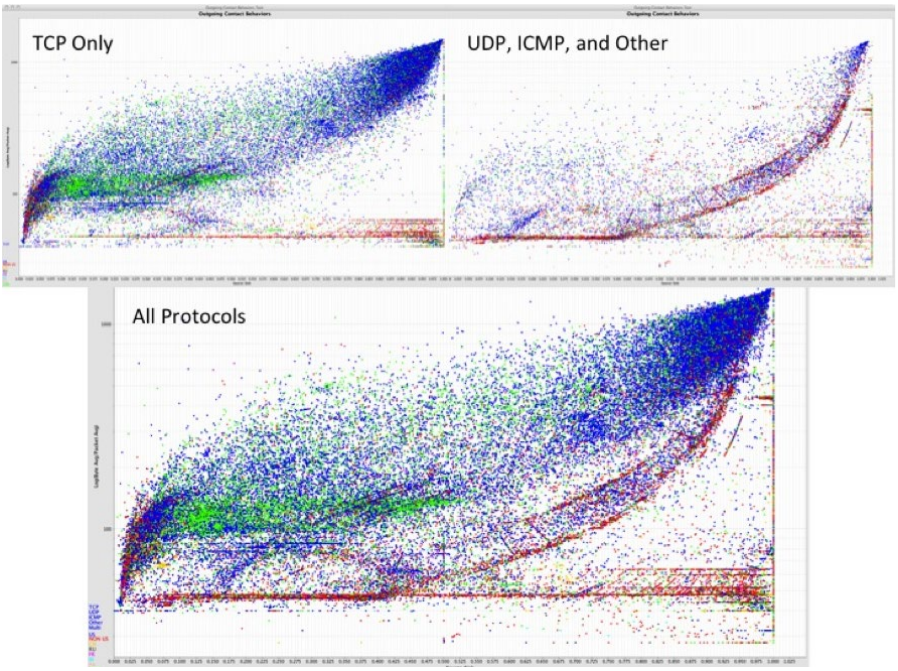

Figure 3. Behavioral Visualization Created From Data Generated by Fusion Prototype. There are three different visualizations depicted showing UDP behaviors (top right), TCP Behaviors (top left) and All protocols bottom. Each dot represents a host in a behavioral feature space. These diagrams show the behaviors of hosts going from "source to sink behaviors," where hosts receiving data from our system are to the left, and hosts sending data to our system are to the right.

In previous years, Sonalysts started the development of a disruptive cyber fusion approach based on aggregate behavioral analysis. Our approach transforms this data, Layer 2, first into a rich multivariate features space before we apply classification models (Layer 3).

Layer 1 CND technologies cannot scale well when faced with the increasing amount of network traffic. By transforming this raw data into behaviors we can aggregate it into multiple time periods and provide a data reduction technique that can begin to scale to the increase in traffic volumes.

3. CHARACTERIZING BEHAVIOR PRIMITIVES

This section highlights the overall methodology for model development that is being employed to detect behavioral primitives enumerated in the ground truth data set. We evaluate the feasibility of our methodology by applying to three different types of classification models focused on the identification of pinging, or beacon-like behaviors. This is one of many types of behavioral primitives that we are working on quantifying as part of the ongoing research.

Figure 4. Behavior Primitive Taxonomy

The behavioral primitives are captured in a behavioral taxonomy (as shown in Figure 4). Each node represents a primitive that can be measured in terms of a set of features created by our fusion engine and by a classification model. For example, in this paper beaconing behavior is modelled using a Support Vector Machine (SVM) in terms of three aggregate features: outgoing work, outgoing byte variance, and source sink. Outgoing work is defined as the average bytes per packet that leaving a network device e.g. host. Outgoing byte variance is measures the changes in bytes per packet in outgoing flow traffic from hosts. Source sink is a measure of the directionality of traffic from network device and has a value of 0 to 1. Where purely beaconing devices have a value of 1.

A. GROUND TRUTH

The research is leveraging ground truth behavioral data gathered between the months of December 2010 and to February 2011. This data is derived from live network flow traffic that we continually capture on our networks and transform into a behavioral features space. The goal in leveraging this ground truth is produce a set of behavioral primitives that can be used to perform predictive analytics using a number of learned models.

The behavioral data for the work is gathered from a number of discrete vantage points: External to the firewall focused on non-assets hosts (not managed by the client), internal focused on non-assets hosts, and internal vantage point focused on assets. We have been gathering behavior data actively since 2009 and to date we have shared ground truth data with institutions to promote aggregate behavior analysis (2010, Oakridge National Laboratory[1].)

1) Meaningful Indicators

We have identified a number of meaningful indicators during the analysis of the three ground truth data sets. Some of these indicators are highlighted in this action of the document.

a) External Vantage Point Non-Assets

There are over 1.7 million hosts being followed in the external vantage ground truth data set. The data set is rich with host behaviors found in both monthly and daily time aggregates. The following picture highlights abnormal activity, against policy of a host running a Unreal Tournament client and having it beacon out to a number of external server hosts. This asset is compromised a few weeks later.

b) Internal Vantage Point Non-Assets

The internal vantage point provides insight to actual communications between assets and non-assets, without the noise inherent from outside the firewall. In (Figure 5) there is a cluster of behaviors associated with internal hosts performing a heartbeat out to Japan. There are multiple machines that are sending a consistent amount of bytes and packets to this server. These machines are on a internal subnet through which there where known compromised machines.

[1] Oakridge National Laboratory, Computational Intelligence Behavior Modeling Laboratory, promoting the use of scalable algorithm development using High Performance Computing technologies, http://csiir.ornl.gov/

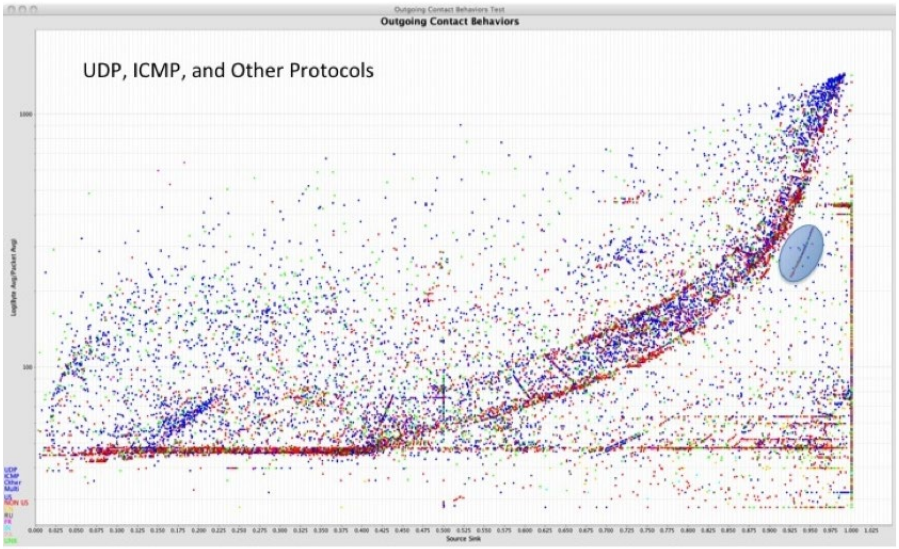

Figure 5. Visualization of Monthly External (non-assets) Host behaviors from the External Vantage Point. All traffic in this view is filtered out except for UDP, ICMP, and Other. See Figure 3 for a comparison between the protocol specific behaviors. The highlighted points are that of a single host in mid December launching Unreal Tournament and beaconing to sites around the globe

c) Internal Vantage Point Assets

The Internal vantage point ground truth data set offers the highest fidelity of behavioral features. We are only tracking 1400 hosts from this vantage point compared to 1.7M hosts on the external one. Having a smaller amount of contacts can allow us to focus on finer grained temporal features looking into both the quantification of normal and abnormal behaviors that provide a side-by-side comparison of host behaviors looking at byte and packet usage.

B. BEHAVIOR TAXONOMY

In a paper delivered to NATO in 2010 (and based on our work for DHS S&T from 2006), we established two taxonomies facilitating the understanding of trust in end-to-end systems [19]: sensor taxonomy, and a behavioral taxonomy. The sensor taxonomy provides a basis for which we associate what behavioral features are derived from the various sensors employed by the system. We are further refining the two sets of taxonomies to support our work. In the future, these taxonomies will be developed into feature Ontologies with the addition of meaningful attributes to each node. The goal of this work is to identify behavioral primitives derived from the analysis of sensor data.

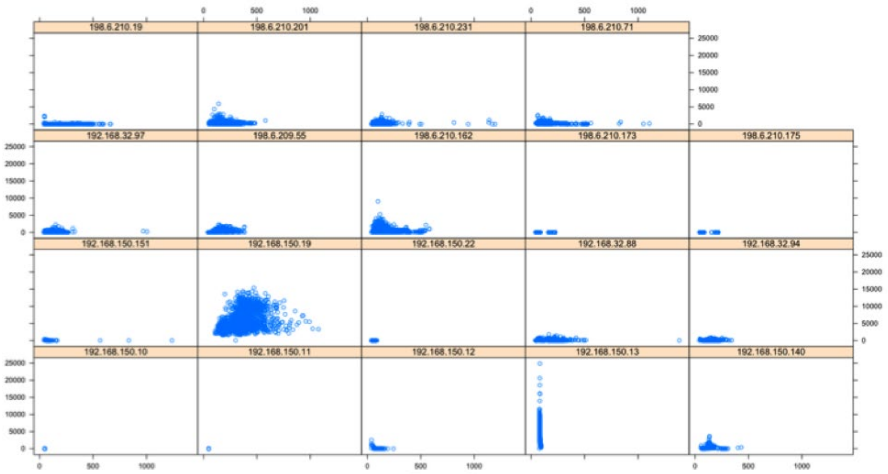

Figure 6. *Comparative Behaviors of Internal Hosts (Assets).* In this graph we are looking at byte versus packet behaviors. This side-by-side visualization of internal asset behaviors presents the degree of behavioral differences between hosts. The host in row 3, from the top, and column 2 from the left is the email server. The host in row 4 and column 4 is a DNS

To date, we are only focused on network flow. In the future, we plan on integrating other types of sensor technology into the prototype. The prototype already has an extensible sensor management framework.

We expect this development to be iterative in nature and mature as we begin to apply multiple classification models to derive meaningful behavioral primitives.

The behavior taxonomy (Figure 4) serves as a way to organize the various behavioral primitives that are being researched within the ground truth data set. Our goal is to have a way to score each of the behavioral primitives found within the taxonomy based on a specific classification model. A first attempt in modeling behavioral primitives is focused on beaconing behaviors as addressed in the previous section. We will plan on choosing multiple features for each model. Our goal is to be able to correlate multiple behavioral features to create graphical narrative describing threat agent behaviors.

C. BEHAVIOR MODEL DEVELOPMENT

Our techniques differ for a number of approaches that focus on the detection of specific classes of applications and attacks using models such as SVMs. Instead of classifying each individual flow of communication from a host we focus on the aggregation of transformed features to one specific host. By taking a host-

centric approach in our methodology we are able to collect meaningful behavioral aggregations of hosts, subnets, and geographic regions.

Li *et al.* [20] have applied the use of SVMs to detect seven classes of applications with optimized yields of 96.4% accuracy with un-biased training data. Their work has classified the following types of applications: Bulk (ftp), interactive (ssh, telnet, rlogin), mail (pop, smtp, imap), service (x11, dns), www (http, https), p2p (kazaa, bittorrent, gnutella), multimedia (voice, video streaming), game (half-life), attack (worms, virus), and other. The approach, although accurate, is high grained.

Instead of using a SVM to classify an application, our approach is finer grained in that by decomposing an application, or a threat agent, into a set of behaviors we will create behavioral language, or narrative, used to describe the threat actions over time. Lastly, instead of focusing on one specific model we are researching a number of models that operate over various time-based behavioral apertures or granularities.

1) Model Development using Support Vector Machine

Our initial goal is to focus on the predictive performance associated with ability to score behavioral primitives. There are a number of existing criteria that exists for evaluating models: predictive performance, interoperability, and computational efficiency. One reason for this choice is that our methodology allows for the concurrent processing of multiple models, which can be an area we focus on in future spirals. Our ultimate goals is to develop a set of primitives using supervised learning methods and then to augment this approach with unsupervised learning methods with the larger data sets. Essentially deriving new models, or variations of models, adding to our behavioral Ontology. For example, there can exist different variations of beaconing used by threat agents as they penetrate our systems. We will evaluate our models using receiver operator characteristic (ROC) curves.

2) Support Vector Machine Model Evaluation

A Support Vector Machines (SVM) represent a supervised pattern recognition algorithm used for binary classification problems. Being a supervised method, we are using our ground truth data set to train a SVM to detect various types of behavioral primitives, beaconing being the first. We are using the LibSVM library and R to apply SVM to our data set[2]. Unlike previous research done in our community [20], we are applying SVMs to host-centric behavioral features. Most of the research to date has applied these models to network communications and raw flow data. Within our methodology we have transformed the data into a

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

host-centric features space before we apply our models. Scaling the data input into the SVM is important. Without doing so the attributes with the higher numeric ranges can dominate the models output. This is especially true when using linear or polynomial kernels.

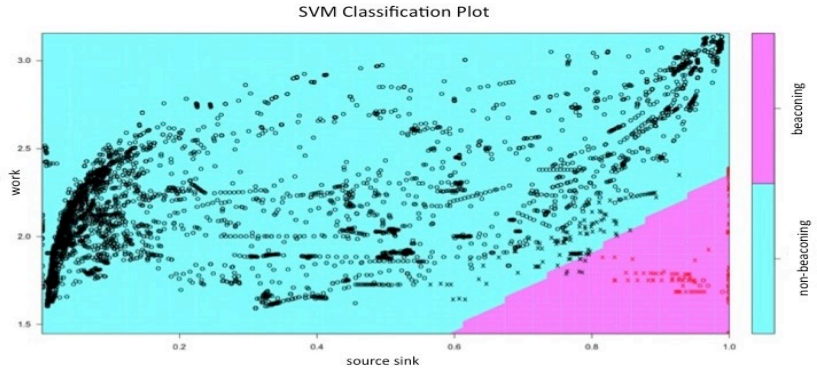

Figure 7. Trained SVM Visualization. Beacon behavior is below and to the right of the hyperplane

To assess Beaconing event (Figure 7) detection accuracy in a threshold-independent manner we use Receiver Operating Characteristic (ROC) curves (Figure 8), i.e., plots of achievable sensitivity vs. false positive rates, where the Sensitivity/True Positive Rate (TPR) is defined as the ratio between the number of Beaconing events (TP) flagged by the algorithm and the total number of Known Beaconing events (P), and the False Positive rate (FPR) is defined as the ratio between the number of non-Beaconing events (FP) flagged by the algorithm and the total number of non-Beaconing events (N).

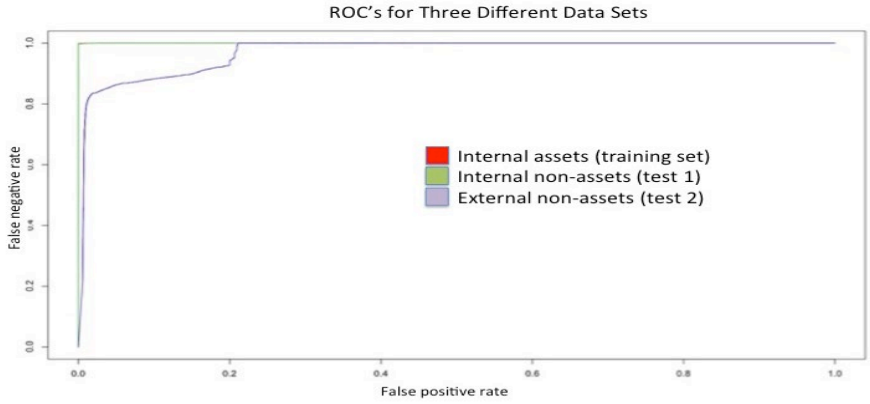

Figure 8. ROC Curve Results. Three different data sets were used: internal assets, internal non-assets, and external non-assets. Both the internal assets and internal non-assets exhibited a clean seperation between beaconing and non-beacon like behavior.

Our work focuses on two-class prediction problems (binary classification) where on class is a positive outcome and the other class has a negative outcome. We present contingency tables in the evaluation of a SVM in classifying pinging, or beacon-like behaviors.

Each point in this visualization (Figure 7) represents a days worth of host behaviors. The block circles are the non-pinging behaviors, and the red circles represent pinging behaviors. The light regions to the left and up are the predicted non-pinging behavior regions and the ping is the predicted pinging region. We are leveraging the use of SVMs to identify behaviors within the data set. We have selected work and source sink features to run against the model to detect beaconing behavior. We train the model with from our ground truth data sets focusing on pure beaconing behavior that exists, where source sink has a value of 1.

The training data set, taken from the internal vantage point, contained 6,635 hosts. The number of hosts exhibiting beaconing was 480. The following contingency table relates the true positive results to the predicted results and shows that 2 hosts where incorrectly predicted in the model. We labeled the data based on source sink and work feature values. This data is biased based on our labeling. We will run the data later on more unbiased data sets.

The unknown data set 1 contained 68,165 hosts. There were very few hosts having behavior indicative beaconing. The number of hosts exhibiting beaconing was 39 and had no false positive or negative errors in this data set.

The Test Data Set 2 (by see Table I) results show that our model was 86.9% accurate using the model developed from the internal training set. The contingency table provides an overview of the false positives (FP) 12,147 hosts, and false negatives (FN) of 50,031 hosts.

Table I. Contingency Table for Data Set 2

		Observed		
		0	1	Total
Predicted	0	195902	12147	208049
	1	50031	214583	264614
Total		226730	245933	472663

4. CONCLUSION

In this paper, we introduced a methodology for establishing behavior primitives in facilitating the creation of long-view situation awareness. Beaconing is just one primitive we will identity, and in our research are looking to grow that list of primitives to a few hundred

We discussed the concept of a behavior aggregation and its use in accurately measuring beaconing. In our work, the establishment of behavior primitives as an integral step leading to future detection, trust and risk models detecting and anticipating emergent behavior of compromised networked devices.

System behaviors can be used to develop models of trust to secure complex network [6], [19], [21], where trust is modeled from changes in past behaviors.

We have presented a classification model that utilizes aggregate features to create behavior profiles using a prototype cyber data fusion system. Since Denning proposed an alert-centric intrusion detection model back in 1987 protecting hosts from threats [12], new detection models are needed to advanced persistent threats (ATPs) that are realized from multiple ingress points within a network. The foundation of our work resides in the use of profiles in:

- The realization of behavior primitives to be later used in the knowledge discovery system,
- Collaboration between the discovery system and the fusion system, and
- The future establishment of threats in terms of behavior graphs in the fusion system.

Acknowledgment

Sonalysts would like to acknowledge support from of the Cyber Security Program Area of the Command, Control and Interoperability Division within the Science and Technology Directorate of the U.S. Department of Homeland Security, especially the support from Dr. Douglas Maughan.

REFERENCES

[1] NATO, “NATO Architecture Framework,” NATO, Technical Report 2007.

[2] Kaspersky. (2013, Jan.) www.securelist.com. [Online]. http://www.securelist.com/en/analysis/204792262/Red_October_Diplomatic_Cyber_Attacks_Investigation

[3] Verizon, “2012 Breach Investigation Report,” Verizon, Technical Report 2012.

[4] T. Strayer et al., "An Architecture for Scalable Network Defense," BBN, Technical Report 2009.

[5] W. Heimerdinger, "Scyllarus intrusion detection report correlator and analyzer," in *DARPA Information Survivability Conference and Exposition, 2003. Proceedings*, vol. 2, 2003, pp. 24-26.

[6] Martin Rehak et al., "Dynamic information source selection for intrusion detection systems," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, vol. 2, Richland, SC, 2009, pp. 1009-1016.

[7] B Czejdo, E Ferragut, J Goodall, and J Laska, "Network Intrusion Detection and Visualization Using Aggregations in a Cyber Security Data Warehouse," *Int. J. Communications, Network and System Sciences*, vol. 5, pp. 593-602, Sept 2012.

[8] Arner Heller. (2010, Jan) str.llnl.gov. [Online]. https://str.llnl.gov/JanFeb10/matarazzo.html

[9] O McCusker, A. Kiayias, D. Walluck, and J. Neumann, "A Combined Fusion and Mining Strategy for Detecting Botnets," in *ATCH '09: Proceedings of the 2009 Cybersecurity Applications and Technologies Conference for Homeland Security*, Washington, DC, 2009, pp. 273-284.

[10] Timothy Shimeall, Sidney Faber, Markus DeShon, and Andrew Kompanek. (2010, Jan) Using SiLK for Network Traffic Analysis. [Online]. http://tools.netsa.cert.org/silk/analysis-handbook.pdf

[11] Anup K. Ghosh, Aaron Schwartzbard, and Michael Schatz, "Learning Program Behavior Profiles for Intrusion Detection.," in *In USENIX Proceedings of the Workshop on Intrusion Detection and Network Monitoring*, Santa Clara, California, USA, April 9-12, 1999.

[12] Dorothy E. Denning, "An Intrusion Detection Model," in *Symp. on Security and Privacy*, Feb 1986, pp. 118-133.

[13] Eugene H. Spafford. (1988, Dec) spaf.cerias.purdue.edu. [Online]. http://spaf.cerias.purdue.edu/tech-reps/823.pdf

[14] Guofei GU, Phillip Poras, Vinod Yegneswaran, Martin Fong, and Wenke Lee, "BotHunter: detecting malware infection through IDS-driven dialog correlation," in *SS'07: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, 2007, pp. 1-16.

[15] Martin Roesch, "Snort: Lightweight Intrusion Detection for Networks," in *Proceedings of LISA '99: 13th Systems Administration Conference*, 1999, pp. 229-238.

[16] M.E. Locasto et al., "Collaborative Distributed Intrusion Detection," Columbia University, Technical Report CUCS-012-04, 2004.

[17] Wenke Lee and Salvatore J. Stolfo, "Combining Knowledge Discovery and Knowledge Engineering to Build IDSs," in *Recent Advances in Intrusion Detection*, 1999.

[18] We Li, "Using Genetic Algorithm for Network Intrusion Detection," in *In Proc. United States Department of Energy Cyber Security Group 2004 Training Conference*, 2004, pp. 24-27.

[19] Owen McCusker et al., “Combining Trust and Behavioral Analysis to Detect Security Threats in Open Environments,” in *NATO/OTAN*, 2010, RTO-MP-IST-091.

[20] Zhu Li, Ruixi Yuan, and Xiaohong Guan, “Accurate Classification of the Internet Traffic Based on the SVM Method,” in *2007. ICC ‘07. IEEE International Conference on Communications*, 2007, pp. 1373 -1378.

[21] O McCusker, B Gittens, J. Glanfield, S. Brunza, and S. Brooks, “The Need to Consider Both Object Identity and Behavior in Establishing the Trustworthiness of Network Devices within a Smart Grid,” in *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, vol. 54, 2010, pp. 1-4, 10.1145/1852666.1852724.

[22] Vern Paxson, “Bro: a system for detecting network intruders in real-time,” in *7th USENIX Security Symposium*, 1998.

[23] H. S. Javitz and A. Valdes, “The SRI IDES Statistical Anomaly Detector,” in *IEEE Symposium on Security and Privacy*, 1991, pp. 316-326.

[24] John McHugh, “Sets, Bags, and Rock and Roll: Analyzing Large Data Sets of Network Data,” in *ESORICS, 2004*, 2004, pp. 407-422.

[25] CERT. System for Internet Level Knowledge. [Online]. http://tools.netsa.cert.org/silk/

[26] Carrie Gates and John McHugh, “The Contact Surface: A Technique for Exploring Internet Scale Emergent Behaviors,” in *DIMVA, 2008*, 2008, pp. 228-246.

[27] Gerhard Munz and Georg Carle, “Real-time Analysis of Flow Data for Network Attack Detection,” in *Integrated Network Management*, 2007, pp. 100-108.

[28] CESNET, “Network Security Monitoring and Behavior Analysis: Best Practices Document,” CESNET, Technical Report 2011.

[29] Shu Yun Lim and Andy Jones, “Network Anomaly Detection System: The State of Art of Network Behaviour Analysis,” in *Proceedings of the 2008 International Conference on Convergence and Hybrid Information Technology*, 2008, pp. 459--465.

[30] Calvin Ko, “Execution monitoring of security-critical programs in distributed systems: A specification-based approach,” in *In Proceedings of the 1997 IEEE Symposium on Security and Privacy*, 1997, pp. 175--187.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Context-based Access Control Model for Smart Space

Alexander Smirnov
Laboratory of Computer Aided Integrated Systems
SPIIRAS
St.Petersburg, Russia
smir@iias.spb.su

Alexey Kashevnik
Laboratory of Computer Aided Integrated Systems
SPIIRAS
St.Petersburg, Russia
alexey@iias.spb.su

Nikolay Shilov
Laboratory of Computer Aided Integrated Systems
SPIIRAS
St.Petersburg, Russia
nick@iias.spb.su

Nikolay Teslya
Laboratory of Computer Aided Integrated Systems
SPIIRAS
St.Petersburg, Russia
teslya@iias.spb.su

Abstract: The smart space is an aggregation of devices, which can share their resources (information and services) and operate in coalitions. This nature of smart space enables of appearance of cyber conflicts between different smart space devices (or participants) which can have different goals and situation understanding but common information space for trusted cyber relationships. Therefore, one of the main security problems of coalition operations in smart spaces is a support of dynamic access control for decreasing cyber risks. In particular, a new access control model for accessing resources is needed. The model should describe the current situation via a context. Therefore, the research and development of the context-based access control mechanisms for smart space resources is an essential task.

The paper proposes a model of the context-based access control for the information shared in a smart space. Micro virtualization mechanisms represented by virtual private micro smart spaces are the basis for the model, which is built on the combination of the role-based and attribute-based access control models. Roles are assigned dynamically based on the smart space participant's trust level. The role separation allows simplifying policies and makes them human-readable and easy to configure. The trust level calculation is based on the participant's context, which includes identification attributes; location; current date; device type, etc. Also, three kinds of access control rules have been proposed. These rules are used to calculate the trust level, to assign roles based on the trust level, and to grant permissions to the smart space resources.

Keywords: *context, access control, smart space, smart-m3*

The paper presents research results partly developed under projects funded by grants #13-01-00286, #13-07-00271, and # 13-07-00336-a of the Russian Foundation for Basic Research; project #213 of the Russian Academy of Sciences; and project #2.2 of the Nano- & Information Technologies Branch of the Russian Academy of Sciences.

1. INTRODUCTION

The cyber physical environment (such as smart building, smart car, etc.) encapsulates both information and physical spaces and provides shared use of information and allows devices to join and leave the environment [1]. Thereby, smart space can be considered as a part of cyber physical environment, where acting, computational & information resources and virtual community members interact with each other as services to share information (Figure 1).

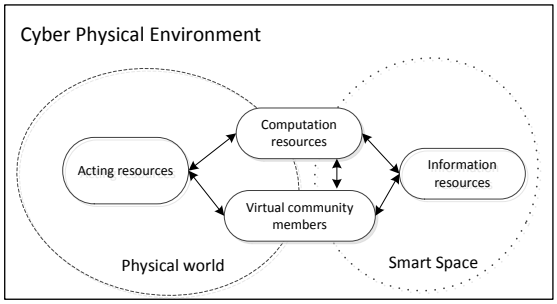

Figure 1. Smart space as a part of cyber physical environment

The smart space paradigm is a basis for the "Internet of Things" concept. This concept helps to make daily human life easier through automation of the routine actions. It allows multiple devices to provide coordinated support to users based on their preferences and current situation in the cyber physical environment (formalized by the context). The smart space is an evolution of the cloud computing concept, which combines the ideas of distributed computing and Semantic Web. In [2] the following features of the smart space are presented and compared with those of cloud computing (see Table I).

The following smart space features affect the information security: information distribution across space devices, ownership issues in information sharing, computational and information storage capacities are limited by those of space devices and services, user controlled information sharing, and large amount of applications and services operating in the smart space. The distribution of information in the smart space makes it difficult to provide access to resources using the existing classical access control models, such as discretionary access control (DAC), mandatory access control (MAC), and role-based access control (RBAC). Limited storage and computational capacities of space devices may be the object of denial of service (DoS) attacks. A large amount of unverified applications may be dangerous, because they may include unknown vulnerabilities or backdoors,

which may enable access to private information for unauthorized participants. In the cloud computing, solving similar problems is the responsibility of the provider. For the users, the cloud computing resources are provided as services, such as IaaS, PaaS, SaaS, etc. The access control system is included into the cloud service infrastructure and all client applications are verified for the potential vulnerabilities and backdoors by the provider.

Table I. Comparison of cloud computing and smart space paradigms

Cloud computing paradigm	**Smart Space paradigm**
Vendor Specific	User specific
Centralised to user (but distributed across provider servers)	Distributed across space devices
Requires network	Network not required continuously
Data privacy and ownership issues	Data is private but some ownership issues (sharing, citation, accreditation)
Unlimited computing resources Unlimited storage resources Cost	Computational and storage capacities are limited by those of space devices and services·(but can extend to clouds)
Not personal, vendor controlled	Personal, user controlled
Partial user responsibility·see licensing agreement, T&C's	User responsibility
Applications decided by vendor	Flexible applications
Interoperable within vendor's context	Interoperable

Both the smart space and the cloud computing paradigms facilitate coalition operations. Coalition operations are very likely to be based on a number of different, quasi-volunteered, vaguely organized groups of people, non-government organizations, institutions providing humanitarian aid and also army troops and official governmental initiatives [3]. In the proposed approach acting, computational & information resources and virtual community members are considered as coalition operation participants. Every participant is characterized by a context, which describes its activities in the smart space. The context is defined as any information that can be used to characterize the situation of an entity, where an entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [3]. For example, the context can include a type of the network which is using to access to the smart space, date and time of activity, company and/or community for which coalition belongs to, position of the participant in company, etc. The union

of contexts of all participants is the context of the corresponding cyber physical environment.

Considering the described above features of the smart space it can be concluded, that one of the main information security problems in coalition operations is a support of the dynamic access control. In particular, it is needed to develop a new access control model based on the coalition operation participant's context. It is proposed to use micro virtualization mechanisms including a virtual private micro smart space for this purpose. This space is a smart space available only for two participants used for private information sharing between them. It is named virtual and micro, because it is created and used only for information transfer between two participants. After that the space is destroyed.

The paper proposes a model of the context-based access control for the information shared in a smart space. The model is built based on the combination of the role-based and attribute-based access control (ABAC) models. Roles are assigned dynamically based on the user trust level and help to manage access to the resources. The trust level calculation is based on the participant's context, which includes attributes, identifying the user (user ID and public key); user location; current date; device, which requests the information, etc. A special smart space service has been proposed for this model. This service grants access to the resources for the smart space services guided by the access control policies. It is needed to note that the public information can be published to smart space and processed by all participants, but the private information is provided only for appropriate participants through the virtual private micro smart spaces when the corresponding access permissions are granted.

The rest of the paper is organized as follows. Section 2 describes the smart space platform features and presents requirements to the smart space security. Section 3 presents some existing works that introduces access control in Semantic Web and smart spaces based on the context of the participant. Section 4 introduces the proposed model and general scheme of the context-based access control for the smart space based on Smart-M3 platform. Section 5 presents main characteristics of the access control module, based on the presented approach.

2. SMART SPACE PLATFORM

Presented work is based on the open source Smart-M3 platform [5], [6], which provides implementation of the smart space methodology. The main difference of this platform compared with other existing solutions described in [8, 9, 10, 11] is that the Smart-M3 is an open source platform, it is accessible for downloading and

testing, supported by development community (last accessible version has been uploaded on the 04.02.2013), and supports modern mobile platforms (Android , Symbian, Harmattan).

This platform was first released at the NoTA conference in October 1, 2009 in San Jose. The Smart-M3 is being developed at ARTEMIS JU programme in SOFIA (smart objects for intelligent applications) [7] and in Finnish national DIEM (Device interoperability ecosystem) research projects. The Smart-M3 platform was applied in other European projects, for example, eHealth, eMobility.

The key idea of this platform is that the formed smart space is device, domain, and vendor independent. Smart-M3 assumes that devices and software entities can publish their embedded information for other devices and software entities through simple, shared information brokers. Information exchange in the smart space is implemented via HTTP using Uniform Resource Identifier (URI) [12]. Semantic Web technologies have been applied for decentralization purposes. In particular, ontologies are used to provide for semantic interoperability.

The Smart-M3 platform consists of two main parts: information agents and kernel (Figure 2) [5]. The kernel consists of two elements: Semantic Information Broker (SIB) and data storage. Information agents are software entities installed on the mobile devices of the smart space users. These agents interact with SIB through the Smart space Access Protocol (SSAP) [5]. The SIB is the access point for receiving the information to be stored, or retrieving the stored information. All this information is kept in the data storage as a graph that conforms to the rules of the Resource Description Framework (RDF) [13]. In accordance with these rules all information is described by triples “Subject - Predicate - Object”. More details about Smart-M3 can be found in [5].

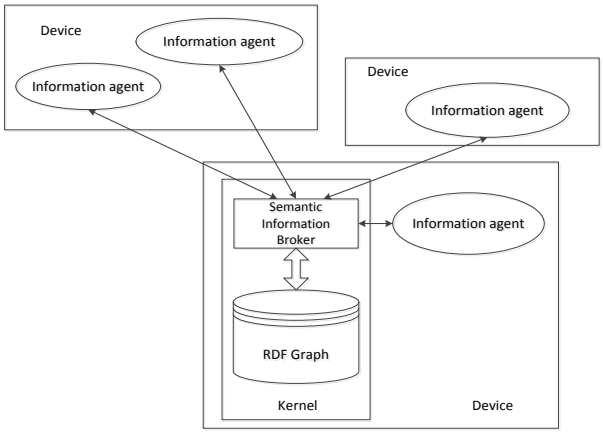

Figure 2. Smart-M3 reference model

Smart spaces extend computing to physical spaces, thus, information and physical security become interdependent. Moreover, the dynamism and interoperability that smart spaces advocate can give additional leverage for cyber-criminals, techno villains, and hackers by increasing opportunities to exploit vulnerabilities in the system without being observed. The following requirements to access control in the smart space have been developed based on security requirements proposed in [14]:

- The access control has to be multilevel, i.e. able to provide different levels of access control depending on predefined policies, current situation in smart space, and available resources.
- The access control model has to support an access control policy that is descriptive, well-defined, and flexible and easy configurable.
- Since a lot of the smart space services are placed on mobile devices, private information has to be transferred through the special secure information channel because smart space available for each participant and information encoding and decoding requires a significant amount of mobile device energy resources.
- Authentication should not be limited to authenticating human users, but rather it should be able to authenticate mobile devices that enter and leave the smart spaces, as well as applications and mobile code that can run within the smart spaces.

3. STATE-OF-THE-ART

J. Al-Muhtadi et al. [14] propose a mechanism that integrates context-awareness with automated reasoning to perform authentication and access control in space-based computing environments. The authors use this mechanism in the core service of the Gaia project, which provides the infrastructure for constructing smart spaces. The access control is based on the user's confidence value calculation. This value is calculated by the user's context (using simple probabilities, Bayesian probability, and fuzzy logic) and associated with different strengths of authentication which allows different activities in the smart space. Such approach is rather flexible and suitable for dynamic system like the smart spaces.

D. Kuhn et al. [15] propose to integrate two access control models: RBAC and ABAC. Three ways of integration are discussed: (i) with dynamic roles, where user's roles are set by attributes, (ii) attribute-centric, where roles are just attributes, not a set of permissions, (iii) role-centric, where attributes are added to constrain of RBAC. Constraint rules that incorporate attributes can only reduce permissions available to the user, but cannot expand them. The integration of roles and attributes

in one model enables to grant access depending on the current situation (context), for example, date and time or location of the user.

Extending this idea, A. Mohammad et al. [16] propose an ontology-based access control model. Usage of ontologies enables access level decisions and provides automated search of information related to the access control.

B. Carminati et al. [17], [18] propose an access control system based on the Semantic Web technologies for social networks. The approach presented in the paper enables granting access based not only on "friendship" relation with the resource owner but also on evaluation of the confidence level of the user. The authors propose policies for filtering available resources specified both by the rules and access control policies. With these policies, the person providing the access can control the information provided to the target users.

Semantic Web technologies are also used by Z. He et al. [19]. They propose access control based on the model of the RBAC using some of the ideas of attributive control, namely, the extending the RBAC with attributes of identity (certificates X.509 [20], public key, etc.). The authors propose the system architecture which implements the described model and discuss its implementation.

S. Verma et al. [21] compare RBAC and ABAC models with respect to the Semantic Web. The authors describe each model and analyze its strongest and weakest features. One of the advantages of the attribute-based access control model noticed by the authors is the support of context by attributes, which enables considering the current situation for granting the access permission.

K. Yudenok in [22] proposes an access control model for the smart spaces which are based on the Smart-M3 platform. The author describes algorithms of the identification, authorization and access control. For the identification and authorization the usage is of the Host Identity Protocol (HIP) [23] is proposed. For the access control the author proposes creation of the mapping between the smart space resources and virtual file system with further usage of the discretionary access control model for granting the access permissions. In this file system every term from the smart space is mapped to the file and the term's hierarchy is represented by the folder structure. A module which implements this model author embeds in the Smart-M3 platform.

The above models (except one described in [14]) are aimed to adaptation of existing access control models to the Semantic Web technologies specifications. Smart space combines the ideas of the distributed computing and Semantic Web, thus, its access control model should provide for interoperability, flexibility and simplicity of the access control rules, decentralization of the resources and access permission based on the user's context. Some of the above requirements are met by the model based

on the combination of the RBAC and ABAC models and by the scheme proposed by J. Al-Muhtadi et al. [14]. The model proposed in [22] can not provide support for the user's context and it is very difficult to configure because it uses the discretionary access control model. Moreover, mapping smart space resources to the virtual file system requires significant computational capacities and will certainly affect the system performance.

4. CONTEXT-BASED ACCESS CONTROL MODEL FOR THE SMART SPACE RESOURCES

As it has been noted, the following specific features of the smart space affect the information security: distribution across user devices, ownership issues computational and storage capacities are limited by those of space devices, and user controlled information sharing. The mechanisms addressing these issues are presented in (Table II).

Table II. Security mechanisms for the smart space security

Smart space specific features	**Security mechanisms**
Distribution across user devices	Share encoded information
Ownership issues	Context management
Computational and storage capacities are limited by those of space devices and services	Access control and context management
User controlled	Context management

All these mechanisms require introduction of the identification and authentication techniques for the services which request information. The participant is identified by the system when registering in the smart space. At this step the unique identifier is generated and saved in the Access Control Service (Figure 3). At the next steps this identifier is used as a part of the participant's context to authorize in the smart space. Additionally, the public and private keys are generated (for example using the RSA algorithm). These keys are needed for participant's authentication in the smart space and providing private information through the virtual private micro smart space.

The context of the smart space participant consists of the physical and virtual components. The physical component includes: geographical location of the device, date and time, type of a device. Using this information, the smart space services

can determine the current network type of the device, and time of the information access. It enables granting different access permissions from the corporate and public networks in different ways. The virtual component of the context includes software used by the participant for accessing the smart space, digital signature (the participant's identifier and the identifier encoded by the private key), and public key. This information enables authentication and authorization of the participant and provides encoding of the private data. For the web-community the participants add a social component to the context. This component includes, for example, position in the company, social relationships. The social component of the context enables granting access to the employees at different positions with the different trust levels, some private data can be shared only between friends, etc. All components of the context are collected and stored on the smart space devices. They become available upon the request of the Access Control Service.

Participant's context is used to define the trust levels assigned with its role. The role separation allows simplifying policies and makes them human-readable and easy to configure. Each component of the context is associated with the trust level. The level is represented by a number in the range [0, 1] and depends on the context of the current situation. For example, the trust level of "0.2" and "0.9" can be assigned for access from the public network and from the private network respectively. The logical function taking into account trust levels of all appropriate context components is used to assign a role to the participant. For example the role "trusted_participant" can be assigned only if the participant is authenticated, its network trust level is in the range [0.8, 1] and current time trust level is in the range [0.3, 1]. According to this, there are three sets of access control policy rules.

TrustValue rules are used to assign the numeric trust value to the context component. The examples of this rule type are the following:

TrustValue(network = public_network) = 0.2;

TrustValue(network = private_network) = 0.9;

TrustValue("08:00" < current_time < "17:00") = 0.6;

TrustValue(current_time > "17:00") = 0.1;

TrustValue(current_location "in set" [Russia, Estonia]) = 0.8….

TrustValue(current_location "in set" [China, North Korea]) = 0.1….

TrustValue(information_type = pdf _document) = 0.7

TrustValue(information_type = doc_document) = 0.3

These values are set by the access control service and based on the estimations of the access control service provider's experts according to the features of the particular smart space service.

Assign_role rules are used at the time of logging in or authentication. This set includes rules in the form of logic equations:

Assign_role(corresponding_author) = (TrustValue(network) ∈ (0.8, 1)) & (TrustValue(information_type) ∈ (0, 1)).

Assign_role(coauthor) = (TrustValue(network) ∈ (0, 1)) & (TrustValue(current_location ∈ (0.7, 0.9)) & (TrustValue(current_time) ∈ (0.3, 1)) & (TrustValue(information_type) ∈ (0, 1)).

Assign_role(reader) = (TrustValue(current_location) = 0.1 & TrustValue(information_type) ∈ (0.6, 0.8)).

Permissions rules are contain access control policies, which determine whether a participant with a certain role is allowed to access a particular resource type or not:

Permission(author) = "pdf_read","doc_read", "doc_write";

Permission(coauthor) = "pdf_read","doc_read", "doc_write";

Permission(reader) = "pdf_read".

General scheme of the request process is presented in Figure 3 and described below.

A device sends the request to access some private information (in the RDF notation) to the public smart space and subscribes to the corresponding response about the access granting:

device.smart_space.insert("participant_ID", "request", "resource");

device.smart_space.subscribe("participant_ID", "access_granted", None);

The smart space service accepts the request and calls the Access Control Service for the access permission.

service.smart_spase.insert("service_name", " participant _requested", "user_ID");

service.smart_spase.insert("service_name", "resource_type", "type");

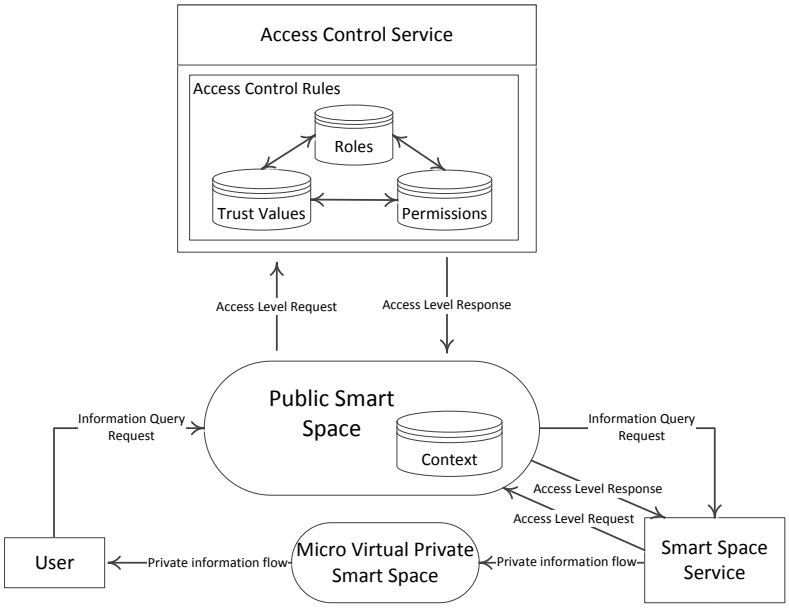

Figure 3. General scheme of context-based access to Smart space resources

The Access Control Service reads the participant's context and verifies its digital signature using the open key. If the signature is correct, the broker confirms that this user is authenticated and applies the rules from the access control policies to assign the role to the participant. The assess permission is granted based on the role of the participant and then is sent to the smart space service, which requested it.

access_control_service.smart_space.insert("Access Control Service", "participant", "participant_ID");

access_control_service.smart_space.insert("Access Control Service", "access", "granted" or "denied");

If the access to the resource is granted, the smart space service creates a virtual private micro smart space. The information requested by the participant is transferred to this private smart space. The connection information (space IP, space port and space name) is encrypted via the open participant's key and is sent to the public smart space.

service.smart_space.insert("participant_ID","access_granted", "Encrypted(IP,Port,Name)");

If the access was denied, the service sends the corresponding notification to the smart space participant.

service.smart_space.insert("participant_ID","access_granted","Denied");

Participant, who sends the information request, gets the notification via the subscription. If access is granted the participant decodes the encoded data with its private key and creates a connection to the specified virtual private micro smart space. When the requested information is transferred the virtual private micro smart space is destroyed.

5. TESTING OF A CONTEXT-BASED ACCESS CONTROL SERVICE FOR THE SMART SPACE RESOURCES

The basic ideas of context-based access control model for smart space resources have been implemented in access control service for ridesharing system [24]. This model has been evaluated by the following main parameters:

- Response time means the total time spent by the system, starting from the moment of sending the user's query and ending with answer of the service with obtaining information.
- Used RAM indicates total cost of the memory on one user's device user and Access Control Service.
- Network load indicates the number of calls to the smart space using SSAP protocol for response time.

A test result shows (Table III) that for information exchange between participant and Access Control Service is 20 ms.

Table III. The main parameters of the access control module working

Parameter	**Value**
Response time	20 ms
Used RAM	Client software additionally needs 1.1.Mb Access Control Service - 4.5 Мб
Network load	4 additional queries from the client software 3 queries from the Access Control Service

In the ridesharing system the response time with the Access Control Service is around 130 ms, and 110 ms without the service. This increase is reasonable for the system since only a few of operations requires access control permissions.

6. CONCLUSION

The paper proposes a context-based access control model for smart spaces. The Smart-M3 information platform is used as a smart space infrastructure for prototyping and testing of the proposed model. Usually in smart spaces the information sharing is implemented without any restrictions. However, some information in real applications can be private and should be shared in secure way. For this purpose a context-based access control model has been developed. It implements mechanisms based on the participant's context, which helps to reduce cyber risks arising from smart space features. The model proposes a service which makes access permission for the requested information using predefined rules. Implementing access control as a separated service that contains all smart space service permission makes it easier to configure rules for access control. All rules are human readable form and easy to set up in a fairly wide range. The rules are quite strict: non-compliance with at least one of the terms of appointment of the role will be assigned to a different role, more precisely satisfying for smart space participants' context. Computation resources used by Access Control Service are not so high and it is possible to optimize its usage. Usage of the context makes the model more flexible and appropriate for such systems.

REFERENCES

[1] M. Mohsin Saleemi, Natalia Diaz Rodriguez, Johan Lilius and Ivan Porres. «A Framework for Context-aware Applications for Smart spaces," *Smart spaces and Next Generation Wired/Wireless Networking. 11th Int. Conf., NEW2AN 2011, and 4th Conf., ruSMART 2011*, St. Petersburg, Russia, August, 2011, pp. 14-25.

[2] Ian Oliver. *"Clouds, Spaces and Information Sharing - A Future for the Semantic Web,"* 5th Conf. of Open Innovations Framework Program FRUCT. [Online]. Avaliable: http://www.fruct.org/sites/default/files/files/seminar5/s5_Fruct_IanOliver_29April2009.pdf.

[3] Smirnov, A., Pashkin, M., Chilov, N., Levashova, T., Krizhanovsky, A. «Agent-Based Intelligent Support to Coalition Operations: A Case Study of Health Service Logistics Support,» *Information & Security. An International Journal. IT in Coalition and Emergency Operations*. ProCon Ltd., Sofia, ISSN: 1311-1493, Volume 16, 2005, 41-61.

[4] Dey, A. K., Salber, D., and Abowd G. D. (2001). *"A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications,"* Context-Aware Computing, A Special Triple Issue of Human-Computer Interaction, 16. Retrieved August 13, 2012. [Online]. Available: http://www.cc.gatech.edu/fce/ctk/pubs/HCIJ16.pdf, Lawrence-Erlbaum.

[5] Honkola, J., Laine, H., Brown, R., Tyrkko, O.: *"Smart-M3 Information Sharing Platform,"* Proc. IEEE Symp. Computers and Communications (ISCC'10). IEEE Comp. Soc.; Jun. 2010, pp. 1041-1046.

[6] *Smart-M3 at Sourceforge*, 2012. [Online]. Available: http://sourceforge.net/projects/smart-m3

[7] Liuha, P., Lappeteläine, A., Juha-Pekka Soininen. "Smart Objects for Intelligent Applications," *ARTEMIS mag.*, no. 5, pp. 27-29. October 2009.

[8] Johanson, B., Fox, A., Hanrahan, P., Winograd, T., The Event Heap: An Enabling Infrastructure for Interactive Workspaces, Proceedings of the Fourth IEEE Workshop on Mobile Computing Systems and Applications, 2001.

[9] Xie, W., Shi, Y., Xu, G., Mao, Y., Smart Platform - A Software Infrastructure for Smart Space (SISS), Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, 429-434, 2002.

[10] Martin, D., Cheyer, A., Moran, D. The Open Agent Architecture: A framework for building distributed software systems. Applied Artificial Intelligence: An International Journal. 91-128, Vol.13, No.1-2. 1999.

[11] Coen, M., Phillips, B., Warshawsky, N., Weisman L., Peters, S., Finin, P., Meeting the Computational Needs of Intelligent Environments: The Metaglue system. In Proceedings of MANSE'99, Dublin, Ireland, 1999.

[12] Berners-Lee, T., Fielding, R., Masinter, L.: *RFC 3986 – Uniform Resource Identifier (URI): Generic Syntax*. [Online]. Available: http://tools.ietf.org/html/rfc3986.

[13] *Resource Description Framework (RDF).* W3C standard, 2004. [Online]. Available: http://www.w3.org/RDF/.

[14] Al-Muhtadi, J. Ranganathan, A. Campbell, R. Mickunas. «Cerberus: a context-aware security scheme for smart spaces,» *Pervasive Computing and Communications, 2003. (PerCom 2003). Proc. of the 1st IEEE Int. Conf.*, pp. 489-496, 23-26 March 2003

[15] D. R. Kuhn, E. J. Coyne, T. R. Weil. «Adding Attributes to Role-Based Access Control.» *IEEE Computer*, vol. 43, no. 6, pp. 79-81, 2010.

[16] A. Mohammad, G. Kanaan, T. Khdour, S. Bani-Ahmad, "Ontology-Based Access Control Model for Semantic Web service", *J. of Inform. And Computing Sci.*, vol. 6, No. 3, pp. 177-194, 2011.

[17] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu, B. Thuraisingham. "A Semantic Web Based Framework for Social Network Access Control," *Proc. of the 14th ACM symp. on Access control models and technologies*, pp. 177-186, 2009.

[18] B. Carminati, E. Ferrari, R. Heatherly, M. Kantarcioglu, B. Thuraisingham. «Semantic Web-based social network access control,» *Comp. & Security*, vol. 30, issues 2–3, pp. 108–115, March–May 2011.

[19] Z. He, L. Wu, H. Li, H. Lai, Z. Hong. ”Semantics-based Access Control Approach for Web Service,” *J. of Comp.*, vol. 6, no. 6, pp. 1152-1161, 2011.

[20] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, W. Polk. *“RFC 5280: Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile,”* [Online]. Available: http://tools.ietf.org/html/rfc5280

[21] S. Verma, M. Singh, S. Kumar. ”Comparative analysis of Role Base and Attribute Base Access Control Model in Semantic Web,” *Int. J. of Comput. Applicat.*, vol. 46, No.18, pp. 1-6, 2012.

[22] Kirill Yudenok. «Smart-M3 Security Model,». *Proc. of 11th Conf. of Open Innovations Assoc. FRUCT*, April, 23-27, St.Petersburg, Russia, pp.210–211, 2012

[23] R. Moskowitz, P. Nikander, P. Jokela, T. Henderson. *“RFC 5201: Host Identity Protocol,”* [Online]. Available: http://tools.ietf.org/html/rfc5201

[24] Alexander Smirnov, Nikolay Shilov, Alexey Kashevnik, Nikolay Teslya, «Smart Logistic Service for Dynamic Ridesharing,» *Internet of Things, Smart Spaces, and Next Generation Networking, 12th International Conf., NEW2AN 2012, and 5th Conf., ruSMART 2012*, St. Petersburg, Russia, August 27-29, 2012, pp. 140-151.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Information Sharing Models for Cooperative Cyber Defence

Jorge L. Hernandez-Ardieta
Cybersecurity Unit, Indra
Madrid, Spain
jlhardieta@indra.es

Juan E. Tapiador
COSEC Lab, Dept. of Computer Science
Universidad Carlos III de Madrid
Leganes, Madrid, Spain
jestevez@inf.uc3m.es

Guillermo Suarez-Tangil
COSEC Lab, Dept. of Computer Science
Universidad Carlos III de Madrid
Leganes, Madrid, Spain
guillermo.suarez.tangil@uc3m.es

Abstract: The globalisation and increasing complexity of modern cyber security operations have made it virtually impossible for any organisation to properly manage cyber threats and cyber incidents without leveraging various collaboration instruments with different partners and allies. This is especially relevant in certain areas of national security, like the protection of critical infrastructures, where the partnership amongst public and private sectors is paramount to adequately protect those infrastructures from emerging threats.

Over the last years consensus has emerged that sharing information about threats, actors, tactics and other cyber security information will play a central role in deploying an effective cooperative cyber defence. Near real-time information sharing has recently gained momentum as a means to redress the imbalance between defenders and attackers. In practical terms, the majority of current efforts in this area revolve around the idea of developing infrastructures and mechanisms that facilitate information sharing, notably through standardization of data formats and exchange protocols. While developing and deploying such an infrastructure is certainly essential to solve the problem of "how" to effectively share information, we believe that some key aspects still remain unaddressed, namely those related to deciding on "what" to share, "with whom", "when", as well as reasoning about the repercussions of sharing sensitive data.

In this paper, we argue that effective policies for near real-time information sharing must rely on, at least, two pillars. First, formal models to estimate the subjective value of the information shared should be developed. Second, trust/reputation models that consider the dynamic behaviour and changing factors of the sharing community have to be identified. For the latter, we propose to model information sharing communities as directed graphs, with nodes representing community members and edges modelling sharing relationships among them. Relevant properties of both nodes and edges are captured through attributes attached to each of them, which subsequently facilitate reasoning about particular data exchanges.

Keywords: *Cyber security, Cyber defence, Information sharing, Cooperation*

1. INTRODUCTION

Cyber conflicts are intensifying at a steady pace, both in prevalence, complexity and potential impact on individual organisations, nations, and the society at large. Besides, they have largely gone global, and globalisation has brought about a number of complications to cyber defence operations. On the one hand, interdependences amongst networks and information systems make localised and uncoordinated countermeasures rather ineffective, as they cannot ensure that no weak links are left in the chain. On the other hand, the attack landscape has evolved considerably in the last years, with a substantial rise in attacks involving a large number of distributed entities (e.g., botnets and DDoS) [1]; the emergence of markets where zero-day vulnerabilities are bought and sold on a regular basis [2]; or the advent of remarkably complex pieces of malware and cyber weapons [3][4][5], to name just a few. One major consequence of this new state of affairs in cyber security is a serious imbalance between the capabilities of attackers and defenders. As a matter of fact, at the moment it is virtually impossible for any organisation to prepare for and respond to cyber incidents without leveraging various collaboration instruments with other partners and allies. Examples abound in some areas of national security, such as the protection of critical infrastructures, where partnerships amongst public and private sectors are paramount to adequately mitigate risks and manage cyber attacks.

Over the last years consensus has emerged that sharing information about threats, actors, tactics and other cyber security information will be key to succeed in cyber defence. This sentiment has certainly not emerged from one day to the next, as proved, for example, by the efforts conducted over the last decade or so to categorise cyber security information, standardise data formats and exchange protocols, and develop infrastructures and mechanisms that facilitate sharing (see, e.g., [6] or the Cyber Defense Data Exchange and Collaboration Infrastructure (CDXI) being built by NATO [7]). While this is clearly essential to solve the problem of *how* to effectively share, some other relevant dimensions of the problem have received far less attention, notably those related to deciding on *what* to share, *with whom*, *when*, as well as reasoning about and adapting to the *repercussions* of sharing. One plausible cause for this is the fact that cyber security information sharing has largely been –and still is– a human-driven activity, where decisions are made one at a time and, in many cases, without an explicit elucidation of the rationale that motivates the decision. We believe, however, that addressing most of these questions will eventually become vital, particularly for scenarios where prompt responses to cyber threats are mandatory and, therefore, sharing decisions need to be made on a policy basis, in near real time, and with very little human involvement.

In this paper, we argue that the problem of sharing cyber security information can

be reformulated as one of *risk-based decision-making*. Thus, we seek procedures to answer questions such as: what are the benefits and the risks of sharing right now this piece of information with such party? Our choosing of this approach is motivated by two main facts:

a. On the one hand, taking an algorithmic approach on sharing will force us to quantify factors such as risks (and, implicitly, the value of information) and trust on sources and recipients. Even though these are challenging issues, a body of work in other contexts is slowly emerging. We believe that the cyber security community should adapt and adopt some of these techniques, particularly in scenarios where there is a need-to-share but the risks of doing so are not properly managed.

b. On the other hand, policies for information sharing must be elucidated and formally analysed. But policy making is a complex issue, and a given set of rules might well have unforeseen consequences, hence the need for automated techniques that provide optimal responses.

However, the ability to automatically making sharing decisions requires reasoning over formal structures (models) of most of the relevant elements involved, including the information itself, its value, the risks associated with disclosure (not only by us, but afterwards by partners receiving the information, either inadvertently or on purpose), our perception of the sharing community and the relationships among partners, etc.

In the remaining of this paper, we attempt to elucidate some of these questions, discuss challenges and identify areas where more efforts are needed. In Section 2, we review a number of research lines where problems similar to those appearing in this domain have been explored for a number of years. In Section 3 we formalise sharing communities as graphs and reformulate some key properties of partners and exchanges among them in graph-theoretical terms. This allows us to define sharing policies as algorithms running at each node. Section 4 develops the basis for a network-based model of cyber security information. Building upon the formats already developed, we point out the need for richer models where individual pieces of information can be annotated with labels reflecting, for example, our perception of its value or the trust we have on it being true. Moreover, connections among data need to be construed and made explicit, offering a view of an *information network* rather than a (more or less structured) list of items. In Section 5 we propose and discuss a risk-aware sharing algorithm. Section 6 concludes the paper by pointing out open problems and some lines of work that we are currently exploring.

2. RELATED WORK

In this section we review a number of research areas connected with the general problem of cyber security information sharing. In some cases, the connection is straightforward, although related to very concrete problems; in others, challenges similar to those appearing in this domain have been approached with techniques that might prove useful if conveniently adapted.

A. STRUCTURED MODELS OF CYBER SECURITY INFORMATION

As a discipline, cyber security deals with heterogeneous information related to the assets and configurations present in a system; the threats and tactics used by attackers; indicators of on-going incidents; countermeasures applied to mitigate risks; etc. Over the last decade, considerable efforts have been devoted to categorise such information and standardise data formats and exchange protocols, most notably through the Making Security Measurable (MSM) [6] initiative led by MITRE. Key aims of MSM include *"improving the measurability of security through registries of baseline security data, providing standardized languages as means for accurately communicating the information, defining proper usage, and helping establish community approaches for standardized processes."*[1]

MSM presents a comprehensive architecture for cyber security measurement and management, where current standards are grouped into processes and mapped to the different knowledge areas. Current MSM standards can be grouped into 6 major knowledge areas, each of which refers to a process (put in parentheses): **A**sset definition (inventory); **C**onfiguration guidance (analysis); **V**ulnerability alerts (analysis); **T**hreat alerts (analysis); risk/attack **I**ndicators (intrusion detection); and incident **R**eport (management). MSM standards and knowledge areas. Table I relates current MSM standards to these areas[2]:

[1] See http://measurablesecurity.mitre.org

[2] We refer the reader to *Appendix A for a description of MSM's acronyms, and to MSM's main website for further details.*

Table I. MSM standards and knowledge areas.

	CPE	OVAL	SWID	XCCDF	CCE	OCIL	CCSS	CVE	CWE	CVSS	CAPEC	CVRF	MAEC	CybOX	IndEX	STIX	IODEF	CPE	CEE	RID	RID-T	CYBEX	CWSS
A	•	•	•															•					
C		•		•	•	•	•																
V		•						•	•	•		•											
T								•	•	•	•		•	•	•	•	•		•	•	•		
I	•							•					•	•	•	•	•	•	•	•	•		
R	•	•			•			•	•	•			•		•	•	•			•	•	•	•

In the near future, it seems quite plausible that information sharing activities will be supported by infrastructures and mechanisms based on these standards, either in their current form or in subsequent revisions and developments.

B. COLLABORATIVE ATTACK DETECTION SYSTEMS

Many cyber attacks can only be detected by gathering and correlating evidences obtained at different locations [8]. In some cases, such evidences may come from sources unavailable to us and over which we have little control. This is, for example, the case of organisations that choose to share information about detected security events, possibly in near real-time, so as to minimise risk exposure or the impact of on-going cyber attacks. The so-called Collaborative Intrusion Detection Systems (CIDS) [9][1] constitute a clear example of the benefits that information sharing can offer to modern cyber defence capabilities. In principle, they have the potential to detect attacks that affect different Internet networks by correlating attack alerts. Besides, they also could reduce the costs involved in attack detection by sharing intrusion detection resources among networks.

CIDS consist of multiple distributed detection units logically organised in a network topology. In centralised systems, such as DIDS [10], DShield [11] and NSTAT [12], each sensor shares alerts with a central correlation unit. Hierarchical approaches (e.g., GrIDS [13], EMERALD [14] and DSOC [15]) attempt to address the scalability issues of centralised approaches by organising detection units into a tree-like topology. Finally, fully distributed approaches such as DOMINO [16] or the one proposed in [17] work in a P2P fashion, with nodes participating in a periodic exchange of information. We refer the reader to [1] for a more comprehensive account of existing CIDS technology.

Unfortunately, CIDS involving different partners are rare nowadays, as organisations are particularly reluctant to share sensitive information with almost any other actor. Apart from privacy issues, trust plays an important role in CIDS too. In most cases, the overall detection accuracy depends on all parties exhibiting honest behaviour,

particularly in terms of the trustworthiness of reported alerts. These issues are ignored or inadequately addressed in existing CIDS, in part because most of them were not conceived for an information sharing setting involving multiple and heterogeneous organisations.

C. TRUST AND REPUTATION MANAGEMENT SYSTEMS

In many fully distributed applications there is often a lack of a central authority in charge of monitoring users and reporting about their behaviour. In these scenarios, users often have to make decisions about who to trust for certain tasks (e.g. selecting routes in a MANET). Trust and reputation management systems have proliferated lately as a potential solution to this problem. Roughly speaking, these systems are based on the principle that users might quantify other users' behaviour by collecting and aggregating recommendations referring to past interactions with them. The interested reader can find surveys of trust systems in [18][19][20].

Possibly the central rationale underlying the utility of trust and reputation systems is that the behaviour exhibited by an entity in the past can be used to predict the expected outcome of future interactions. For cyber security information sharing scenarios, we anticipate that trust and reputation will play a key role in tasks such as deciding on whether to share some information with someone or not, or assessing the reliability and accuracy of pieces of data coming from questionable sources (e.g., using the aggregated value of previous data provided by one party as proxy for the a priori value of future information).

D. FLEXIBLE ACCESS CONTROL MODELS BASED ON RISK ESTIMATES

Imposing restrictions on sensitive information flows is a long-established problem in computer security. Traditional models of multi-level security, such as Bell-La Padula [21], deal with this problem by associating security clearances with subjects, security classifications with objects, and providing clear decision rules as to whether an access request should be granted or not. However, such mechanisms encode for a pre-determined calculation of risks and benefits, and in many modern situations preclude effective operations that can be justified on a risk basis when the specifics of the context are taken into account. The JASON Report [22] raised concerns about the inability of many organisations, particularly those in the national security and intelligence arena, to rapidly process, share and disseminate large quantities of sensitive information, in part due to the inflexibility of current access control models. Even worse, organisations are increasingly resorting to ad hoc means to surpass these restrictions, such as granting temporary authorisations for high-

sensitive objects or, as mentioned in [22], to follow the line of the old saying "it is better to ask for forgiveness rather than for permission."

Motivated by these issues, a number of works have proposed in the last years more flexible access control models based on an explicit quantification of the risk associated with every access request. For example, FuzzyMLS [23] replaces the classical binary allow/deny decision in BLP by a risk estimate that extends BLP rules to a continuous case. In [24], the model is extended to support uncertainty in security labels and clearances, and to account for the time dimension of sensitivity. Works in this area have proliferated in the last years, with a variety of proposals, including risk-based access control built on fuzzy inferences [25]; attribute-based risk-adaptive models [26]; role-and-risk based models [27]; benefit and risk access control [28]; and many others. Although the majority of these works explicitly target the particularities of information sharing settings, to the best of our knowledge none addresses cyber security information sharing.

3. A FORMAL MODEL OF INFORMATION SHARING COMMUNITIES

A. *COMMUNITY STRUCTURE*

We represent an information sharing community as a weighted directed graph (digraph) $G = (V, E)$, where V is the set of nodes or vertices that represent the entities that are member of the community, while E is the set of the edges or links that represent the information flows permitted within the community. For each edge $e = (u,v) = uv$, we denote by $e^{-1} = vu = (v,u)$ its inverse, if it exists.

In information sharing terminology, u corresponds to an originator of information, while v is a recipient of information. Therefore, edges restrict not only the members that may share information amongst them but also who distributes the information within the community and with whom. Please note that the originator does not necessarily correspond to the source of the information. The latter is the entity that produces an item of information. As the source does not need to be a member of the community, for simplicity we do not consider them in our model. Thus, an originator u that shares information with a recipient v can transmit information produced on its own (i.e. u is the source as well), forward information received from other nodes (i.e. u behaves as a forwarder of information), or both.

A graph that permits multiple edges between nodes is called a multigraph. We generalize the representation given above for an information sharing community to formally include the multigraph notation:

$G= (V, E,\Psi)$

where $E = \{e_1, e_2, ..., e_m\}$ is a set of symbols representing the edges of the graph, and $\Psi: E \rightarrow E(V)$ is a function that attaches an ordered pair of nodes to each $e \in E$: $\Psi(e) = uv$, u and v being nodes.

In our digraph (information sharing domain), if $\Psi(e_1) = \Psi(e_2)$, then $e_1 = e_2$. As a digraph has directed edges, two different edges that have the same ends (e.g. uv, vu) must have a different predecessor node (originator). In other words, the direction of each edge must be opposite to the other. This restriction conditions the structure of the multigraph, as there cannot be two equally directed edges between two nodes u and v. We do not consider loops (edges with ends uv / $u = v$) either, as sharing information with oneself is given per se.

It should be noted that the graph representing an information sharing community may contain cycles, and this will depend solely on the community structure.

B. LINKS BETWEEN NODES

Definition 1. Let $e_i = u_i u_{i+1} \in E$ *for* $i \in [1,k]$. The sequence $W = e_1 e_2 e_3 ... e_k$ is a walk of length k from u_1 to u_{k+1}. It should be noted that e_i and e_{i+1} must be adjacent $\forall\, i \in [1, k-1]$. For simplicity, we write $W: u_1 \rightarrow u_2 \rightarrow u_3 \rightarrow ... \rightarrow u_k \rightarrow u_{k+1}$ or $W: u_1 \sim^n u_{k+1}$ to represent a walk of length n from u_1 to u_{k+1}.

Definition 2. A walk $W = e_1 e_2 e_3 ... e_k: u \sim v$ is a directed walk, if $e_k \in E$, $\forall\, i \in [1,k]$, $u \in e_1$ is the originator of information and $v \in e_k$ is the latest recipient of the information. A directed path $P: u_1 \rightarrow u_k$ is a directed walk where $u_i \neq u_j$, $\forall\, i \neq j$. A directed cycle is a directed path where $u_1 = u_{k+1}$.

We consider that one of the next four possibilities can occur in an information sharing community between any two indistinct nodes *(u, v)*:

(1) there is no directed path that connects u and v, and therefore they cannot share information between them, neither directly nor indirectly.

(2) there is no edge that connects u and v, that is, there is no direct connection between them. However, they could share information using a directed path that connects them indirectly (e.g. $W: u \rightarrow w \rightarrow v$).

(3) there is a directed edge from u to v or from v to u.

(4) there are two directed edges that connect both nodes, being u and v both originators and recipients of information.

Definition 3. Two nodes are unconnected if there is no directed edge or path that connects both nodes. Two nodes are strongly connected (adjacent) if there is a directed edge that connects them independently of the edge direction. Finally, two nodes are weakly connected if there is a directed path that connects them but where there is no directed edge between them.

C. TYPES OF NODES

We classify the nodes in a community using the indegree ($deg^-(u)$) and outdegree ($deg^+(u)$) properties of a node, which specify the number of head and tail endpoints, respectively, adjacent to a node. Formally:

$$deg^-(u) = |\{e \in G / e = xu\}|$$

$$deg^+(u) = |\{e \in G / e = ux\}|$$

In graph theory, the node with $deg^-(u)$ equals zero is called a source, while a node with $deg^+(u)$ equals zero is called a sink. In our information sharing community scenario, we identify three types of nodes, two of them according to the balance between their indegrees and outdegrees. Let Ω be the difference between $deg^-(u)$ and $deg^+(u)$ of a node n.

$$\Omega(u) = deg^-(u) - deg^+(u)$$

Definition 4. We say that a node $u \in V$ is a distributor if $\Omega(u) << 0$, and $\Omega(u) \in N^-$ (negative integers). A distributor is expected to receive information from a few originators and provide information to many recipients.

Definition 5. We say that a node $u \in V$ is a collector if $\Omega(u) >> 0$, and $\Omega(u) \in N^+$ (positive integers). A collector is expected to receive information from many originators and provide information to few recipients. When $deg^+(u) = 0$ (sink), the information received by the collector is not further shared with other community members.

On the other hand, we use the betweenness centrality property to define the third type of node in a community. The betweenness quantifies the number of times a node is part of the shortest path between two other nodes. This provides a measure of the relevance of that entity within a community in terms of presence in information sharing routes. Given a connected graph G with a weight function α: $E \rightarrow N$, the shortest path between two nodes u and $v \in G$ is the path P with the minimum total

weighted distance between u and v:

$$d^{\alpha}{}_{G}(u, v) = min\{\sum_{e \in P} \alpha(e) / P: u \sim v\}$$

Thus, the betweennes centrality $Cb(u)$ of a node u is given by

$$Cb(u) = \sum_{r \neq k \neq v} \sigma_{r,k}(u) / \sigma_{r,k}$$

where $\sigma_{r,k}(u)$ is the number of shortest paths between any two nodes r and $k \in V$ that pass through u, and $\sigma_{r,k}$ is the total number of shortest paths between any two nodes r and k ∈ V. For example, in a centralized sharing approach (e.g. an Information Sharing and Analysis Center), the central node has values of betweenness much higher than any other node of the community, while sources and sinks are expected to have a zero betweenness centrality.

Definition 6. We say that a node $u \in V$ is a bridge if its betweenness centrality $Cb(u)$ has a value higher than the arithmetic mean of all nodes of the graph G.

$$u \text{ is bridge} \leftrightarrow Cb(u) > Cb(v)/|V|, \forall v \in V$$

Figure 1 exemplifies the properties described above.

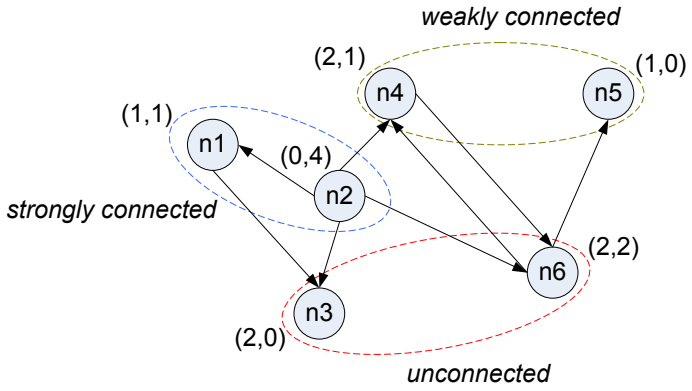

Figure 1. A graph example. The pair of numbers associated with each node indicate the indegree and outdegree values. Node n2 is a distributor, while nodes n3 and n5 are collectors. Node n6 is the only node that is part of a shortest path between two other nodes of the graph (i.e. path between n4 and n5), and thus, is the only one that complies with the bridge definition given above.

D. RISK ASSESSMENT FUNCTION

The edges of the graph have a weight that, in our case, is a richer concept than traditional edge weights. For cyber security information sharing communities, we add metadata to each edge that represents a function that calculates the risk level to which the originator of the information is exposed if certain piece of information is shared with certain strongly connected recipient.

In our scenario, we propose a simple formula for defining the risk to which a node is exposed when sharing information. It basically depends upon two well-differentiated factors that have played a key role in well-recognized risk assessment methodologies:

- First, the value of the information shared, and thus the impact caused on the entity (originator node) in the case that such information is accessed by unauthorised entities. We address this point later in Section 4, including those cases where the value of information varies over time [24].
- Second, the probability that such information is accessed by unauthorised entities (any node in the graph).

For the impact, its value strongly depends on the particularities of the originator, the information shared, and other contextual information.

It should be noted that the quantification of both the impact and the probability values is a difficult task whose precise estimation is generally impossible, as the knowledge required to do so is incomplete. For instance, the calculation of the probability of occurrence may be improved using intelligence obtained from the terrain (e.g. OSINT, HUMINT, trends, market analysis, etc.), but, unfortunately, we usually end up with a rough estimation that will be substantially different to the underlying reality. Notwithstanding, this problem is out of the scope of the present paper, and thus we do not question the trustworthiness of these values when used in our formulae.

Definition 7. Let $u \in V$, and δ be a piece of information originated by u. We define $I(u, \delta) \in [0, 1]$ as a measure of the impact on u caused by an unauthorised access to δ.

On the other hand, a node u may not be able to estimate the second factor above (the probability), especially when the recipient can further share this information with third nodes in the graph (and so forth). Instead, we use the trust that node u (originator) has in node v (recipient). Intuitively, a higher level of trust should pose a lower probably of unauthorised access if such trust has been adequately assessed based on empirical data or any other information derived from past experiences.

Specific mechanisms for trust computation fall out of the scope of the paper, though some proposals can be found in Section 2 C.

If *v* can share this information with a third-level authorised node *w*, then the risk level should consider the probability that *w* leaks the information to an unauthorised entity as well. For this case, the risk value also depends on the trust that *u* has on *w*. If it cannot be directly inferred by *u*, then it can be calculated using indirect means, such as by combining the trust that *u* has in *v* with the trust that *v* has in *w*.

Definition 8. Let $u, v \in V$. We define $T(u,v)$ as the function that calculates the trust that *u* has on *v*, denoted by $T(u,v) = \{t / t \in [0, 1]\}$.

In conclusion, we formalize the risk function β in a directed multigraph as follows.

$$\beta(u,v,\delta) = I(u, \delta) \bullet (1 - \Pi_{s \in S} T(u,s))$$

$$\text{where } S \subseteq V / s \in P(u_i, u_{i+k}), \text{ with } u_i = v, \forall s \in S.$$

In a nutshell, the risk value is computed multiplying the impact by the probability that any node weakly connected to the originator *u* through *v* discloses the information to an unauthorised entity. Please note that the resultant probability is expressed as 1 – the probability that no node discloses the information, and also considers the trust value of *u* on *v*.

Next, the risk function is generalized in order to calculate the risk to which a node is exposed in the case that it shares the information will all its strongly connected nodes:

$$\beta(u,\delta) = I(u, \delta) \bullet (1 - \Pi_{i = v \in V} A_{u,i} \bullet \Pi_{s \in S} T(u,s))$$

where $A_{u,i}$ is the adjacency matrix for *u*, and S the set of nodes reachable through a directed path starting in *i*, for all *i* strongly connected node with *u*.

$$A_{u,i} = \begin{cases} 1 & if\ nodes\ u\ and\ i\ are\ strongly\ connected \\ 0 & otherwise \end{cases}$$

The function above is applicable as long as $|A_{u,i}| > 0$.

For the risk value computation we assume that the nodes do not collude, and thus, the probability is calculated unconditionally.

From the formula $\beta(u,v,\delta)$ above, it can be easily inferred that the minimum and

maximum risk values for an originator *u* that shares a piece of information δ with a strongly connected recipient *v* correspond to 0 (i.e. both *v* and all weakly connected nodes are fully trusted, that is, $T(u,s) = 1 \; \forall s \in S$) and $I(u, \delta)$ (i.e. at least one of the nodes is fully distrusted, that is, $T(u,s_i) = 0$), respectively. Also, it can be observed that the risk value increases with the number of recipients, unless these are fully trusted.

It should be noted that the risk value is a dynamic value that has to be updated (re-calculated) when any of the forming factors changes, such as the impact value or the trust in any of the related nodes.

4. FORMAL MODELS OF CYBER SECURITY INFORMATION

We represent the knowledge of information security as a weighted graph $G = (V, E)$, where V is the set of nodes or vertices that represent pieces of well structured information from an *element of knowledge*, while E is a relation set given by $E \subseteq V \times V$, in which each of whose members is a pair (i.e., edge or link) representing a relationship between the different pieces of information.

Definition 9. Elements of knowledge κ define a set of ontologies $\kappa = \{\kappa_1, ..., \kappa_n\}$, where each κ_n encodes some knowledge about different domains of information security.

In our domain, nodes represent heterogeneous pieces of information from elements of the set κ. Thus, we represent this variety of information by using vertex-labelled graphs.

Definition 10. Let $\rho^{k_n} = \left\{\rho_1^{k_n}, ..., \rho_m^{k_n}\right\}$ be a set of properties of a specific κ_n. We say that two nodes *u* and *v* are related, denoted $\sigma(u, v)$, if there is a pair of properties (ρ_i, ρ_j) such that $\int(\rho_i^{k_u})$ is equal to $\int(\rho_j^{k_v})$. Formally:

$$\sigma(u,v) = true \leftrightarrow \exists(\rho_i, \rho_j)/\forall\rho_i \in \rho^{k_u}, \forall\rho_j \in \rho^{k_v} : \int(\rho_i^{k_u}) = \int(\rho_j^{k_v})$$

For simplicity, we denote the property throughout κ_u is associated with κ_v as $p^{k_{u,v}}$, and we denote the piece of information that satisfies $\sigma(u, v)$ as $\int(\rho^{k_{u,v}})$.

In our domain, an edge is represented as $e=(u, v, p)$, where *u* and *v* are two pieces of information that satisfy $\sigma(u, v)$, and *p* is a label representing the shared property $p^{k_{u,v}}$. We represent each shared property using edge-labelled graphs.

Henceforth, each node in this model corresponds to an element of the information security knowledge, and each edge corresponds to a relation between those elements.

Figure 2. shows an exemplification of an information knowledge graph structure, where assets are connected to each other, as well as to specific configurations. Additionally, exploits can target different vulnerable configurations held by the several assets. For instance, asset A_3 represents an Oracle Java runtime environment application installed on A_1 –a Red Hat Linux server. Here, certain configuration C_1 allows the asset to execute Java applet scripts (CCE-10083-4).

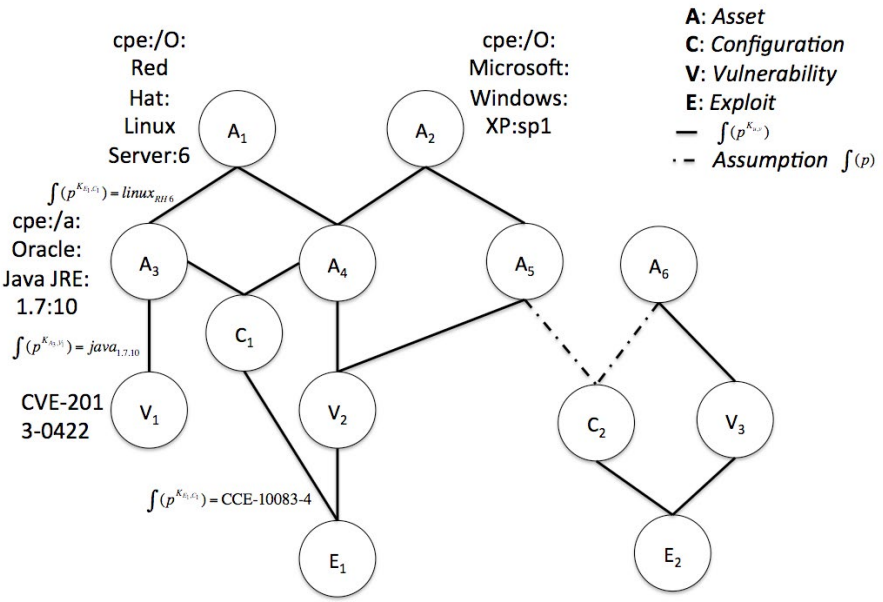

Figure 2. A graph example. Different elements of knowledge (CVE, CPE, CCE) are related by a number of pieces of information.

A. INFORMATION VALUE AND REASONING OVER GRAPH STRUCTURES

We present a quantification of information value based on the relation between different elements of knowledge. More precisely, we position that the value of the information directly depends on the information already owned by a community member, how this information is structured and to what extent is related to other pieces of information. In this regard, identifying missing information can also contribute to the quantification of information value. Furthermore, there are other key attributes for quantifying the information value, such as its relevance, timeliness, and accuracy, to name a few.

Definition 11. Let *c* be the *cost* of a piece of information $\int(\rho^{k_{u,v}})$, *t* the time window where such cost holds (assuming that cost decays over time), *a* the degree of reliability that the piece of information is *accurate*, and *r* how *relevant* the information is for an organization. We define the *value of information* of a node *V* as:

$$Vol_v(c,t,a,r) = \omega \cdot \psi(c_v,t_v,a_v,r_v) + (1-\omega) \cdot \sum_{u \in V : \sigma(v,u)=true} Vol(c_u,t_u,a_u,r_u)$$

where ψ determines the subjective value of a vertex *V* for a given community and the second term factors in the aggregated value of adjacent nodes. By *cost* we mean here the amount of resources (economic, computational, etc.) needed to acquire and process the information. In our model we assume that there are markets where such information can be acquired, and that costs can be known. On the contrary, the *value* is specific to each party and will very likely vary over time. As an example, we suggest the next measure for the subjective value of information:

$$\psi(c_v,t_v,a_v,r_v) = (c_v \cdot a_v) \cdot e^{-k \cdot r_v \cdot \left(\frac{t}{t_v}\right)}$$

where the first term represents the value of the information and the second an exponential decay function over time, *k* being a decay constant weighted by *r*. In other words, the value of relevant pieces of information will be exponentially bigger than non-relevant pieces and it will decay slower over *t*. Note, too, that the relevance may also serve to modulate the risk of disclosure of the information.

Here, the relationship among different nodes could be expressed in a more complex way. For instance, some relations are often due to **causality**, and some others are subject to a perception error, i.e. **uncertainty**. In this regard, graph described on Figure 2. shows how easy could be reasoning using a graph structure. For instance, if we knew that a given asset A_5 is from the same vendor as A_6, and we knew that the latter have an exploitable configuration, we could reason that the same exploit might be applied to A_5 with a certain probability.

Furthermore, graph structures also allow us to establish possible paths from a type of node, e.g. an asset, to all other nodes of the same or different type, e.g. exploits. In this regard, different conclusions might be extracted depending on the type of nodes through the path. On the one hand, if there were a direct connection between nodes of the same type such as A_1 and A_3 in Figure 2. , compromising A_3 would also compromise A_1 and A_4 –as asset A_1 is the operating system executing application A_3 and A_4. On the other hand, if there were a connection between nodes of different type such as A_3 and E_1 through a node of types vulnerability and configuration (V_2 and C_1 in the aforementioned example), we could conclude that A_3 could be compromised using E_1.

Thus, reasoning over heterogeneous graph structures in a complex task, which requires context-based reasoning, i.e., type of node, length of the walk, etc. Note that this section intends to introduce the concept of information value over graph structures, and we refer the reader to forthcoming publications for a deeper treatment of reasoning about cyber security information and its value.

5. EXAMPLE: AN INFORMATION SHARING ALGORITHM

In this section we present an algorithm that aims at achieving the need-to-share concept [29], so as to maximise the information sharing within a community while the risk value for the originator of information is kept below an established threshold (i.e. an acceptable risk level). In the next Subsection we first describe the general aspects behind the algorithm, and in the subsequent, a running example to illustrate its behaviour.

We do not claim that this algorithm is the only one applicable to the information sharing scenario. Actually, there are a number of approaches, where the most appropriate one should be selected depending on the particularities of the community, the policies applicable to the originator, the information to share at each moment, and other contextual information. For instance, the same node may decide to apply a different algorithm for different pieces of information depending on their level of classification. Or the same node may select a different algorithm for the same piece of information at different moments (e.g. a less conservative approach may be followed in a crisis situation).

A. *OVERVIEW*

The algorithm is a greedy algorithm in the sense that it follows the problem solving heuristic of making the locally optimal choice at each stage. The problem to solve at each stage corresponds to whether or not sharing a certain piece of information with an adjacent node depending on the accumulated risk value and the threshold established by the originator.

The algorithm consists of two well-differentiated phases. In the first one, that we call *Decision Phase*, the originator performs a simulation of how the information should be shared across the community in order to keep the accumulated risk value that results from the subsequent sharing actions below the desired threshold. At the end of this phase the originator is able to conclude what nodes of the graph are authorised to access the information.

In the second phase, named *Sharing Phase*, the sharing process itself is undertaken, started by the originator, and by which the pertinent information that allows each sharing node to know who are the authorised nodes amongst its adjacent ones is also transmitted.

During the *Decision Phase*, the simulation orders the adjacent nodes of a certain sharing node n_i by their trust value from higher to lower, discarding those in whom the originator fully distrusts ($T(u, x_i) = 0$) as well as those that have been already marked as authorised node. Then, it calculates the accumulated risk value $\beta_A(u,v,\delta)$, being u the originator and v the node adjacent to u through which n_i has been reached. β_A formula considers the trust values of every node that has already been marked as authorised nodes, plus the adjacent nodes of n_i. If the resultant value of β_A is greater than the established threshold, then the simulator discards the last adjacent node of the ordered list, and recalculates β_A. The analysis is iterated until the obtained β_A is below the threshold. The adjacent nodes that remain in the list when this condition is satisfied are marked as authorised nodes.

The simulation stops analysing a certain sharing node if any of the following conditions is met:

- The sharing node has an outdegree equals zero.
- The ordered list is empty or, after discarding the adjacent nodes during the β_A calculation, there is no one left. This means that the sharing node will not be authorised to share the information with any of its adjacent nodes.
- The sharing node already received that piece of information (the algorithm considers cyclic graphs).

At the end of the *Decision Phase*, a subset of nodes $V' \subseteq V$ and edges $E' \subseteq E$ will have been selected. The resultant subgraph $G' = (V', E')$ is an acyclic directed graph (i.e. tree) where the root node is the originator, the rest of the nodes are those authorised to access the information and the edges represent the sharing links between the nodes. In principle, the search strategy of the *Decision Phase* could be configured to follow either a Breadth- First -Search (BFS) or a Depth-First-Search (DFS) [30] as both approaches have the same time (O|E|) and space (O|V|) bounds. However, the vertex ordering produced in BFS (i.e. the order in which the vertices are explored) better reproduces the behaviour expected in an information sharing community. In these communities, each node is strongly connected to other nodes in which it explicitly trusts. Therefore, it is expected that any originator will preferably share the information with these nodes in the first instance, rather than leveraging on weakly connected nodes the increase of the accumulated risk value β_A.

The difference between the sharing process for the originator and any other node is that, for the latter, they can share the information (as long as the threshold condition

is satisfied) with adjacent nodes with which the originating node has no explicit trust calculated. For these nodes, the algorithm uses an indirect trust computation using the path of nodes from the originator to those nodes. For instance, the indirect trust computation for node *u* on a node *x* weakly connected through the walk *W: u* → *v* → *w* → *x* is as follows:

$$T(u,x) = T(u,v) \bullet T(v,w) \bullet T(w,x)$$

This approach helps maximizing the information sharing by permitting the sharing with unknown nodes as long as the threshold is not exceeded.

B. AN EXAMPLE

In this section we show the application of the *Decision Phase* to the graph example shown in Figure 3. following a BFS approach and considering the table of trust shown in Table II. In our example, the node n1, as the originator, wishes to share some piece of cyber security information with the information sharing community.

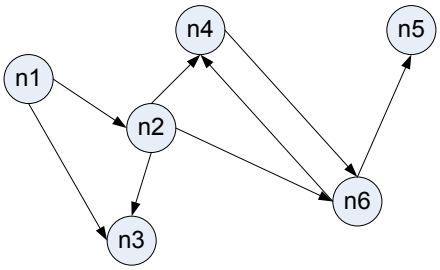

Figure 3. Graph G.

Table II. Table of trust for graph G. A value of 0 means no trust at all; a value within the range (0,1) means relative trust; and 1 means full trust. If a node has no explicit (dis)trust in some other node, then no value is indicated. No edge appears in the graph between those nodes if there is an explicit distrust or when no explicit trust exists.

$T_G(n_i)$	n_1	n_2	n_3	n_4	n_5	n_6
n_1	1.00	0.75	1.00	-	0.00	-
n_2	-	1.00	0.05	0.6	-	0.35
n_3	-	-	1.00	-	-	-
n_4	-	-	-	1.00	-	0.80
n_5	-	-	-	-	1.00	-
n_6	-	-	-	0.25	0.90	1.00

We define $\varphi_G(u,\delta)$ as the risk threshold, that is, the maximum risk value acceptable by u for the piece of information δ within the information sharing community G.

The initial values for our example are the following:

$\varphi_G(n_1,\delta)$: 0.7 $I(n_1, \delta)$: 0.3 $AuthNodes:=\{\}$

The initial values should be result of a risk analysis carried out by the originator, and by which the maximum tolerable risk $\varphi_G(u,\delta)$ and the impact $I(n_1, \delta)$ for the piece of information δ can be estimated. In this example both the initial values and the table of trust shown in Table II have been selected to serve for illustrative purposes only.

It is worth mentioning that in many cases the originator can exert some control over the impact –and, therefore, over the maximum tolerable risk– by selectively removing sensitive parts of the information to be shared. In scenarios other than cyber security information sharing this is commonly achieved by anonymising data, e.g. by removing or aggregating pieces of information.

We next proceed with the example:

1. Analysis of sharing node n_1 (originator)

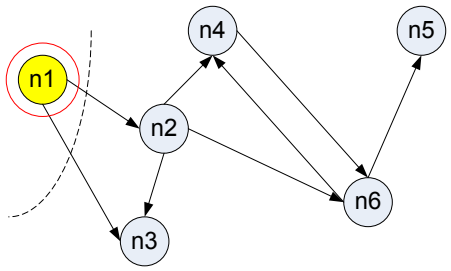

Figure 4. Sharing process for originator n_1

Ordered list of adjacent nodes $(n_1) := \{n_3, n_2\}$

Calculate accumulated risk level:

$$\beta_A(n_1,n_1,\delta) = I(n_1, \delta) \bullet (1 - T(n_1, n_1) \bullet T(n_1, n_3) \bullet T(n_1, n_2)) = 0.3 \bullet (1 - 1 \bullet 1 \bullet 0.75) = 0.075$$

If $0.075 < \varphi_G(n_1,\delta)$ then update *AuthNodes* and share with nodes remaining in the ordered list:

$$AuthNode\sigma := \{n_3,n_2\}$$

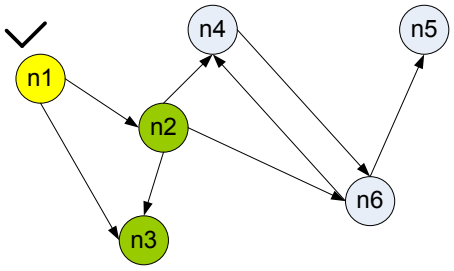

Figure 5. Result of sharing with n_2 and n_3[3]

[3] For clarity purposes, we mark the nodes that have already been analysed (n_2 in this case).

2. Analysis of sharing node n_3

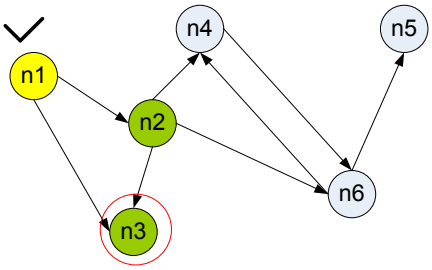

Figure 6. Sharing process for n_3

Stop condition applies: $deg^{+}(n_3) = 0$

3. Analysis of sharing node n_2

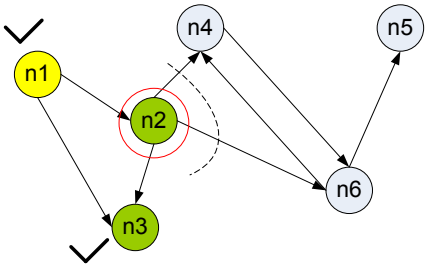

Figure 7. Sharing process for n2

Ordered list of adjacent nodes $(n_2) := \{n_4, n_6\}$[4]

Calculate accumulated risk level:

$$\beta_A(n_1,n_2,\delta) = I(n_1, \delta) \bullet [1 - T(n_1, n_1) \bullet T(n_1, n_3) \bullet T(n_1, n_2) \bullet T(n_1, n_4) \bullet T(n_1, n_6)]$$ [5] $$= I(n_1, \delta) \bullet [1 \quad T(n_1, n_1) \bullet T(n_1, n_3) \bullet T(n_1, n_2) \bullet \underline{(T(n_1, n_2) \bullet T(n_2, n_4))} \bullet \underline{(T(n_1, n_2) \bullet T(n_2, n_6))}] = 0.3 \bullet [1 - 1 \bullet 1 \bullet 0.75 \bullet (0.75 \bullet 0.6) \bullet (0.75 \bullet 0.35)] = 0.273$$

If $0.273 < \varphi_G(n_1,\delta)$ then update *AuthNodes* and share with nodes remaining in the ordered list:

[4] Please note that n_3 is not included as it has already been marked as authorised.

[5] We underline the new factors that are incorporated to the B_A formula.

$AuthNode\sigma := \{n_3, n_2, n_4, n_6\}$

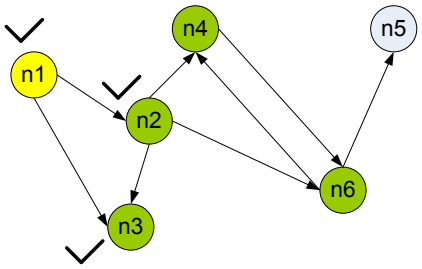

Figure 8. Result of sharing with n_4 and n_6

4. Analysis of sharing node n_4

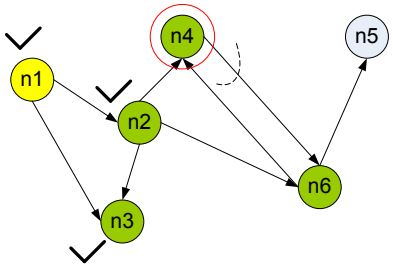

Figure 9. Sharing process for n_4

Ordered list of adjacent nodes $(n_4) := \{\}$[6]

Stop condition applies: list is empty.

[6] In this case, the ordered list is empty as n_6, the single node adjacent to n_4, has already been marked as authorised.

5. Analysis of sharing node n_6

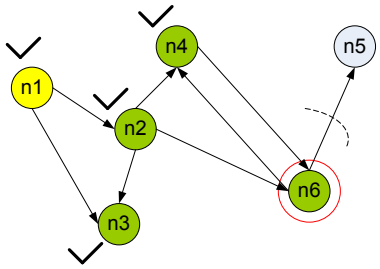

Figure 10. Sharing process for n_6

Stop condition applies: list is empty[7].

After the application of the *Decision Phase*, the list of authorised nodes to which the information can be shared is $\{n_3,n_2,n_4,n_6\}$.

[7] In this case, the ordered list is empty as n_5, the single node adjacent to n_6, is fully distrusted by the origin n_1.

6. CONCLUSIONS, CHALLENGES AND FUTURE WORK

Information sharing will be central to cooperation activities in cyber security operations. But the benefits derived from being a member of an information sharing community are not always perceived in the same way by different entities. Furthermore, organisations might well be reluctant to share sensitive information with partners whose trustworthiness is unclear and/or when the repercussions of sharing are not properly understood. These and other factors have been already identified as major inhibitors for the proliferation of information sharing communities and deterrents to members' active participation when being part of a community. In this paper, we shed some light on a few of these questions and point out the need to attack the problem from a formal perspective. In particular, we suggest analysing the *topology* of sharing by modelling as graphs both the community and the information network. In doing so, we can leverage a number of tools from a number of disciplines –notably graph theory, complex networks, and social network analysis– to study relevant aspects of the problem.

Due to space reasons, in this paper we have not given a deep account of any of these problems. Rather, our aim is to raise awareness about the benefits that such a perspective could bring to information sharing in cyber security. Our formal treatment of the information network and sharing communities, including the sharing algorithm discussed above, attempts to be merely illustrative of the potential that this approach could yield. In fact, this issue have received much attention in other contexts where information sharing is essential for agents that cooperate towards a common goal. For example, Zhu et al. present in [31] an algorithm to share information among a set of agents that operate in an ad hoc fashion. Each agent must decide whether to broadcast sensed and/or received information to neighbouring members. The approach is similar to ours in the sense that the problem is couched as one of optimal decision-making. However, the focus in [32] is on maximising sharing and minimising communication cost, whereas in cyber security risk factors are paramount.

We are currently exploring in greater depth several of the work areas discussed throughout this paper. Specifically:

- Some metrics and techniques well known in complex and social network analysis can easily be reinterpreted in this domain. For example, *information centrality* measures the efficiency of a network in delivering information. Similarly, the *betweenness centrality* of a node measures the importance of node in a network in terms of how many shortest paths between any other pair of nodes pass through it. Both measures, together with other centrality

quantities, can be valuable in establishing efficient sharing policies and assessing attributes of individual participants. Analogously, the centrality of a piece of information could be used as a proxy for, e.g., its relevance.

- Trust- and risk-based algorithms for the dissemination of information through the community. We are currently developing flexible but robust schemes that take as input a description of the sharing context (e.g., need-to-share this data, maximum risk allowed, etc.) and choose paths along the community graph so as to maximise dissemination while keeping risk of disclosure under control. We believe that this issue is particularly relevant when automating information exchange mechanisms, as the risks of sharing too much are apparent. However, unintended disclosures have very different consequences if the receiver is a highly trusted ally or an occasional collaborator, hence the need to explicitly consider trust in the decision making process. Similarly, privacy issues might be a major deterrent when parties face the problem of whether to share or not [31]. In this regard, both technical (e.g., trust-building mechanisms, data anonymisation) and non-technical (e.g., mutual agreements) measures should be further explored.
- Resilient but trusted communities. In many contexts, it is crucial to ensure that information reaches the intended recipients in time and with some minimum guarantees of risk containment. This requires building a community where paths with sufficient trust are always present, avoiding the presence of bridge nodes (i.e., nodes necessarily present in a subset of paths) and cut edges/nodes (i.e., those that make subset of nodes disconnected from each other if removed).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions, which have contributed to improve the quality of this work.

REFERENCES

[1] C. Fung, J. Zhang, I. Aib, and R. Boutaba, "Trust management and admission control for host-based collaborative intrusion detection," *Journal of Network and Systems Management*, vol. 19, no. 2, pp. 257-277, 2011.

[2] B. Schneier. (2012) The Vulnerabilities Market and the Future of Security. [Online]. http://www.schneier.com/blog/archives/2012/06/the_vulnerabili.html

[3] R. Langner, «Stuxnet: Dissecting a Cyberwarfare Weapon,» *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49-51, 201.

[4] B. Bencsath, G. Pek, L. Buttyan, and M. Felegyhazim, «The Cousins of Stuxnet: Duqu, Flame, and Gauss,» *Future Internet*, vol. 4, no. 4, pp. 971-1003, 2012.

[5] D.E. Denning, «Stuxnet: What Has Changed?,» *Future Internet*, vol. 4, no. 3, pp. 672-687, 2012.

[6] R. A. Martin, «Making security measurable and manageable,» in *IEEE Military Communications Conference (MILCOM)*, 2008, pp. 1-9.

[7] L. Dandurand, «Cyber Defense Data Exchange and Collaboration Infrastructure (CDXI),» in *ITU-T Workshop*, 2010.

[8] S. Teng, W. Zhang, X. Fu, and W. Tan, «Cooperative intrusion detection model based on scenario,» in *Proc. 11th International Conference on Computer Supported Cooperative Work in Design*, 2007, pp. 876-881.

[9] C.V. Zhou, C. Leckie, and S. Karunasekera, «A survey of coordinated attacks and collaborative intrusion detection,» *Computers & Security*, vol. 29, pp. 124-140, 2012.

[10] S. Snapp et al., «DIDS (distributed intrusion detection system) – motivation, architecture, and an early prototype,» in *Proceedings of the 14th national computer security conference;*, 1991, pp. 167-176.

[11] Internet Storm Center. [Online]. http://www.dshield.org

[12] RA Kemmerer, «NSTAT: a model-based real-time network intrusion detection system.,» University of California at Santa Barbara, 1998.

[13] S. Staniford-Chen et al., «Grids-a graph based intrusion detection system for large networks,» in *Proceedings of the 19th national information systems security conference*, 1996, pp. 361-370.

[14] P. Porras and P. Neumann, «Emerald: event monitoring enabling responses to anomalous live disturbances,» in *Proceedings of the 20th national information systems security conference*, 1997, pp. 353-365.

[15] RB. Abdoul Karim Ganame, J. Bourgeoisa, and F. Spiesa, «A global security architecture for intrusion detection on computer networks,» *Computers & Security*, vol. 27, pp. 30-47, 2008.

[16] V. Yegneswaran, P. Barford, and S. Jha, «Global intrusion detection in the DOMINO overlay system,» in *Proceedings of network and distributed security symposium (NDSS)*, 2004.

[17] M. Locasto, J. Parekh, A. Keromytis, and S. Stolfo, «Towards collaborative security and P2P intrusion detection,» in *Proceedings of the 2005 IEEE workshop on information assurance and security*, 2005, pp. 333-339.

[18] H. Yu, «A Survey of Trust and Reputation Management Systems in Wireless Communications,» *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1755-1772 , 2010.

[19] A. Josanga, R. Ismailb, and C. Boyd, «A survey of trust and reputation systems for online service provision,» *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007.

[20] J.H. Cho, A. Swarmi, and I.R. Chen, «A Survey on Trust Management for Mobile Ad Hoc Networks,» *IEEE Communications Surveys & Tutorials*, vol. 13, no. 4, pp. 562-583, 2011.

[21] D.E. Bell and L.J. La Padula, «Secure Computer Systems: Unified Exposition and Multics Interpretation,» The MITRE Corporation, ESD-TR-75-306 1976.

[22] MITRE, «Horizontal integration: Broader access models for realizing information dominance,» Availabe at http://www.fas.org/irp/agency/ dod/jason/classpol.pdf, 2004.

[23] P.-C. Chen et al., «Fuzzy multi–level security: An experiment on quantified risk adaptive access control,» in *IEEE Symposium on Security and Privacy*, 2007, pp. 222-230.

[24] J.A. Clark et al., «Risk based access control with uncertain and time-dependent sensitivity,» in *SECRYPT*, 2010, pp. 5-13.

[25] Q. Ni, E. Bertino, and J. Lobo, «Risk-based access control systems built on fuzzy inferences,» in *ASIACCS*, 2010.

[26] S. Kandala, R. Sandhu, and V. Bhamidipati, «An attribute-based framework for risk-adaptive access control models,» in *ARES*, 2011.

[27] J. Hu, R. Li, Z. Lu, J. Lu, and X. Ma, «RAR: A role-and-risk based flexible framework for secure collaboration,» *Future Generation Computer Systems*, vol. 27, pp. 574-586, 2011.

[28] L. Zhang, A. Brodsky, and S. Jajodia, «Toward Information Sharing: Benefit and Risk Access Control (BARAC),» in *POLICY*, 2006, pp. 45-53.

[29] R.A. Best Jr, «Need-to-Know vs. Need-to-Share,» Congressional Research Service, 7-5700, 2011.

[30] A. Gibbons, *Algorithmic Graph Theory*. Cambridge, UK: Cambridge University Press, 1985.

[31] ENISA, «Incentives and Challenges for Information Sharing in the Context of Network and Information Security,» 2010.

[32] L. Zhu, Y. Xu, P. Scerri, and H. Liang, «An Information Sharing Algorithm For Large Dynamic Mobile Multi-agent Teams,» in *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2012.

APPENDIX A: MSM ACRONYMS

CAPEC – Common Attack Pattern Enumeration and Classification.

CCE – Common Configuration Enumeration.

CCSS – Configuration Scoring System.

CPE – Common Platform Enumeration.

CVE – Common Vulnerabilities and Exposures.

CVRF – Common Frameworks for Vulnerability Disclosure and Response.

CVSS – Common Vulnerability Scoring System.

CWE – Common Weakness Enumeration.

CWSS – Common Weakness Scoring System.

CybOX – Cyber Observable Expression.

CYBEX – The Cybersecurity Information Exchange Framework.

IODEF – Incident Object Description Exchange Format.

MAEC™ – Malware Attribute Enumeration and Characterization.

OVAL – Open Vulnerability and Assessment Language.

OCIL – Open Checklist Interactive Language.

RID – Real-time Inter-network Defense.

RID-T – Transport of Real-time Inter-network Defense.

SBVR – Semantics of Business Vocabulary and Business Rules.

STIX – Structured Threat Information Expression.

SWIDs – Software Identification Tags.

XCCDF – Extensible Configuration Checklist Description Format.

Chapter 2.

Attack Modeling – Washing Away the Borders between Cyber and Kinetic Attacks

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

The Vulnerability of UAVs to Cyber Attacks - An Approach to the Risk Assessment

Kim Hartmann
Institute of Electronics, Signal Processing and Communication
Otto-von-Guericke-University
Magdeburg, Germany
kim.hartmann@ovgu.de

Christoph Steup
Department of Distributed Systems
Otto-von-Guericke-University
Magdeburg, Germany
steup@ovgu.de

Abstract: By 2012 the U.S. military had increased its investment in research and production of unmanned aerial vehicles (UAVs) from $2.3 billion in 2008 to $4.2 billion [1]. Currently UAVs are used for a wide range of missions such as border surveillance, reconnaissance, transportation and armed attacks. UAVs are presumed to provide their services at any time, be reliable, automated and autonomous. Based on these presumptions, governmental and military leaders expect UAVs to improve national security through surveillance or combat missions. To fulfill their missions, UAVs need to collect and process data. Therefore, UAVs may store a wide range of information from troop movements to environmental data and strategic operations. The amount and kind of information enclosed make UAVs an extremely interesting target for espionage and endangers UAVs of theft, manipulation and attacks.

Events such as the loss of an RQ-170 Sentinel to Iranian military forces on 4th December 2011 [2] or the "keylogging" virus that infected an U.S. UAV fleet at Creech Air Force Base in Nevada in September 2011 [3] show that the efforts of the past to identify risks and harden UAVs are insufficient. Due to the increasing governmental and military reliance on UAVs to protect national security, the necessity of a methodical and reliable analysis of the technical vulnerabilities becomes apparent.

We investigated recent attacks and developed a scheme for the risk assessment of UAVs based on the provided services and communication infrastructures. We provide a first approach to an UAV specific risk assessment and take into account the factors exposure, communication systems, storage media, sensor systems and fault handling mechanisms. We used this approach to assess the risk of some currently used UAVs: The "MQ-9 Reaper" and the "AR Drone". A risk analysis of the "RQ-170 Sentinel" is discussed.

Keywords: *UAV, Risk assessment, Cyber attack, Security analysis*

1. INTRODUCTION

The targets of concern to cyber conflict researchers are often either civilian infrastructures or military computer systems. However, the increasing level of technology in modern warfare and the reliance on these technical devices enforces the investigation of the vulnerability of advanced military devices against technical attacks.

Unmanned aerial vehicles (UAVs) are currently reascending military aerial devices capable of operating without human pilots on board. Previously predominately used by military services, UAVs are becoming increasingly valuable to civil applications. UAVs may manoeuvre autonomously, relying on on-board-computers or be remotely controlled by pilots from ground stations.

Within the past 5 years several incidents concerning drones have been reported by the public news agencies, showing and increasing the public interest in military and civilian drone applications.

The U.S. military increased its investment in the research and production of UAVs from $2.3 billion in 2008 to $4.2 billion in 2012 [1]. UAVs are currently used for a wide range of operations such as border surveillance, reconnaissance, transport and armed attacks.

UAVs are presumed to be reliable, automated and autonomous machines, providing their services at any time. Based on these presumptions, governmental and military leaders hope that UAVs improve national security. However, reviewing UAVs from a technical point of view, UAVs must be classified as highly exposed, multiply linked, complex pieces of hardware with high strategic and economic value.

It is interesting and bizarre that there is more research done regarding the security of modern cars incorporating car-to-car- and car-to-infrastructure-communication than research regarding the security of UAVs. It is unclear whether this is an effect of the closed-source-politics due to UAVs military origins or if these devices are simply considered to be secure due to their original tasks.

System security should never be considered as a state, but rather as a process. In order to support this process, it is important to be capable of describing and judging the current security status. Furthermore, it is desirable to be able to compare system configurations in terms of security levels. In order to fulfil these tasks, we are confronted with the questions: What is security and how is it measured?

Focusing on the technical aspect of the questions, (information) security is defined in the 44 USC §3542 [4] as " … protecting information and information systems

from unauthorised access, use, disclosure, disruption, modification, or destruction … ". Hence, security is a value describing how good a system is protected against the above named.

In order to determine how good a system is protected, it is important to know its vulnerabilities. Technically, the vulnerability of a system is an aspect of the system that heightens the probability of malfunction due to specific incidents. Depending of the severity of the malfunction, ranging from the complete loss of control/ destruction of the system to mere errors, the vulnerability may impose a threat to the systems security. In other words: A threat is a possible incident with a severe impact on the systems security. An incident may either be an attack or an event [5].

In terms of system security, a risk is a combination of the severity of the impact of an attack on the systems security, multiplied by its probability of occurrence. Hence, risk assessment quantifies the possible severity and likelihood of attacks. It is a crucial value for any high-level security system [6].

Interestingly, attackers searching for targets go the same way as system architects designing a secure system. An attacker is searching for a system vulnerability imposing a high threat, implying a high risk. A system architect is trying to eliminate vulnerabilities imposing high threats and hardens the system through the integration of coping mechanisms.

To heighten the systems security it is essential that the system designer finds vulnerabilities before attackers do. This is achieved by continuous risk analysis and assessment. Risk assessment schemes defined for most types of software- and hardware-components exist. However, none such risk assessment scheme or guideline for UAVs was found. Alarmingly, the reported incidents regarding UAVs indicate that the risk assessment – if used - for UAVs must be deficient. This paper aims at improving this situation through supplying a prototype scheme for the risk assessment of UAVs and the initiation of an academic discussion on the topic.

2. UAV – BASICS

UAVs are highly exposed technical systems. To analyse an UAVs vulnerabilities, it is important to understand what components an UAV is made of and how these components interact. In order to analyse UAVs on a common basis, we described UAVs in terms of component models.

Figure 1 shows a general component model of a standard UAV, without autonomous flight entity and weapons. The model in Figure 1 describes the basic components a UAV must incorporate.

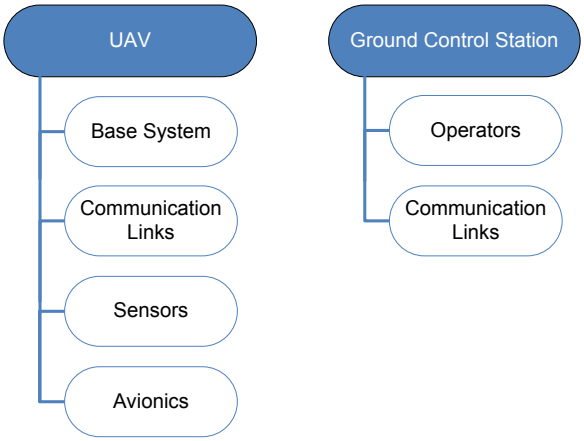

Figure 1. Right: General component model of a UAV. Left: Simple component model of a ground station

The "UAV base system" is the foundation of the UAV linking together the UAV components. It is needed to allow inter-component communication and controls the sensor, navigation, avionic and communication system. It may be considered as an UAV "operating system". The base system also allows the integration of further optional components such as special sensors or weapon systems.

The UAV sensor system consists of the sensory equipment of the UAV together with integrated pre-processing functionalities. For common military UAVs these sensors are often cameras with different capabilities. UAVs may be equipped with further sensors, such as INS, GPS and radar.

The UAV avionic system is responsible for the conversion of received control commands to commands of the engine, flaps, rudder, stabilisers and spoilers.

The in-flight communication of UAVs is always wireless and may be divided into two types: a) direct, line-of-sight (LOS) communication and b) indirect – mostly – satellite communication (SATCOM).

Figure 2 displays the information flow between components of the UAV system.

Newer UAVs, such as the RQ170 Sentinel, are able to operate autonomously. They may be additionally capable of holding and operating weapons as well as weapon supporting systems (e.g. the MQ-9 Reaper).

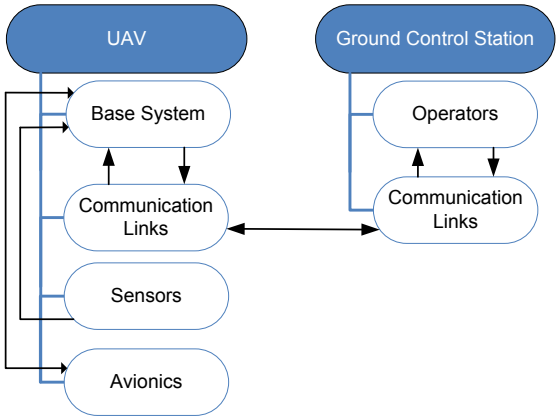

Figure 2. Information flow between the UAV components and the ground station

To account to the above adjustments, an extended UAV component model is given in Figure 3.

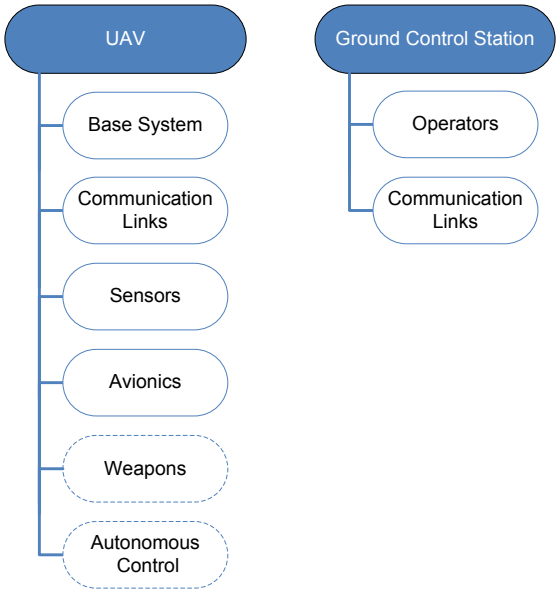

Figure 3. Extended UAV component model

The information flow within the extended UAV component model may differ, according to the UAV type. The exact internal communication may be relevant

for an attacker, if the attacker already has access to the internals of the system. Otherwise it is not essential.

Unless physical access to the UAV is given, an attacker must access and influence the UAV externally. Luckily for an attacker, UAVs are highly dependent on external input and therefore provide multiple input channels. Due to the "wireless nature" of UAVs, these channels are wireless and hence difficult to harden.

There are several information flows between an UAV and its environment, as shown in Figure 4. The two most important operational connections are 1) the bidirectional information flow between the communications system and the ground control station (GCS) and 2) the information flow from the environment to the sensors.

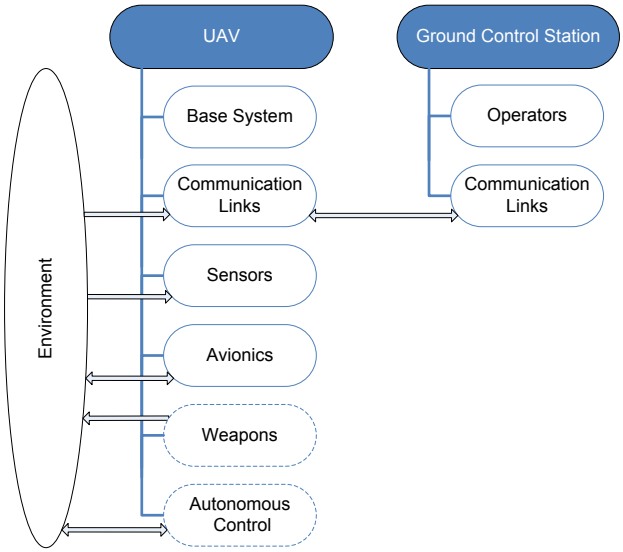

Figure 4. Extended UAV component model with information flow

However, additional influences between the environment and the UAV must be considered. These influences are the changes of the attitude of the UAV induced by the avionics, the result of weapons on the environment and the influence of the environment on the communication links.

The links are diverging in reliability and receptive to manipulation in different ways. While the reliability of sensors and system components are mostly investigated during system design, the consideration of the receptiveness of a sensor or system component to manipulation is not common.

The key to unauthorised control of an UAV is knowledge of the receptiveness of the system components to manipulation. To avoid third parties to take advantage of this knowledge, the receptiveness must be considered during system design.

3. RECENT ATTACKS

The incorporation of UAVs in military services was accompanied by a series of accidents having a broader impact on the overall security of UAVs.

One of the most recent and interesting incidents was the claimed theft an RQ-170 Sentinel by Iranian forces. It is widely accepted that Iranian forces are in the possession of the RQ-170 Sentinel. This claim was implicitly confirmed by a press statement of US-President Obama, asking for the return of the UAV [2] .

However, the circumstances under which the UAV came into the possession of the Iranian forces are controversial. Two popular theories exist that explain how the RQ-170 Sentinel may have been lost.

The first theory supposes that a vulnerability of the UAV sensor system with effects on the navigation system was used to attack the GPS system, discussed by Humphreys [7]. The attack uses details about the GPS functionality which make it easy to attack the GPS system of an UAV by a “GPS-spoofing”-attack. The GPS-satellite-signal is overlaid by a spoofed GPS-signal originating from a local transmitter with a stronger signal. The spoofed GPS-signal simulates the GPS-satellite-signal, leading to a falsified estimation of the UAVs current position. Supporter of this theory suppose that Iranian forces jammed the satellite communication of the drone and spoofed the GPS-signal to land the drone safely on an Iranian airfield.

Although the described attack is difficult to execute, it is not impossible [8]. If Iranian forces possess the knowledge and techniques to complete a GPS-spoofing attack remains and open question.

The second theory explains the loss of the UAV as a result of a technical malfunction. The theory postulates that the UAV may have landed on Iranian territory due to a technical malfunction. This may have allowed Iranian forces to recover the UAV.

Both theories indicate security problems. The GPS-Spoofing theory emphasises the necessity to include further and unusual components (e.g. sensors, input channels) in the risk assessment of UAVs. Partial autonomous systems as UAVs are dependent on their sensor systems in order to operate correctly. Furthermore, the sensor system must be reviewed as a continuously open input channel and may hence be prone to attacks.

Some reported incidents craved the destruction of the UAV to secure the confidentiality of sensitive data, [9], [10]. The technical malfunction theory claims that a self-destruction of the RQ-170 Sentinel was not possible. Regardless whether this theory is correct or not, it shows the necessity to examine the autonomous behaviour of UAVs regarding the security implications. An UAV must be capable of autonomously choosing the right strategy in case of a severe fault to uphold the systems security.

Another threat to UAVs is the exposure of the GCS to viruses as in the keylogging-virus attack [3]. The possible consequences may range from a loss of sensitive data to a loss of control of the assigned UAVs.

Another type of attack reported aimed directly at the communication link between the UAV and its GCS. During this attack live video feeds of an UAV were captured by Iraqian forces. The attack was possible due to a disabled encryption of the communication link. The software used to accomplish the attack was worth $26 [11].

4. PROACTIVE RISK ASSESSMENT SCHEME

We assessed the risk of security violations of UAVs based on our component models. Accordingly, the overall risk assessment of an UAV is the summation of its components risk assessment.

The risk assessment result of the provided scheme is multi-dimensional. It provides the risk assessment according to the type and intensity of security needed. It is a component-wise, probability-based evaluation of integrity, confidentiality and availability of the UAV [5]. A high score in the risk assessment scheme corresponds to a high risk regarding the loss of confidentiality, integrity or availability.

The scheme provides information on the susceptibility of components to attacks on the integrity, confidentiality or availability of the component, respectively of the UAV. According to the level of susceptibility, values between 0 and 1 are appointed to the component (0 meaning “not susceptible”, 1 corresponds to “highly susceptible”).

The values given by the scheme represent the susceptibility of the investigated component to attacks influencing integrity, confidentiality or availability. To calculate the risk, the specific probabilities of the occurrence of an attack are multiplied with the susceptibility value [12]. The result must be evaluated according to the severity of the loss of integrity, confidentiality or availability of the investigated component/ UAV [6]. The aspects of security may be in conflict.

The multi-dimensional risk assessment considers the different requirements of UAVs. According to the general task of the UAV, different aspects of security play varying roles and must be weighted accordingly. Therefore, the risk assessment of UAVs is always mission-bound.

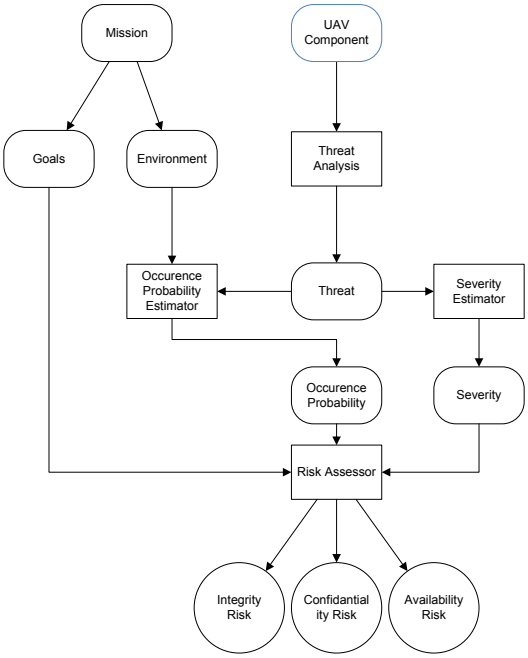

Figure 5. General overview of the proposed UAV risk assessment scheme.

A. ENVIRONMENT

As seen in the component model in Figure 1, the environment influences the UAVs sensors, its communication links and avionics. Hence, the environment must be considered in the UAV risk assessment. It is important to distinguish between political and physical factors of the environment, as these influence security aspects differently.

The landforms may be classified according to geomorphological categories. We considered two types of landscape (lowland and mountainous) and two political states (friend or enemy). This selection is only for demonstration purposes.

The influence of environmental factors on the UAVs security level in terms of availability, confidentiality and integrity is shown in Table I. The physical factors described are not capable of influencing the UAVs confidentiality or integrity.

However, other factors such as weather conditions, altitudes etc. may influence integrity.

The two political factors considered have influences on all aspects of the systems security. An UAV moving in enemy territory may lose its availability due to a heightened threat of destruction, takeover, signal disturbances etc.. Additionally, the UAV is exposed to the threat of confidentiality or integrity loss due to the risk of takeover, theft or manipulation.

Table I. Prototype environmental influence on UAV

Landscape	Integrity	Confidentiality	Availability
Lowland	0	0	0
Mountainous	0	0	0.9
Friendly territory	0	0	0
Enemy territory	0.9	0.9	0.9

B. COMMUNICATION LINKS

For the investigated UAVs, the satellite link tends to use the K_u-Band. The LOS-communication with the GCS is often based on the C-band or WiFi b-/g- or n-standard.

The following subsections give a short introduction on common communication types.

1) TCDL Ku-band communication

The TCDL (Tactical Common Data Link) is a secured data link developed by the U.S. military, capable of deriving data from different sources. It may furthermore route, encrypt, de-/multiplex, encode and transmit data at high speeds.

The TCDL uses a narrowband uplink at 15.15 GHz – 15.35 GHz and a wideband downlink at 14.40 GHz – 14.85 GHz. The TCDL may be operated both with directional and omnidirectional antennas and has ranges of 200 km at rates from 1.5 Mbit/s to 10.7 Mbit/s and low bit-error-rates. It may be used to transmit sensor data of any kind, especially radar, images and video signals.

One characteristic of K_u-band-based communication is that it is susceptible to rain/

snow fade. Due to the high frequencies used the signal may become disturbed by air humidity.

However, K_u-band-based communication is harder to overhear and hence harder to actively disturb than other comparable communication links, as required by [13].

2) LOS Communication: C-Band

Generally, the C-band describes the electromagnetic spectrum ranging from 4 GHz to 8 GHz. The C-band is used by a wide range of applications, such as weather radar systems, satellite communication, cordless phones and WiFi communication.

The frequencies relevant to uplink/downlink of the UAV communication systems investigated are 4.4 – 4.94 GHz and 5.25 – 5.85 GHz.

The C-band communication is less susceptible to air humidity than K_u-band communication. Nevertheless, due to the variety of applications, several COTS-devices exist that may interfere the radio signal and cause signal distortion.

UAVs tend to use omnidirectional antennas for C-Band communication, heightening the threat of interception by third parties.

3) LOS Communication: WiFi a/b/g/n

WiFi, synonymously described as "WLAN", refers to any communication based on the IEEE 802.11-standard. The frequencies used and the transmission rates differ according to the used standard. WiFi a, referring to the IEEE 802.11 a standard, ranges from 5.15 GHz – 5.75 GHz at transmission rates of 54 Mbit/s. The b and g standard operate in the frequency range of 2.4 GHz – 2.4835 GHz at 11 Mbit/s (b), respectively 54 Mbit/s (g). The WiFI n standard may operate both at 2.4 GHz as well as in the 5 GHz range. Due to the use of MIMO (Multiple Input Multiple Output), the n standard may transmit over longer distances and higher rates (up to 600 Mbit/s). To cover longer distances and achieve higher rates, the n standard uses multiple data streams and up to 4 antennas.

Due to its multiple applications and free usage, the b and g standard must expect signal interference. The frequencies above 5 GHz are restricted; hence interferences through civil applications are less likely. However, this may change in the near future (5-GHz-WLAN)

Because of the omnidirectional antennae used in the WiFi standards, WiFi is susceptible to eavesdropping. Precautions such as tunneling and encryption may be taken, but the general risk of eavesdropping – compared to other media – is still heightened as no knowledge of the signals direction is needed to tap the signal.

4) Summary - Scheme for communication links

The result of the general risk assessment scheme for communication links is shown in Table II. It is important to note that - although all communication links impose security threats to all aspects of security - the degree of susceptibility varies greatly. The overall risk depends on the specific task.

Table II. Risk assessment results for commonly used communication links

Link type	Integrity	Confidentiality	Availability
K_u-band	0.1	0.1	0.1
C-Band	0.1	0.5	0.5
WiFi a	0.1	0.9	0.9
WiFi b	0.1	0.9	1
WiFi g	0.1	0.9	1
WiFi n	0.1	0.9	0.9
No encryption	0	0.9	0
No signature	0.9	0	0

C. SENSORS

Sensors may be classified according to the type of reference used. References can be external or internal. An external reference is e.g. a GPS satellite. INS on the other hand relies only on internal references of physical parameters, such as acceleration or angular rates.

To determine the risks of the individual sensor systems, the characteristics of the sensor, the importance of the aspect observed and the mechanisms to detect spoofed or false sensor values must be considered.

Sensors with external references are more susceptible to jamming and spoofing than sensors with internal references. External references generally impose a risk to the integrity of the system.

Sensors relying on internal references must cope with value drifts, a certain deviation from the correct value over time. This phenomenon is due to the lack of external synchronisation and inherent errors. Reliable coping strategies exist and an external synchronisation may additionally take place when appropriate. It is widely accepted, that internal reference systems impose no additional risk.

Aspects of the environment that are crucial to the correct execution of the mission

must be observed correctly and reliably. If such an aspect is observed solely by a sensor with an external reference, a risk for the integrity and the availability of the system may emerge. An UAV relying on GPS-based navigation is prone to attacks on the GPS-sensors, which may be jammed or spoofed. In this case, due to the reliance on the external reference and the lack of control and coping mechanisms, the correct autonomous behaviour of the UAV cannot be guaranteed [7].

However, sensors observing non-critical aspects of the environment may also impose security threats. If the values delivered by the sensors are incorrect and other components rely on these values, the implications may be severe. Hence, all sensor data needs to be checked before used. Consequently, only optional sensors with reliable failure and attack detection mechanisms impose no additional risk for the integrity.

The redundancy mechanisms used to compensate sensor values may additionally contribute to the systems security. If several - but different - sensors are used to observe one aspect of the environment, the acquired values are considered more reliable. It is less likely that multiple, different sensors are jammed or spoofed collectively. Therefore, it may be concluded that single sensor observations impose an additional threat for the systems integrity. If one sensor observes a crucial value, such as flight attitude, this imposes a threat to the availability* of the system, as jamming or spoofing of this sensor may lead to the loss of the UAV.

The above observations lead to the risk assessment according to Table III.

The risk assessment must be done for each sensor in the UAV system as well as every observed mission aspect. Since depending on the mission, different aspects need to be considered and different aspects are critical, a mission specific sensor setup will provide better options to lessen the risk for the UAV system. Also the application of sensor fusion mechanisms, as described in [14], for cross-checking and enhancement may lessen the risk of integrity or availability loss.

Commonly combined sensor systems as GPS, INS, camera and radar will now be discussed based on the results of the general analysis.

INS is a traditional sensor to observe positional data and flight attitude for planes. INS is often paired with GPS as an additional sensor to acquire absolute position data. GPS relies on external references, creating +1 for integrity. However, a navigation system based on an INS and a camera system are combined to observe optical feature - see [15] - it poses no immediate security risk, even though the increasing deviation is still present. If all three systems are combined, jammed/spoofed GPS values are overruled by the INS and the optical features. This combined sensor system poses no additional security risk.

Table III. General sensor risk assessment, overview

Sensor system property	Integrity	Confidentiality	Availability
Sensor with external reference	0.9	0	0
Mandatory sensor with external reference	0	0	0.9
Mandatory sensor without redundancy	0.9	0	(0.9)*
Optional sensor without attack or fault detection	0.9	0	0

To control an UAV, awareness of the UAVs current situation is needed. This accounts to autonomous and human control. In current UAVs the situation awareness is created by camera or radar systems. The multiple camera system MTS-B that is used in the MQ9-Reaper consists of infrared, daylight and light enhancing cameras, which are automatically fused to provide an optimal image. This heterogeneous setup decreases the risk of jammed or spoofed sensor data due to cross-checking and mutual enhancement. Although it is theoretically still possible to jam the cameras, the used light would need to cover a wide frequency spectrum, making it impractical and unlikely.

The results of the sensor system discussion are shown in Table IV.

Table IV. Risk assessment results for different sensor combinations and mission aspects

Aspect	Sensor System	Integrity	Availability
Navigation	INS	0	0.9
Navigation	GPS	1.8	0.9
Navigation	INS + GPS	0.9	0
Navigation	INS + Optical Flow	0	0
Navigation	INS + GPS + Optical Flow	0	0
Flight Attitude	INS	0	0.9
Flight Attitude	INS + Optical Flow	0	0
Situation Awareness	Single Camera	0.9	0
Situation Awareness	Multiple Cameras	0	0

D. DATA STORAGE

The risk assessment of data storage mechanisms considers three main aspects:

1. Volatility
2. Encryption
3. Signature

The usage of volatile storage imposes a risk to the availability of the stored data. If appropriate coping strategies are lacking, this may also lead to an inconsistent storage state and hence result in a loss of integrity of the stored data. However, the sole use of volatile storage does not impose an additional risk to the confidentiality of the stored data.

The use of encryption mechanisms may preserve the confidentiality of stored data. The lack of encryption generally heightens the risk of confidentiality loss. Encryption mechanisms do not prevent the stored date from being overwritten, which implies a risk for data integrity. To secure the integrity, mechanisms such as signatures or forgery detection must be integrated. These mechanisms have no influence on the confidentiality or availability of the data.

Using the above considerations, the resulting observations are:

- The availability of the data is based on the volatility of the storage medium.
- Solid state storage imposes no risk on the availability, as it is considered robust.
- Hard drive based storage and magnetic tapes are susceptible to force and magnetic fields, resulting in a higher risk of data loss.
- Volatile memory such as RAM is considered to impose no risk for the confidentiality but may impose a risk of availability and integrity loss.

We considered magnetic tapes, hard drive storage, solid state storage and temporary storage through RAM. The risk assessment for the considered storage media is shown in Table V.

Table V. Risk assessment of common data storage media

Storage type	Integrity	Confidentiality	Availability
Analog magnetic tape	0.9	0.9	0.9
Hard drive based storage	(0.9)	(0.9)	0.9
Solid state based storage	(0.9)	(0.9)	0
RAM	0.9	0	0.9

The numbers in brackets imply that the actual value depends on the encryption and signature used and may be 0. The values converge to zero if the data stored is signed and encrypted using strong encryption mechanisms.

E. FAULT HANDLING MECHANISMS

Fault handling mechanisms are difficult to assess regarding their "usefulness" in terms of security aspects. Although it is obvious that a "good fault handling mechanism" should improve the systems overall security, it is not obvious what good fault handling mechanisms for UAVs are. This is a common research problem of UAVs.

UAVs are technical systems and prone to faults in all of their components. Faults create errors, unhandled errors lead to malfunctions and disrupt the mission. To prevent this, the emerging of faults must be prohibited or faults must be masked by appropriate fault handling mechanisms [16].

Examples for fault handling mechanisms are "triple modular redundancy" or "fail-safe states". These mechanisms may cause restrictions on the functionality of the UAV, but enable the continuation of the mission. However, the fail-safe state may impose new threats to the security if the state is chosen unwisely.

Consider the following example: An UAV which is controlled remotely through a communication link must switch into a fail-safe state if the communication link is lost. One possible fail-safe state is to maintain the current position until the communication link is restored. In this case the UAV needs to aviate based on its on-board sensors, making the impact of manipulated sensor data tremendous. An example of this type of attack is the GPS-signal spoofing [7].

To assess the threats imposed by the fault handling mechanisms of an UAV it is necessary to categorise the possible faults. A fine grained categorisation is discussed in [17]. We categorise security threats by severity of the fault and fault type (transient or permanent).

Transient faults are often external temporary disturbances, such as communication interferences due to weather conditions. Permanent faults are mainly hardware damages.

The risk assessment of fault handling mechanisms in UAVs considers transient and permanent mission critical fault handling mechanisms and analyses their implications on integrity, confidentiality and availability.

Different fault-handling strategies for mission critical faults exist, examples are "self-destruct", "automatic-return", "land" and "hover". Not all strategies may be equally appropriate for all faults [18].

The possible fault handling mechanisms for severe faults of general UAV components are shown in Table VI.

The "hover" strategy requires working avionics and navigation. For transient faults "hover" provides the ability to continue the mission after recovery. However, due to possibly limited sensor and communication facilities the UAV is more likely to be attacked through spoofed or manipulated data. This invokes threats to the integrity of the mission.

The "automatic-return" strategy provides the best chance of retrieving a functional UAV, but it imposes the same risks as the "hover" strategy.

Table VI. Component-dependent fail-safe states

Component	Fault handling mechanism
Base system	self-destruct
Data Storage	land, self-destruct, (automatic-return)
Sensors	hover, (automatic-return), land, self-destruct
Communication	hover, automatic-return, land, self-destruct
Avionics	Land, (automatic-return), self-destruct

The "land"-strategy needs a minimal set of working components and is also applicable in the case of engine failure. However, in enemy territory it imposes a risk on integrity and confidentiality.

The "self-destruct" strategy has the lowest risk of misuse or exposure of sensitive data, but it destroys the availability of any UAV component or data.

The deduced risk assessment values are shown in Table VII.

Table VII. Fail-safe state risk assessment results

Strategy	Integrity	Confidentiality	Availability
Hover	0.9	0	0
Land	0.9	0.9	0.9
Automatic-return	0.9	0	0
Self-destruct	0	0	0.9

The risk assessment shows that the security aspects are hardly compatible. This implies that fault handling mechanisms should be adapted to the preferred security aspect.

5. RESULTS

This section presents the results of applying the described scheme to modern UAVs.

A. AR.DRONE

The parrot AR.Drone is a remotely controlled quadrocopter originally designed for augmented reality video games. Meanwhile, the AR.Drone is commonly used as a research platform [19]. Apart from research institutions, the AR.Drone was also used during the "occupy wall street" actions to realise a robust police reconnaissance system [20].

The basic hardware setup incorporates a single IEEE 802.11b/g [21] compatible wireless communication link and an android or IOS based smartphone as GCS. The antenna is omnidirectional and the link is usually not encrypted.

Apart from RAM (used to buffer video streams), the AR.Drone does not possess any storage media. It contains two video cameras, an ultra-sonic range finder, a low-altitude altimeter and an INS as sensory equipment.

The fault handling mechanism in case of an error of the communication link is to enter the hover mode. Every other error results in instantaneous landing manoeuvres (land mode).

The results of our risk assessment for the AR.Drone are shown in Table VIII.

Table VIII. AR.Drone risk assessment results

Component	Integrity	Confidentiality	Availability
Communication links	1.1	2.7	2
Data storage	0.9	0	0.9
Sensors	2.7	0	0.9
Fault handling	1.8	0.9	0.9
Total	**6.5**	**2.6**	**4.7**

The sensor risk value results from the following observations: The used INS is accompanied by an optical flow measurement of the ground to track the position [15], which represents a checked mandatory sensor. The additional low-altitude

distance sensor can be used to manipulate the flight height of the drone, which is a risk comparable to an unchecked mandatory aspect sensor with external reference. The cameras never overlap, prohibiting image cross-validation.

B. MQ-9-REAPER

The General Atomics MQ-9 Reaper is a remotely controlled UAV. It is the successor of the MQ-1 Predator. It uses the TCDL satellite communication system (SATCOM) as well as a direct LOS C-band communication.

The control of the UAV is done by a GCS. The default equipment of the UAV consists of several cameras bundled in a multi-spectral targeting system (MTS-B). These cameras detect infrared, daylight and intensive light. The data is automatically pre-processed and fused by the MTS-B. The navigational sensors are INS and GPS.

The MQ9-Reaper contains digital storage for video data. The encryption and signature mechanism are unknown.

The results of the risk assessment are shown in Table IX.

Table IX. MQ-9-Reaper Risk assessment results

Component	Integrity	Confidentiality	Availability
Communication links	0.2	0.6	0.6
Data storage	0.9	0.9	0
Sensors	0.9	0	0
Fault handling	0.9	0.9	0.9
Total	**2.9**	**2.4**	**1.5**

The communication system uses two independent links, which are both encrypted and signed.

The data storage is non-volatile, the encryption and signature methods used are unknown. For our calculations we presumed the worst-case-scenario; no encryption or signature methods.

The used camera system is redundant and uses fusion. The used combination of INS and GPS poses a risk for the integrity of the data as the GPS uses an external reference.

The accident described in [18] shows that the remote pilot must cope with permanent faults manually. Furthermore, the self-destruct mechanism is activated manually. This may lead to uncontrolled landings or flights and imposes threats to the availability, integrity and confidentiality of the system.

C. RQ-170 SENTINEL

Due to the current investigations of the Iranian claim to have attacked an RQ-170 Sentinel, publically available and reliable sources regarding the equipment of the RQ-170 are rare. The data available allows only a partial risk analysis of the UAV.

The sensory equipment of the UAV consists of infrared and daylight cameras as well as GPS and INS. The equipment is similar to the MQ-9-Reaper. The risk assessment of these sensors and the combinations are equal to the MQ-9. It is likely that the scores of the Sentinel are similar to the MQ-9-Reapers scores, if not better.

The data storage is non-volatile; the encryption and signature mechanisms are unknown. The communication link and the fault handling mechanisms are unknown.

6. CONCLUSIONS

The risk assessment of UAVs is a complex task consisting of vulnerability and threat analysis and is additionally dependent on mission details. The discussed UAV related incidents imply that risk assessment schemes for UAVs are lacking or insufficient.

The provided scheme is a first attempt to describe and formalise the risk assessment of UAVs. A component model of UAVs was designed to categorise and define a component-based risk assessment.

The components “communication system”, “data storage” and “sensor system” were analysed based on the used technology and known vulnerabilities. Environmental factors and fault handling mechanisms were additionally investigated. Security was defined following the definition in the 44 USC $ 3542.

The provided scheme was applied to the AR.Drone and MQ-9-Reaper. A brief risk analysis of the RQ-170 Sentinel was done, however the currently public available data is insufficient to draw any further conclusions. It appears that the RQ-170 Sentinel will at least score at the same rates as the MQ-9-Repear. However, depending on the further system setup, it is equally likely that this impression is false.

The calculated values give an indication of the susceptibility of the investigated UAV to attacks influencing availability, integrity or confidentiality.

Within this scope, risk was defined as the result of the susceptibility of an UAV multiplied by the probability of occurrence of a specific attack on a component's vulnerability, multiplied by the severity of the attack. It was shown that the risk assessment of an UAV is highly dependent on the assigned task/mission.

The described method is a first approach to a general scheme for the risk assessment of UAVs. The risk analysis and assessment of each of the named components describes an individual research area. This paper understands itself as a basic but crucial introduction to the risk assessment of UAVs in terms of structure, tactics and analysis.

REFERENCES

[1] Lolita C. Baldor, «Flashy drone strikes raise status of remote pilots,» *The Boston Globe*, pp. online at 01.11.2012: http://www.bostonglobe.com/news/nation/2012/08/11/air-force-works-fill-need-for-drone-pilots/ScoF70NqiiOnv3bD3smSXI/story.html, 2012.

[2] CNN Wire Staff, «Obama says U.S. has asked Iran to return drone aircraft,» 2011.

[3] Noah Shachtman, «Computer Virus Hits U.S. Drone Fleet,» *Wired*, pp. online at 01.11.2012: http://www.wired.com/dangerroom/2011/10/virus-hits-drone-fleet, 2011.

[4] Cornell University Law School. [Online]. http://www.law.cornell.edu/uscode/text/44/3542

[5] Matt Bishop, *Introduction to Computer Security*, 1st ed., Addison-Wesley, Ed. Boston, USA: Pearson Education, 2004.

[6] Andrew Jaquith, *Security Metrics: replacing Fear, Uncertainty, and Doubt*, 1st ed., Addison-Wesley, Ed. Boston, USA: Pearson Education, Inc., 2010.

[7] Todd Humphreys, «Statement on the vulnerability of civil unmanned aerial vehicles and other systemes to civil gps spoofing,» Austin, 2012.

[8] David Cenciotti. (2011, December) The Aviationist. [Online]. http://theaviationist.com/category/captured-stealth-drone/page/2/

[9] US Air Force. (2007, November) United States Air Force. [Online]. http://usaf.aib.law.af.mil/ExecSum2008/MQ-1L_AOR_29Nov07.pdf

[10] US Air Force. (2007, December) United States Air Force. [Online]. http://usaf.aib.law.af.mil/ExecSum2008/MQ-1B_AOR_17Dec07.pdf

[11] British Broadcasting Corporation. (2009, December) BBC News. [Online]. BBC News: online at http://news.bbc.co.uk/2/hi/world/middle_east/8419147.stm

[12] Carl Young, *Metrics and Methods for Security Risk Management*.: Syngress Media, 2010.

[13] Erdal Torun, «UAV Requirements and Design Consideration,» in *RTO-MP-44*, Ankara, Turkey, 2000, pp. B4-1 - B4-8.

[14] D.L. Hall and J. Llinas, «An introduction to multisensor data fusion,» in *Proceedings of the IEEE*, 1997, pp. 6 -23.

[15] Pierre-Jean Bristeau, François Callou, David Vissière, and Nicolas Petit, «The Navigation and Control Technology Inside the AR.Drone Micro UAV,» in *18th IFAC World Congress*, Milano, Italy, 2011, pp. 1477-1484.

[16] Jane W. S. Liu, *Real-Time Systems*, 1st ed., Prentice Hall, Ed., 2000.

[17] Algirdas Avizienis, Jean-Claude Laprie, and Brian Randell, «Fundamental Concepts of Dependability,» Newcastle, 2001.

[18] US Air Force. (2009, March) United States Air Force. [Online]. http://usaf.aib.law.af.mil/ExecSum2009/MQ-9_FortIrwin_20Mar09.pdf

[19] Vojtěch Vonásek,Daniel Fišer,Jan Faigl Tomáš Krajník, «AR-Drone as a Platform for Robotic Research and Education,» in *Research and Education in Robotics - EUROBOT 2011*, Prague, Czech Republic, 2011, pp. 172-186.

[20] N. Sharkey and S. Knuckey. (2011, December) The Guardian. [Online]. http://www.guardian.co.uk/commentisfree/cifamerica/2011/dec/21/occupy-wall-street-occucopter-tim-pool

[21] IEEE, «IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,» *IEEE Std 802.11-2012*, pp. 1-2793, Mar 2012.

[22] Laurence R. Newcome, *Unmanned aviation: a brief history of unmanned aerial vehicles*. Michigan, USA: American Institute of Aeronautics and Astronautics, 2004.

[23] P. W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*.: Penguin Books, 2009.

[24] Bill Yenne, *Birds of Prey: Predators, Reapers and America's Newest UAVs in Combat*. Pasadena, CA, United States: Specialty Pr, 2010.

[25] Ian Palmer, *Unmanned Aerial Vehicles: Robotic Air Warfare 1917-2007*. Essex, United Kingdom: Osprey Publishing, 2008.

[26] Kimon P. Valavanis, *Advances in Unmanned Aerial Vehicles*. Dordrecht, The Netherlands: Springer Netherland, 2008.

[27] Army UAS CoE Staff, «»Eyes of the Army» U.S. Army Roadmap for Unmanned Aircraft Systems 2010-2035,» U.S: Army UAS Center of Excellence, Fort Rucker, Alabama United States, 2010.

[28] J. R. Wilson, «A new generation,» *Aerospace America*, pp. 28-32, January 2007.

[29] T. J. Nugent and Kare J.T., «Laser Power for UAVs,» LaserMotive, Kent, WA United States, White Paper.

[30] Secretary of Defense, «Unmanned Systems Roadmap 2007 - 2032,» U.S. Department of Defense, Washington D.C., USA, Roadmap December 2007.

[31] John R. Vacca, *Computer and Information Security Handbook*.: Morgan Kaufman, 2009.

[32] Douglas L. Landoll, *The Security Risk Assessment Handbook: A Complete Guide for Performing Security Risk Assessments*.: Crc Pr Inc, 2011.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

A Cyber Attack Modeling and Impact Assessment Framework

Igor Kotenko
Laboratory of Computer Security Problems
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
Saint-Petersburg, Russia
Email: ivkote @comsec.spb.ru

Andrey Chechulin
Laboratory of Computer Security Problems
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
Saint-Petersburg, Russia
Email: chechulin@comsec.spb.ru

Abstract: The paper suggests a framework for cyber attack modeling and impact assessment. It is supposed that the common approach to attack modeling and impact assessment is based on representing malefactors' behavior, generating attack graphs, calculating security metrics and providing risk analysis procedures. The main aspects outlined are achieving near-real time mode, event analysis and prognosis mechanisms, security and impact assessment. To optimize the attack graph generation and security evaluation we apply an anytime approach to have the result at any time by applying a set of algorithms with different timelines and precision. The architecture of the Cyber Attack Modeling and Impact Assessment Component (CAMIAC) is proposed. We present the prototype of the component, the results of experiments carried out, and comparative analysis of the techniques used.

Keywords: *attack modeling, attack graphs, security metrics, impact assessment, anytime algorithms*

1. INTRODUCTION

Currently, computer networks are playing an important role in many areas. The increasing size and complexity of networks lead to the growth of complexity of their security analysis. Possible financial, political, and other benefits, which can be gained by cyber attacks, lead to substantial increase of the number of potential malefactors. Despite these facts, the existing security analysis is a process which still dependents mainly on the experience of security administrators. All these problems define the importance of the research and developments in the area of automated security analysis of computer networks.

There are many approaches to the network security analysis. One of the promising approaches consists in cyber attack modeling and impact assessment based on attack graphs (or trees). Studies related to building and analysis of attack graphs have been conducted for over 20 years. *Attack graph* is a graph which represents all possible sequences of the malefactor actions that lead him/her to the goals established. These action sequences are also called *attack traces*. The main disadvantage of this approach is its *computational complexity*. Building a complete attack graph for a malefactor (and especially for a set of malefactors) is a computationally complex problem and usually takes a long time. For instance, for a small size network the attack graph can be formed quickly, however, when the graph must be built for a network which includes hundreds or even thousands of hosts and the result should be obtained in a limited time (or even in real time), the graph based algorithms require a very large amount of computational resources. Moreover, over time the composition of hosts and links between them can be changed, and the attack graphs will require reconstruction. Thus, at the present time, the usage of attack graphs in systems operating in near real-time mode, for example, in Security Information and Event Management (SIEM) systems, is very complicated.

This paper suggests a framework for designing the Cyber Attack Modeling and Impact Assessment Component (CAMIAC) which implements the attack graph generation, real-time event analysis techniques, prognosis of future malefactor steps, attack impact assessment, and anytime approach for attack graph building and analysis. In contrast to the existing works of the authors [17-19], the paper describes the attack modeling and impact assessment solutions directed to optimization of attack graph building and analysis process with the goal to enable their usage in the systems operating in near real time. The main contributions of the paper are as follows: two stages based algorithm for attack graph building, the basic principles of real-time event analysis, the approach to identify possible malefactors by analyzing the compliance between security events and attack graphs, the application of anytime approach for the attack graphs analysis.

The paper is organized as follows. Chapter 2 analyzes the state-of-the-art in attack modeling. Chapter 3 outlines briefly key elements of suggested techniques for attack modeling and impact assessment. Chapter 4 presents a prototype of the CAMIAC. The architecture and implementation are described. Chapter 5 describes a case study and an example of the experiments with the prototype of the CAMIAC. It is shown how suggested approaches could be applied to a real use case. Also the comparison with existing tools is presented. Conclusion surveys main results and next steps of the research and development activities.

2. RELATED WORK

There are a lot of papers, which consider different approaches to attack modeling and security evaluation taking into account various classes of attacks. We analyze briefly current state-of-the-art in representation of attack scenarios and malefactors, generation of attack graphs, determining security metrics, combining service dependency graphs with attack graphs, representing zero day attacks and specification of platforms, vulnerabilities, vulnerability scorings, attacks, weaknesses and configurations.

In [1, 20, 30] *attacks are represented* in a structured and reusable tree-based form. In [30] a high-level conceptual model of attacks based on the intruder's intent (attack strategy) is suggested. The comprehensive work using the so-called tree based approach is proposed in [1]. This paper describes means for documenting attacks in a form of attack trees. One of the most important problems in security analysis is the malefactors' classification and model construction. In [38] the task of modeling and simulation of intelligent, reactive attackers is described. The suggested computer network attack model uses an action representation based on the GOLOG situation calculus [13] and goal-directed procedure invocation.

Different approaches, which use *attack graphs and trees for security analysis*, have been suggested. S. Hariri et al. [40] calculate global metrics to analyze and proactively manage the effects of complex network faults and attacks. S. Noel, S. Jajodia et al. [25, 41] propose a technique based on determining the minimum-cost network hardening via exploit dependency graphs. I. Kotenko and M. Stepashkin [14-16] are focused on security metrics computations based on attack graph representation of malefactor behavior. R. Lippmann and K. Ingols [35] propose to use attack graphs to detect firewall configuration defects and host critical vulnerabilities. This approach was extended by taking into modern network attacks threats (zero-day exploits and client-side attacks) and countermeasures (intrusion prevention systems, personal firewalls, and host-based vulnerability scanners) [23]. J. Ryan and D. Ryan [21] suggest calculating metrics based on failure-time analysis. L. Wang, S. Jajodia et al.

[24, 26] propose to calculate attack resistance metrics based on probabilistic scores by combining Common Vulnerability Scoring System (CVSS) scores [11]. N. Kheir et al. [32] suggest an implementation of confidentiality, integrity and availability metrics using the notion of privilege, which is inspired by access permissions within access control policies.

There is a new trend of research in attack modeling, which is to *combine attack graph models and service dependency models*. In their essence, attack graphs represent possible attacker actions in the light of current system configuration. Meanwhile, they do not represent service dependencies and their underlying connection requirements. N. Kheir et al. [31] propose to extend the use of CVSS metrics in the context of intrusion response, by supplying this metric with dynamic information about system configuration and service dependencies structured within dependency graphs. The dependency graph is further used to evaluate the overall impact of an attack, thus replacing the informal environmental parameters in the CVSS vector. Nonetheless, the problem with this approach is that it does not provide clear evidence on how to interface service dependency graphs with attack graph models.

The analysis of network security against unknown zero day attacks is also a relatively new topic of research. Zero day attacks can be defined as attacks which use unknown vulnerabilities. E. Bursztein [12] extends the security analysis approach, based on game theory, by taking into account zero day exploits. L. Williams [27] presents a practical realization of the approach to calculate the possible number of zero day vulnerabilities. M. McQueen et al. [29] attempt to evaluate the total number of possible zero day vulnerabilities for one day. K. Ingols et al. [23] suggest ordering different applications by the seriousness of consequences of having a single zero day vulnerability. L. Wang et al. [24] propose a security metric called k-zero day safety. It is based on how many unknown vulnerabilities are required to compromise a network asset, regardless of the type of vulnerabilities.

Very important relevant work is connected with research and developments in coherent *description of vulnerabilities, attacks, weaknesses, security policies and configurations, lists of the software/hardware installed on each platform, events, countermeasures*, etc. Common Vulnerabilities and Exposures (CVE) [9] contains the list of known information security vulnerabilities and exposures. Usage of the National Vulnerability Database (NVD) [33] based on CVE dictionary is the basis for constructing of attack graph via known vulnerabilities. Common Vulnerability Scoring System (CVSS) [11] is an open and standardized vulnerability scoring system for vulnerabilities rating. Common Weakness Enumeration (CWE) [10] contains a unified, measurable set of software weaknesses. Usage of the database of weaknesses can improve the quality of the zero-day based attack graph generator

module. Common Platform Enumeration (CPE) [7] provides a unified description language for information technology systems, platforms, and packages. Common Configuration Enumeration (CCE) [6] gives common identifiers to system configuration issues. Common Attack Pattern Enumeration and Classification (CAPEC) [5] helps to capture and use the attacker's perspective. Usage of attack patterns allows applying sequences of known and zero-day vulnerabilities in one attack action. Common Remediation Enumeration (CRE) [8] defines a security-related set of actions that result in a change to a computer's configuration. Remediations may be motivated by discovered vulnerabilities or misconfigurations.

3. PROPOSED TECHNIQUES

Let us consider the main techniques that are suggested in the CAMIAC: attack graph generation, real-time event analysis, prognosis of future malefactor's steps, attack impact assessment, and anytime approach. These techniques are based on using a comprehensive security repository, efficient attack graph (tree) generation techniques, taking into account known and new attacks based on zero-day vulnerabilities, stochastic analytical modeling, and interactive decision support to choose preferred security solutions (countermeasures). Not all these aspect are outlined in the paper due to limited volume.

A. ATTACK GRAPH GENERATION

Let us consider at first the basic definitions needed for attack graph generation.

Basic objects define the graph *vertexes*. They are linked to each other by *edges* to form different sequences of malefactor's actions. Basic objects and the links between them are included in the *network model* which is used for attack graph generation. Basic elements can belong to two types: "host" and "attack action". The objects of the "*host*" type describe hosts discovered and attacked by malefactors, while the objects of the type "*attack action*" describe all distinguishable actions of malefactors.

The algorithm of generating common attack graph is based on implementation of the following sequence of actions: (1) preparatory actions which allow a malefactor to move from one host to another; (2) reconnaissance actions for detection of "live" hosts; (3) reconnaissance scenarios for detected hosts; (4) attack actions based on vulnerabilities and (5) auxiliary actions.

As the result, all *attack actions* are divided into the following classes: (1) reconnaissance actions; (2) preparatory actions within the limits of

malefactor's privileges (these actions are used for creation of conditions needed for implementation other attack actions; (3) actions for gaining the privileges of local user and of administrator; (4) confidentiality, integrity and availability violation.

As in [45], we use three necessary conditions for adding a potential attack in the attack graph: (1) the protected system has vulnerabilities; (2) an attacker needs knowledge and resources to perform attacking activities; (3) an attack accomplishment facilitates the achievement of the malefactor goal. The first condition is determined completely by properties of the protected system. The second one is defined by both the system and the malefactor model properties. The third one is determined by the malefactor goals.

Respectively, at the first stage - the *stage of preparation and construction of attack trees* - a *3-dimension matrix* is formed for each host according to the following information:

(1) class of attacks (data gathering, preparation activities, escalation of privileges, attack goal realization);

(2) needed access type (remote source without access rights, remote user of the system, local user of the system, administrator);

(3) restriction for malefactors (by malefactor knowledge, zero-day vulnerabilities, etc.).

As a result for each host a set of corteges (attack action class, access type, and malefactor knowledge level) is formed, for each cortege in its turn a list of particular attacks and vulnerabilities needed for these attacks implementation is generated. The total list of vulnerabilities is formed on the base of host software and hardware description using CPE [7] and public vulnerability databases such as NVD [33]. Additional data sources for the detected vulnerabilities are the reports of security scanners such as Nessus, MaxPatrol, etc. In the CAMIAC the vulnerabilities are stored in the CVE format [10].

When constructing an attack graph, particular attack patterns described in the CAPEC format [5] are used. The CAMIAC uses these patterns not only as input information, but also allows producing new ones. They can correspond to the most often used sequences of vulnerability exploitations and other actions of the attacker. The patterns also contain attack descriptions that do not use vulnerabilities, for example at the initial stage of an attack the malefactor could gather information on available hosts. To specify in this case the attacker actions, the CAPEC-292 (Host Discovery) entry is used. It describes a group of various ways of scanning hosts and ports. This group contains, for example, such entries as CAPEC-285 (ICMP Echo

Request Ping), CAPEC-296 (ICMP Information Request), CAPEC-299 (TCP SYN Ping), etc.

The second stage is a *search of vulnerable software*. The examples of patterns used to describe malefactor actions are as follows: CAPEC-310 (Scanning for Vulnerable Software), CAPEC-311 (Fingerprinting Remote Operating Systems), CAPEC-300 (Port Scanning), etc.

On the third stage of *attack graph generation*, both particular vulnerabilities from the CVE dictionary and patterns like CAPEC-233 (Privilege Escalation) are used.

After forming matrixes of potential attacks, for each host the *possible malefactor type and his/her initial location* are chosen for the analyzed network.

The examples of the malefactor type are as follows:

(1) *External hacker*, a user having significant knowledge in information security field, but lacking any direct possibility to connect to the internal network; possible intrusion points, for example, are servers which can be accessed via the Internet (web servers, mail servers, etc.);

(2) *Internal user*, a user having basic knowledge in information security field with local user or administrator rights;

(3) *Worm/virus/botnet*, a program that can use a set of vulnerabilities specified in advance. It is supposed that in this case a part of internal network can be already infected.

The full *malefactor model* includes following parameters: type (internal, external, complex); initial privileges for each host of the network (none, remote user, local user, administrator); possible access points in the network; knowledge level (defines possible attack actions).

Further for each chosen malefactor model a list of possible goals is generated. For example, for the internal user it could be a revenge (causing maximum damage for the company). The goal of the external hacker could be the access to confidential information located on a server inside the network. For a worm at the first stage a goal can be its distribution, while at the second one it could be carrying out DDoS attacks.

Therefore, the malefactor in the CAMIAC is presented by a pair "malefactor model, goal", which determines constraints on the usage of attacking actions and possible initial intrusion point into the network.

The *key elements of the suggested approach* are as follows: (1) for all malefactor

models the attack graph is formed in the same time on the basis of information gathered; (2) for each malefactor model the security metrics are evaluated; (3) for each malefactor who can successfully realize attacks, the list of possible countermeasures is formed.

Due to the fact that the attack modeling cannot be often fulfilled in real-time, its usage in real-time processes is limited. However, the generated attack graphs keep their actuality for a certain period of time (until significant changes in the security policy or physical network topology occur).

Thanks to this in the frame of the general system of event analysis it is suggested to use the attack graphs constructed in advance. These attack graphs can be used when solving two main tasks: (1) predicting subsequent malefactor actions and (2) analysis and detection of their past actions which led the system into its current state.

However, in some cases the attack modeling system needs to *update attack graphs*. For example, this necessity occurs when host characteristics (software and hardware, criticality, etc.), network topology and a list of possible malefactors are changed, as in these cases key objects (malefactor models, matrixes of host properties, etc.) are changed.

However in this case attack graphs are updated partly as the changes are calculated only for particular elements of matrixes. Due to this fact the computational complexity of the update decreases significantly.

B. REAL-TIME EVENT ANALYSIS

Attack graphs produced by the attack modeling component allow us to specify the ways of system security violation using different attacks. For a given malefactor, the attack graph corresponds to an attack tree, where the root represents an initial location of the malefactor and the leaves depict conditions which allow achieving the malefactor goals. That means that all paths from the root to the leaves are the sets of potential attacks. An attack scenario in its turn represents a minimum range of conditions the malefactor should meet to achieve the goal.

Depending on the kind of the graph, a scenario can represent a sum of some leaves or a subset of the graph elements including at least the root and one terminal leaf. Therefore, the graphs and the attack scenarios are supplementary notions. While the former allows getting some knowledge on potential attacks, the latter represent detailed information on particular attack type which can be included in a great number of graphs.

In fact the *task of real-time event analysis* consists in detection of a set of attack trees this event can belong to. In the end, an ideal result of the CAMIAC operation should be an attack scenario (i.e. a path from the root to its leaf) completely determining the malefactor model (i.e. goals and possibilities of the malefactor) and its possible subsequent actions.

In existing literature a range of approaches to find out the most probable attack scenario on the base of analysis of particular security events is proposed.

W. Lee et al. [48] suggest an approach allowing determining a scenario, which lower level events belong to. The paper also describes a technique for evaluation of probability of revealed malefactor goals and its subsequent steps. In the CAMIAC we apply this approach.

Three potential attack conditions described in [45] allow us evaluating a probability the attacker chooses exactly a given attack (proceeding from a fact that the malefactor uses a way of maximally quick achievement of his/her goal). On the basis of the probability of vulnerability exploitation by a particular malefactor model, it is possible to detect what generated attack tree the detected event relates to.

C. PROGNOSIS OF FUTURE MALEFACTOR'S STEPS

The most part of existing attack detection systems revealing current attacks cannot predict subsequent malefactor actions. It is obvious that the prediction of subsequent malefactor steps allows increasing the protection level of the system against malefactors. Thus, the main function of the security system becomes the detection of particular malefactors and generation of targeted protection measures rather than the discovery of particular attacks.

Let us consider the following example. If a malefactor conducts an attack against some host in a network, he/she can have the goal (1) to capture the control over the host for carrying out further attacks or (2) to assess data on the host.

To implement the system protection correctly, it is necessary to define the malefactor goals and predict its future actions. Let us consider these variants in detail.

The first goal supposes the malefactor will use the attacked host as an intermediate one, thus, it is not the object of the final goal. Therefore, we could predict the next actions of the malefactor: he/she will concurrently look for other intermediate hosts. If the malefactor captures the attacked host, the attack will be progressing from it. In this case it is worth to increase the sensitivity of rules for attack detection of the attacked host as a protection mean in order to detect the capture of a given host. It allows gathering additional information on the malefactor and his/her methods.

The second goal supposes the attacked host is the final goal of the malefactor and contains valuable information. It is likely that all hosts in the network controlled by the malefactor participate in the attack. In this case it is reasonable to block temporarily all suspicious or all available connections to this host as all hosts connected to the attacked host should be considered as potentially infected.

Thus, depending on the malefactor goal the response of the security system on one event detected in the network (attack) can differ. In fact the wrong malefactor goal detection can even assist the malefactor in the goal achievement. For example, when the attack goal is detected as one of the second type, the protection mechanism will allow the malefactor conducting DDoS attack, which may lead to successful implementation of some other attack, for instance IP spoofing.

We propose an approach that allows determining the future malefactor actions on the basis of the reports from an event correlation subsystem or an intrusion detection system (IDS) and the analysis of attack graphs created in advance for different models of potential malefactors. This approach also facilitates conduction of a retrospective analysis of events in the network, which allows revealing of unknown vulnerabilities (0-day).

The proposed approach includes the following steps:

(1) An attack graph(s) is (are) formed on the basis of the network and malefactor models;

(2) In a real network the network of connected sensors is formed. These sensors allow detection of particular attacking actions. A monitoring system allows producing a general picture of events occurred in the network on the base of information gathered by sensors. The events should be normalized, prioritized and correlated by an event correlation subsystem (or an intrusion detection system);

(3) Further the management subsystem finds correspondences between attack graphs and events in the real network. Thus, according to the analysis of incidents and data received from the attack modeling subsystem it is possible to conclude that there exists a sufficient probability that an incident “scanning of host C by host B” is followed by some undetected incident “host B was attacked by host A” and the next action of the malefactor is “host C is subjected to attack from host B”.

In the literature a number of approaches to determine a current attack scenario on the base of security events is described. For example in [3] the PHATT (Probabilistic Hostile Agent Task Tracker) algorithm using Bayes models is suggested. When revealing a scenario, this algorithm uses probabilities of attack fulfillment, contradictions in scenarios and works with a range of malefactors having different goals.

We use the following admissions simplifying the work of an algorithm are assumed:

(1) the malefactor uses only a single scenario at a particular moment;

(2) elements of a scenario have no strict time limitation;

(3) if the malefactor uses several scenarios, he/she runs them sequentially.

D. SECURITY AND IMPACT ASSESSMENT

Attack modeling needs to represent not only the sequences of actions, but also the attack consequences (in terms of impacts), as well as how countermeasures can mitigate these impacts and for which cost [32]. According to this principle we realize five main groups of security and impact assessment metrics.

The first group includes *metrics that are connected with topology, criticality and vulnerabilities of the analyzed system (hosts)*: the level of the host vulnerability which is defined on the base of the known vulnerabilities (we suggest to exploit CVE and NVD for its assessment); the level of the host criticality which is defined by its position and functions in the system (we suggest to define it on the base of CVSS); and the vulnerability of host to the zero-day attacks. The last metric allows considering the potential to compromise the system via unknown vulnerabilities. Here we use CWE to detect "weak places" of the system, and on the base of the approach in [23], depending from the malefactor type, we suppose the existence of zero-day vulnerabilities on the hosts where they lead to the maximum impact. By extension of the attack graph and taking into account weak places we define and use for impact assessment a set of possible unknown vulnerabilities.

The second group includes *metrics characterizing the attack, for example, attack potentiality.* When we define the *attack potentiality* (probability), the attack graph is used. These metrics are based on CVSS and the metrics of the first group, and allow calculating the integrated complexity and severity of the sequence of steps that are necessary to compromise the system assets.

The metrics of *the third group* characterize *the malefactor's potential* and are intended to define possibilities of the attack development. As the basis for such calculation we consider the malefactors position in the system and his/her skills (malefactor profile).

The metrics of *the fourth group* are *response efficiency and response collateral damage*. Response efficiency measures the response ability to reduce attack impacts. Response collateral damage evaluates negative influence of countermeasures on the system efficiency. To evaluate response collateral damage we use both attack graphs and service dependencies.

The last group includes *integral spatial characteristics of the system security* and a score of the system risk level. For example, the approach of qualitative express assessment of network security level uses the following basic metrics: *Criticality*(h) - criticality level of host h; *Severity*(a) - criticality level of attack action a; *Mortality*(a,h) - damage level caused by attack action, taking into account the criticality level of host; *Mortality*(S) and *Mortality*(T) - damage level of route S and threat T; *AccessComplexity*(a), *AccessComplexity*(S), *AccessComplexity*(T) - "access complexity" of attack action a, route S and threat T; *Realization*(T) - admissibility of threat realization; *RiskLevel*(T) - risk level of threat T; *SecurityLevel* – general security level of the computer network.

E. ANYTIME APPROACH

To improve the efficiency of construction, modification and analysis of attack graphs the usage of the anytime approach is proposed. The main goal of the anytime approach is to have the result at any time by applying a set of algorithms with different timelines and precision. Summarizing, the anytime algorithms suppose the following peculiarities: a opportunity to obtain the solutions, possibly not precise, as soon as they are needed during the process of solving the problem; a solution found is to be of a sufficient level of adequacy (but it may be either incomplete or approximate); with the lapse of time, the obtained solutions are getting closer to the final result (i.e. improving the precision).

Application of the anytime algorithms for the cyber attack modeling and impact assessment includes the procedures for constructing and analyzing attack trees, including calculating security metrics. Such application enables continuous security monitoring and decision-making.

In the suggested approach the security evaluation is conducted with the use of security metrics, which can be obtained by means of complex analysis including the detection of hosts, network interfaces, operating systems, taking into account different kinds of communication, etc. Such analysis can take a lot of time and hence exact values of the integral metrics (e.g. express evaluation of the protection) can be available in due time.

For security level evaluation by anytime approach, the following groups of algorithms were selected (they are sorted in the order of time expenses and precision increase):

(1) analysis of security level on the base of lists of vulnerabilities detected on hosts without determination of a particular malefactor model (or taking into account a simplest model – a level of complexity of vulnerabilities the malefactor can use) and without considering an attack graph;

(2) construction of an attack graph, where all sub-networks are grouped according to the criticality level, meanwhile the vulnerability groups are computed as a sum of vulnerabilities of individual hosts in them;

(3) construction of an attack graph for the complete network;

(4) dynamic attack simulation as a technique that allows obtaining more precise attack modeling.

Also, in order to construct an anytime algorithm for integral security metrics computation, one can conduct their calculations for some subnet available at the given moment and for hosts for which all needed information is already known or for subnet which is changed.

4. ARCHITECTURE AND IMPLEMENTATION

The task for this stage of research and development was to develop a version of the CAMIAC prototype, which uses the techniques suggested in the paper. In this section we describe the CAMIAC prototype architecture and its current implementation.

The CAMIAC prototype is aimed to demonstrate the approach to modeling attacks at various levels. To do this the prototype has to implement the following functionality:

(1) generation of basic attack trees (not in real-time mode);

(2) construction of the dynamical simulation model (not in real-time mode) imitating stochastically different attacks and countermeasures;

(3) computation of the security level of the network, determining the possible bottlenecks and defining other network security metrics (anytime mode);

(4) updating attack trees taking into account changes in input data (near real-time mode);

(5) detection of scenarios of the current attack (near real-time mode).

The input data for the prototype is as follows: network (system) configuration and the list of hosts' software and hardware; list of existing vulnerabilities from External database (DB) - National Vulnerability Database (NVD); security scanners reports, that can consist information about network configuration and vulnerabilities; real-time events from the external correlation engine.

The output of the prototype consists of a list of security metrics calculated for the network and a recommendation for increasing its security level.

The general architecture of the implemented CAMIAC is shown in Figure 1.

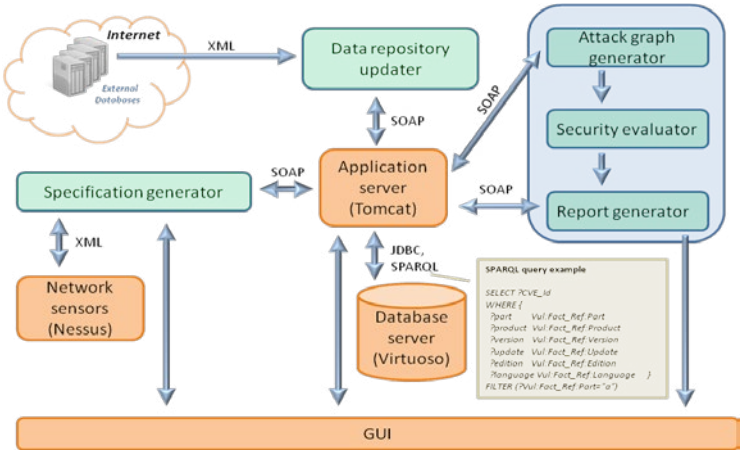

Figure 1. CAMIAC prototype architecture

This figure contains three main components of the CAMIAC: (1) Data repository updater; (2) Specification generator; (3) Attack graph generator and analyzer. Additionally the prototype includes the database and the files for storing the tested network elements and links between them. The first component allows updating the internal database of vulnerabilities, using information obtained from NVD (National Vulnerability Database) [33]. The second one mainly aims to create and modify models of computer networks and helps operator to form all input values like malefactor models and security requirements. The third component demonstrates the suggested approach for attack graph building and analysis.

To organize the fast interaction with the database, the ontology for representation of the CAMIAC data has been developed. The model is based on the SCAP protocol. The approach to the vulnerability presentation for the CAMIAC allows getting significantly lesser amount of data from the database and getting rid of a necessity of program processing, delegating analysis task to a logical reasoning system. The Virtuoso server [47] is used as the storage for ontologies. Interaction with the main CAMIAC module is carried out through the Repository Application Server (RAS), representing web services for demands to the logical reasoning system. The client part is written in Java. When implementing RAS, the Jena framework is used. The logical reasoning system is embedded into the Virtuoso server.

For the interaction between CAMIAC components, SOAP v1.2 [43] is used. The client generates a query in the form of XML document, which is transported using HTTP. Apache Tomcat [2] is chosen as a servlet container as it satisfies all requirements. The web-services are implemented in Java programming language [37]. The service implementation level uses the Hibernate library [36], which supports Java Persistence API v. 2.0 (JPA) [46].

The example of the CAMIAC dashboard showed in Figure 2 is used to setup initial data.

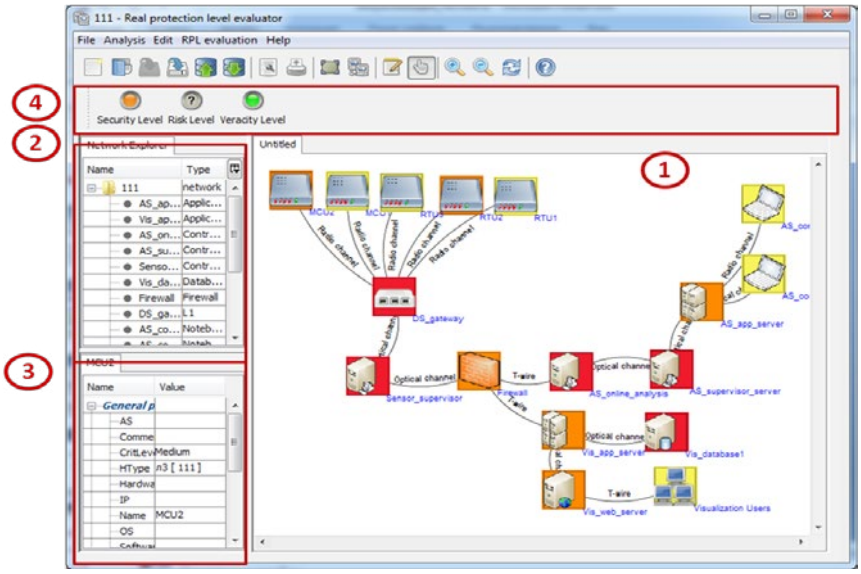

Figure 2. Example of CAMIAC dashboard

This dashboard can be divided into four subviews. The main view 1 shows the topology of the studied network, while view 2 reflects the hierarchical structure of the network, depicting domains or specified networks zones. The graph based techniques are used to represent the network topology. Each network object is represented by an icon. The user has possibility to define icons for each type of the network objects. The background color of the icon is used to encode values of the security metrics calculated for the given host, such as Criticality, Mortality, Risk Level. These metrics are chosen by the user from the predefined list. The brief information about each host is available a via tool tip which appears when mouse hovers over the network object.

The user can configure each host and network using the property view 3. It can specify predefined properties of the host such as IP address, host type (web server,

ftp server, database server, router, firewall, etc.), installed software and hardware, user-defined host criticality. These properties are necessary for attack graph generation. There is also a possibility to define user properties. This property view is updated whenever a particular state node is selected. Thus user always has details at hand.

The view 4 shows the security metrics calculated for the network itself: Security level, Risk Level, and Veracity level. As these metrics can have value from the predefined set of values {Low, Medium, Above Medium, High, Undefined}, they are presented in a form of the semaphore signal. We think that such dashboard design gives a general overview about security analysis of the network and communicate a lot of information in a glance. Thus, the user can analyze calculated host security metrics in the context of initial host configuration; all information is available in different views, but on one dashboard panel.

5. CASE STUDY

We performed the following experiments with the prototype implemented to show the advantages of the proposed CAMIAC framework:

(1) Formation of the attack graph for a computer network;

(2) Evaluation of network security by the attack graph analysis;

(3) Import of the report containing the IDS security events;

(4) Analysis of the security events related to the changes in the source data (changes in the list of installed software, host list and links between hosts) to show the modification of attack graphs;

(5) Analysis of the security events for detection of attack actions in order to recognize possible malefactor models.

The network of a small company was selected for experiments. This network includes the hosts of several types: user computers, a database, a web server and network equipment. For each host in the network the software and hardware were defined.

Several malefactor models were selected:

(1) An *external malefactor with medium knowledge in security area*. His/her knowledge enables usage of the attack action with low and medium complexity. The external type means that he/she is located in the Internet and has access only to the web server. The aim of this model is to collect information about the company

network. This malefactor does not know any zero-day vulnerability and does not have any rights in internal network;

(2) An *external malefactor with high knowledge in security ar*ea. His/her knowledge enables usage of the attack actions with any complexity. The external type means that he/she is located in the Internet and has access only to the web server. The aim of this model is to destroy all information in the internal database of the company. This malefactor knows several zero-day vulnerabilities, but initially does not have any rights in the internal network;

(3) The *internal malefactor with low knowledge in security area*. His/her knowledge limited with the list of attack actions which can be performed by some security tools. The aim of this model is to modify some information in the internal database of the company. This malefactor does not know any zero-day vulnerability, but has initial access to the users' computers in the internal network and has user and remote user rights for several hosts.

To make clearer the illustration of the CAMIAC prototype possibilities, let us consider a case with the following software for network hosts: operating system (OS) Windows Server 2003 is installed on all hosts, DBMS MySQL 5.0 - on the host Database, Apache HTTP Server 1.3.6 - on the Web Server host.

After constructing the attack graph, the CAMIAC provides the following information: the malefactor knowledge after all possible attacks, the attack tree in the graphic form and the log of the malefactor's actions.

For the malefactor 2, the attack graph example is depicted in Figure 3.

The malefactor, carrying out attack actions, is located on the top of the graph. The other icons are as follows: "A" – an attack action, "S" – a scenario which does not use vulnerabilities (for example, host discovery (PING)), "V" – an attack action which exploits some vulnerability.

According to the attack graph the chain of malefactor's actions and their results are as follows:

(1) Detection of nodes connected with the initial malefactor host. Web Server host is detected; (2) Detection of the software installed on the Web Server host. Windows Server 2003 is detected; (3) Usage of the vulnerability CVE-2007-0214. Malefactor compromises of the Control Web Server; (4) Detection of the nodes connected with the Web Server. Application Server host is detected, etc.

Thus, the malefactor starts to perform attack actions from the host "External Web Server Users". This host is a starting point because the malefactor has all privileges

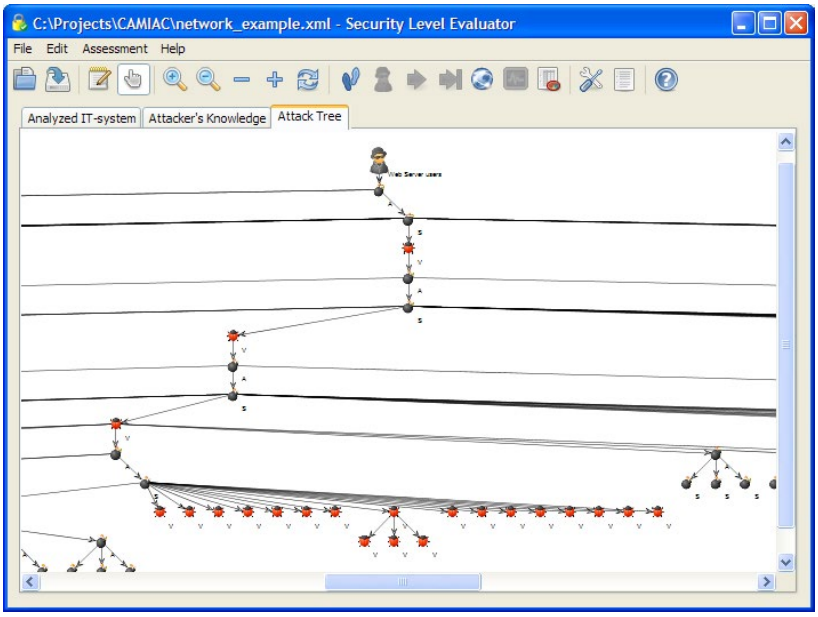

Figure 3. Attack graph example

in the host according to the specified malefactor model. The selected malefactor is an external for the network, and he/she can connect only to the "Web Server".

Firstly the malefactor gathers the information about the host "Web Server" and performs attack actions without any privileges on this host. After several attack actions the malefactor obtains the remote and local users privileges and continues the information gathering. The final step of the malefactor on this host is to obtain administrators privileges. Then the malefactor scans for accessible hosts and starts new attack actions for a new host.

According to the suggested metrics the security level of the tested network is evaluated. For each node the criticality level is determined, for example for the nodes "Web Server Users" and "Application Server" it is LOW while for the node "Firewall" it is HIGH.

There are 12 hosts in the attack graph. For these hosts eight different successful attack actions were discovered and modeled. The attack graph contains 207 different attack routes. These routes contain 31 security violations (confidentiality, integrity and availability) for different host.

For each attacker's action and each possible attack route the security metrics Access Complexity (AC) and Mortality (M) are calculated.

Thus, the Route parameters Access Complexity and Mortality equal LOW (the minimal values for each host in the route). These metrics form the basis for the general network level evaluation. In this use case the Security Level is ORANGE, what means that countermeasures need to be implemented. The weak place in the analyzed network for the selected malefactor model is the host "Web Server" – all 207 routes passed through it. The main recommendation for the system administrator is to increase the protection of the host "Web Server".

The next stage of the experiment is the analysis of security events that are received from the analyzed network. To do this a simple report is created. This report contains two types of the security events – the events describing changes in the network and the events including the recognized attacking actions. The following rows are the examples of the events content:

(1) 192.168.0.107	192.168.0.2	SCAN nmap TCP {tcp}
(2) 192.168.0.2	installed	cpe:/a:mysql:mysql:5.1.33

Event 1 contains information about the detected scanning process. The Database host (192.168.0.2) was scanned by the Nmap tool from some user's host in the network (192.168.0.108). If there are no other attack actions in the security event report, then the CAMIAC make a decision that with high probability the malefactor of type 3 was detected in the network. The reason of this decision is the fact that the malefactors of type 1 and 2 should firstly perform attacks on Web Server, after that they can attack some user host, and only then, they are able to attack database host.

Event 2 specifies new software installed to the host Web Server. It contains the CPE description of the MySQL server (DBMS MySQL 5.1.33 was installed instead of MySQL 5.0). This event leads to the fact that some vulnerabilities that are specific to the previous version of the database may disappear. For instance, this stipulates that the malefactor of type 3 will not be able to modify data in the database, and thus the evaluation of network protection for the malefactor of this type is changed to the GREEN.

6. COMPARISON WITH RELATED COMMERCIAL SYSTEMS

As it was mentioned the number of the implemented security evaluation systems based on attack graph analysis is very small in contrast to amount of the theoretical papers. In this section, several related systems of different classes, which can fulfill security analysis functions, are analyzed: COMNET III [4], OPNET [36], Amenaza SecurITree [42] and Nessus [34]. There are other related systems, but they have

similar disadvantages which we are trying to overcome in the CAMIAC solutions.

Stochastic discrete event simulation systems like COMNET III [4] and OPNET [36] allow creating the detailed model of computer networks. The results of simulation are the evaluation of network protection against a variety of attacks including resource depletion. Disadvantage of these systems is the high resources needed for development. Detailed simulation of the network activity of all services and hosts requires a long time and, therefore, the use of such systems for security analysis is very complicated. In addition, after the changes of network topology and services, it is necessary to fulfill repeated simulation. Thus, taking into account the requirements of operative near real time security analysis, these systems are worse than the CAMIAC by efficiency and resource consumption parameters.

Amenaza SecurITree [42] is an example of commercial software which uses attack trees for security analysis. This tool is designed for attack tree building and analyzing, it has a friendly interface and very detailed documentation. The disadvantage of this system in comparison with the CAMIAC is the lack of possibility to investigate specific malefactors with his/her capabilities and goals. Also there is no support for real-time event analysis in this tool.

Nessus security scanner [34] interacts with the real network and during the scanning cannot penetrate internal network from the external network, if some security system is installed. That is why it usually recognizes only a small number of vulnerabilities. The approach based on malefactor modeling and attack graphs analysis, implemented in the CAMIAC, allows detecting all currently known vulnerabilities in the network, regardless of the original location of the malefactor. Also Nessus can detect changes in the analyzed network, but it requires full rescanning for that.

7. CONCLUSIONS

The paper suggests a framework which allows using attack graphs to evaluate security and provide impact analysis for detection of malefactors and determining the countermeasures in near real time. To achieve this goal the graph generation process is divided on two stages. On the first stage it is suggested to generate the graph of potential attacks for a general malefactor model. This stage should be performed at the time of network deployment or as an offline process, when there are no severe time limitations. On the second stage the attack graph is modified according to the changes in the analyzed computer network. During this stage, modification and analysis of the attack graph should allow to obtain the results in a limited time. The detection of malefactors by their attack action is also performed during the second stage.

This paper gives consideration to the state-of-the-art in attack modeling, the essence of the approach to analytical attack modeling and impact analysis, as well as the architecture of Cyber Attack Modeling and Impact Assessment Component. The techniques suggested are based on the attack graph generation which represents possible attack scenarios taking into account the current security situation, including network configuration, security policy, events and alerts, probable malefactor's location, knowledge level and strategy, known and possible new vulnerabilities.

The developed prototype of the CAMIAC is described. The main difference between the proposed approaches and the existing ones is the possibility of the work in near real-time. Thus, the new results obtained in this investigation are the algorithms and methods of attack graph constructing and analyzing that excel in performance existing ones. All elements of attack modeling, described in the paper, will be extended and detailed in the next steps of research and development.

Acknowledgement

This research is being supported by grant # 13-01-00843 of the Russian Foundation of Basic Research, Program of fundamental research of the Department for Nanotechnologies and Informational Technologies of the Russian Academy of Sciences (contract #2.2), State contract #11.519.11.4008 and partly funded by the EU as part of the SecFutur and MASSIF projects.

REFERENCES

[1] A.P. Moore, R.J. Ellison, R.C. Linger. Attack Modeling for Information Security and Survivability. Technical Note CMU/SEI-2001-TN-001. *Survivable Systems*, 2001.

[2] "Apache Tomcat." http://tomcat.apache.org/. [Dec. 28, 2012]

[3] C. Geib, R. Goldman. "A Probabilistic Plan Recognition Algorithm Based on Plan Tree Grammars", *Artificial Intelligence.* vol. 173(11), 2009, pp. 1101-1132.

[4] "CACI Products Company." http://www.caciasl.com/, [Dec. 28, 2012]

[5] "Common Attack Pattern Enumeration and Classification (CAPEC)." http://capec.mitre.org/. [Dec. 28, 2012]

[6] "Common Configuration Enumeration (CCE)." http://cce.mitre.org/. [Dec. 28, 2012].

[7] "Common Platform Enumeration (CPE)." http://cpe.mitre.org/. [Dec. 28, 2012]

[8] "Common Remediation Enumeration (CRE)." http://scap.nist.gov/events/2010/saddw/presentations/remediation.pdf. [Dec. 28, 2012].

[9] "Common Vulnerabilities and Exposures (CVE)." http://cve.mitre.org/. [Dec. 28, 2012]

[10] "Common Weakness Enumeration (CWE)." http://cwe.mitre.org/. [Dec. 28, 2012].

[11] "Common Vulnerability Scoring System (CVSS)." http://www.first.org/cvss/. [Dec. 28, 2012]

[12] E. Bursztein. "Extending Anticipation Games with Location, Penalty and Timeline." LSV, ENS Cachan, CNRS, INRIA, France, 2008, pp. 272-286.

[13] H.J. Levesque, R. Reiter, Y. Lesperance, F. Lin, R.B. Scherl. "GOLOG: A Logic Programming Language for Dynamic Domains." *Journal of Logic Programming*, vol. 31, 1997, pp. 59-83.

[14] I. Kotenko, M. Stepashkin. "Attack Graph based Evaluation of Network Security." *Lecture Notes in Computer Science*, vol. 4237, 2006, pp. 216-227.

[15] I. Kotenko, M. Stepashkin, E. Doynikova. "Security Analysis of Computer-aided Systems taking into account Social Engineering Attacks.", in *Proc. of PDP 2011*, Los Alamitos, California. IEEE Computer Society, 2011, pp. 611-618.

[16] I. Kotenko, M. Stepashkin. "Network Security Evaluation based on Simulation of Malefactor's Behavior." *In Proc. of International Conference on Security and Cryptography (SECRYPT-2006)*, Portugal, 2006, pp. 339-344.

[17] I. Kotenko, A. Chechulin, E. Novikova. "Attack Modelling and Security Evaluation for Security Information and Event Management." *In Proc. of International Conference on Security and Cryptography. Proceedings (SECRYPT 2012),* 2012, pp. 391-394.

[18] I. Kotenko, A. Chechulin. "Common Framework for Attack Modeling and Security Evaluation in SIEM Systems." *In Proc. of 2012 IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing.* Los Alamitos, California. IEEE Computer Society, 2012, pp. 94-101.

[19] I. Kotenko, A. Chechulin. "Attack Modeling and Security Evaluation in SIEM Systems." *International Transactions on Systems Science and Applications*, vol.8, December 2012, pp.129-147.

[20] J. Dawkins, C. Campbell, J. Hale. "Modeling network attacks: Extending the attack tree paradigm." In *Proc. of the Workshop on Statistical and Machine Learning Techniques in Computer Intrusion Detection*, Johns Hopkins University, 2002.

[21] J. Ryan, D. Ryan. "Performance metrics for information security risk management." *IEEE Security and Privacy*, vol 6, 2008, pp. 38-44.

[22] K. Ingols, M. Chu, R. Lippmann, S. Webster, and S. Boyer. "Modeling modern network attacks and countermeasures using attack graphs" In *Proc. of the 2009 Annual Computer Security Applications Conference*, Washington, DC, USA, 2009, pp. 117–126.

[23] K. Ingols, M. Chu, R. Lippmann, S. Webster, S. Boyer. "Modeling modern network attacks and countermeasures using attack graphs." In *Proc of Annual Computer Security Applications Conference (ACSAC '09)*, Washington, D.C., USA, IEEE Computer Society, 2009, pp. 117-126.

[24] L. Wang, S. Jajodia, A. Singhal, S. Noel. "k-Zero Day Safety: Measuring the Security Risk of Networks against Unknown Attacks." In *Proc. of ESORICS'10*, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 573-587.

[25] L. Wang, S. Noel, S. Jajodia. “Minimum-cost network hardening using attack graphs.” *Computer Communications*, vol. 29, 2006, pp. 3812-3824.

[26] L. Wang, T. Islam, T. Long, A. Singhal, S. Jajodia. “An attack graph-based probabilistic security metric.” In *Proc. of the 22nd annual IFIP WG 11.3 working conference on Data and Applications Security*. Springer-Verlag Berlin, pp. 283-296, 2008.

[27] L. Williams. “GARNET: A Graphical Attack Graph and Reachability Network Evaluation Tool.” In *Proc. of the 5th international workshop on Visualization for Computer Security*, Springer-Verlag Berlin, 2008, pp. 44-59.

[28] M.M. Gamal, D. Hasan, A.F. Hegazy. “A. Security Analysis Framework Powered by an Expert System.” *International Journal of Computer Science and Security*, vol. 4(6), pp.505–526, 2011.

[29] M. McQueen, T. McQueen, W. Boyer, M. Chaffin. “Empirical estimates and observations of 0-day vulnerabilities.” In *Proc. of Hawaii International Conference on System Sciences*, 2009.

[30] M.Y. Huang, T.M. Wicks. “A Large-scale Distributed Intrusion Detection Framework Based on Attack Strategy Analysis.” *Computer Networks*, vol. 31, New York, NY, USA, pp. 2465-2475, 1999.

[31] N. Kheir, H. Debar, N. Cuppens-Boulahia, F. Cuppens, J. Viinikka. “Cost evaluation for intrusion response using dependency graphs.” In *Proc. of IFIP International Conference on Network and Service Security (N2S)*, IEEE, Paris, France, 2009, pp. 1-6.

[32] N. Kheir, N. Cuppens-Boulahia, F. Cuppens, H. Debar. “A service dependency model for cost-sensitive intrusion response.” In *Proc. of ESORICS 2010*, Athens, Greece, 2010, pp. 626-642.

[33] “National Vulnerability Database (NVD).” http://nvd.nist.gov/. [Dec. 28, 2012]

[34] “Nessus scanner software.” http://www.tenable.com/products/nessus/nessus-product-overview, [Dec. 28, 2012]

[35] O. Sheyner, J. Haines, S. Jha. “Automated generation and analysis of attack graphs.” In *Proc. of IEEE Symposium on Security and Privacy*, Berkeley, California, 2002, pp. 273.

[36] “OPNET Technologies, Inc.” http://www.opnet.com/. [Dec. 28, 2012]

[37] “Oracle Java SE.” http://www.oracle.com/technetwork/java/javase/downloads/index.html. [Dec. 28, 2012]

[38] R.P. Goldman. “A Stochastic Model for Intrusions.” *Lecture Notes in Computer Science*, vol. 2516. Springer Verlag, 2002, pp. 199-218.

[39] “Relational Persistence for Java and .NET.” http://www.hibernate.org/. [Dec. 28, 2012]

[40] S. Hariri, G. Qu, T. Dharmagadda, M. Ramkishore, C.S. Raghavendra. “Impact Analysis of Faults and Attacks in Large-Scale Networks.” *IEEE Security and Privacy*, vol. 1, pp. 49-54, 2003.

[41] S. Noel, S. Jajodia, B. O'Berry, M. Jacobs. "Efficient minimum-cost network hardening via exploit dependency graphs." In *Proc. of ACSAC'03*, 2003, pp. 86-95.

[42] "SecurITree – Attack graph analysis software. Amenaza Technologies Limited." http://www.amenaza.com/. [Dec. 28, 2012]

[43] "SOAP." http://www.w3.org/TR/soap/. [Dec. 28, 2012]

[44] "Symantec Enterprise Security Manager." https://www.symantec.com, [Dec. 28, 2012]

[45] T.L. Amenaza. "Fundamentals of Capabilities-based Attack Tree Analysis." Calgary, Canada, November, 2005.

[46] "The Java Persistence API." http://glassfish.java.net/javaee5/persistence/. [Dec. 28, 2012]

[47] "Virtuoso universal server." http://virtuoso.openlinksw.com/. [Dec. 28, 2012]

[48] W. Lee, X. Qin. "Attack Plan Recognition and Prediction Using Causal Networks." In *Proc. of the 20th Annual Computer Security Applications Conference (ACSAC 2004)*, Tucson, Arizona, December, 2004.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Exploring the Prudent Limits of Automated Cyber Attack

Jeffrey L. Caton
President
Kepler Strategies LLC
Carlisle, Pennsylvania, U.S.A.
Jeff.Caton@keplerstrategies.com

Abstract: The notion of cyber conflict occurring at network rates that surpass the speed of decision-making by national leaders has bolstered the possibility of introducing automated cyber attacks as part of their spectrum of response. This paper's objective is to identify some prudent limits to govern the incorporation of automated cyber attack as an instrument of policy in national and collective defense. For this paper, the concept of automated cyber attack focuses on nations' in-kind responses to strategic-level attacks by actors that use cyber means. The main aspects of the paper explore the theoretical roles of critical thinking in the development and operation of such systems. Topics include the context, points of view, and cognitive biases of the cyber actors; the assumptions and inferences inherit in their decision making; and the implications of decisions related to automated cyber attack.

The structure of research utilizes the Gerras critical thinking model to identify the factors to evaluate. It outlines how techniques such as the analysis using *Tallinn Manual* criteria may be used to identify assumptions and inferences for categorizing national response actions as cyber attack. It examines several historical incidents involving decisions related to strategic attack for implications to automated cyber attacks. It also investigates the implications of adopting a policy of cyber resilience, focusing on how it could be integrated with automated cyber responses measures. Finally, it studies the implications of automated cyber attack connected to the philosophy and ethics of evolving Just Cyber Warfare theory, such as that proposed by Taddeo.

Keywords: *critical thinking, escalation, resilience, automated response, attack assessment*

The views expressed are those of the author and do not necessarily reflect official policy or position of any Department or Agency within the U.S. Government. All information and sources for this paper were drawn from unclassified material.

1. INTRODUCTION

When contemplating the topic of cyber warfare, there is general consensus supporting the primacy of offensive over defensive actions [1]. In more common parlance, it is often said that "the best defense is a good offense." But should such a tenet be implemented in service of a nation's security in cyberspace? And how should this tenet be characterized in an environment where thrusts and parries may occur at network speeds? This paper's objective is to identify some prudent limits to govern the incorporation of automated cyber attack as an instrument of policy in national and collective defense. A key aspect of the paper is to explore the role of critical thinking in the development and operation of such systems.

2. CRITICAL THINKING

The framework for analysis in this paper utilizes the Gerras [2] model (derived from the work of Paul and Elder) which defines critical thinking as "deliberate, conscious, and appropriate application of reflective scepticism." Gerras applies the context-dependent school of thought and focuses on factors important to the decision making of strategic leaders. The model is broken into six main elements: clarify concern, point of view, assumptions, inferences, evaluation of information, and implications. These elements are considered to be interactive and are not necessarily linear or sequential in application of assessing the deliberate use of critical thinking.

The element of *clarify concern* concentrates on the desire to separate the root causes of problems from their symptoms; this should be done in such a way as to not preclude or limit potential responses. When evaluating the actions of nations, a significant aspect of the *point of view* element is egocentrism, which Gerras calls the "tendency to regard oneself or one's own opinions or interests most important." He offers four specific applications of this principle—egocentric memory (forgetting information that does not support one's thoughts); egocentric myopia (narrowing point of view in assessment to support one's thoughts); egocentric righteousness (considering one's thoughts to be superior); and egocentric blindness (disregarding information that does not support one's thoughts). Making *assumptions* is an inherent trait that humans use to provide boundaries for decision making; clearly stating and understanding such assumptions aids the critical thinking process. *Inferences* are logical perceptions of how available facts and evidence fit together in the environment being considered. In ideal applications, the *evaluation of information* is an objective process. However, decision makers often employ cognitive strategies such as heuristics ("rules of thumb") to simplify the process; but these useful tools may also introduce unknown and undesired biases.

Considering *implications* of any decision should include potential effects (desired and undesired) beyond or collateral to the anticipated outcomes.

3. RESPONSE AND ESCALATION

For this paper, the concept of automated cyber attack focuses on nations' in-kind responses to strategic-level attacks by actors that use cyber means. Such automated responses would go beyond merely defending or mitigating the effects on an ongoing attack, but would instead be an offensive or proactive counter-strike to thwart any future attacks. The intent to have a cyber attack response capability is made clear by such statements as General Keith Alexander's recent testimony [3] to the U.S. Senate as Commander, U.S. Cyber Command: "We feel confident that foreign leaders believe a devastating attack on the critical infrastructure and population of the United States by cyber means would be correctly traced back to its source and elicit a prompt and proportionate response." This paper is a theoretical study that assumes that the desire and technical capability to automate such cyber attacks is feasible in the near future.

A. *ASSESSING POTENTIAL ATTACKS*

It is critical to ensure a cyber attack has occurred before considering a cyber attack as an in-kind response. How does one differentiate a coincidental incident in cyberspace with negative consequences from an actual attack? One of the best tools supporting this complex task is the framework of the Schmitt [4] criteria which considers the intensity of damage in each of seven areas to provide a composite assessment of the effects of a potential cyber attack. These are considered in the perspective of *jus ad bellum* and compared to international norms and agreements such as those established by charters of the United Nations and the North Atlantic Treaty Organization (NATO) as well as humanitarian law [5]. These criteria have been further refined and expanded to eight areas in their recent adoption as an integral part of the *Tallinn Manual on the International Law Applicable to Cyber Warfare* [6]. Figure 1 [7] depicts the *Tallinn Manual* criteria and related elements as a framework to assess incidents in cyberspace which may put them into categories of hostile events ranging from use of force to armed conflict. If the determination is made for cyber attack, then any response should apply *jus in bello* tenets, such as those codified in the Law of Armed Conflict.

Even learned scholars may disagree on the practical application of this framework in complex and dynamic geopolitical environments. The implications of cyber attack characterization are potentially dangerous, as Ziolkowski [8] notes, "the threshold of endangering the (physical) security of a State is a high one and should not be

diluted." It may become a mostly academic issue if a nation opts to implement an automated cyber attack responses based on pre-determined indicators and criteria.

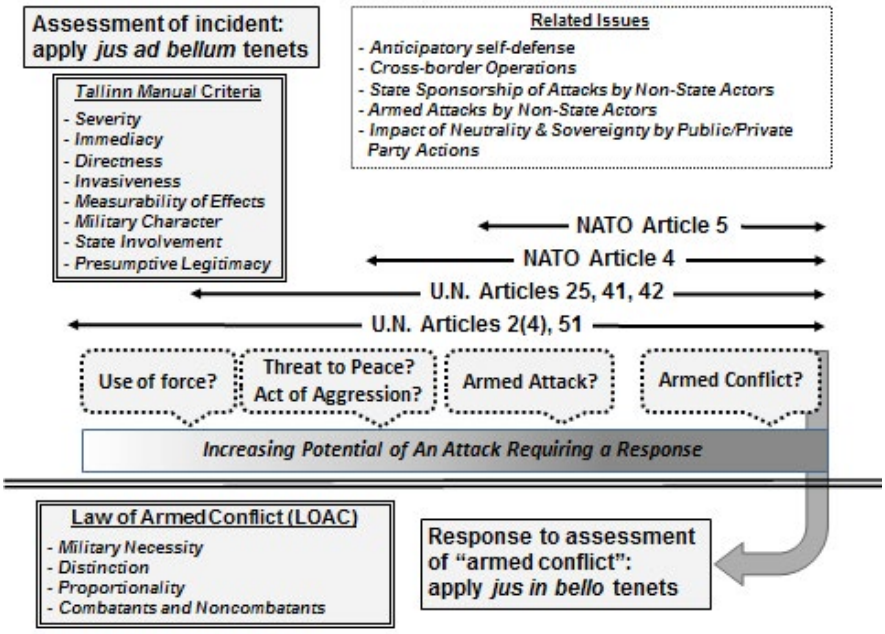

Figure 1. Cyber incident assessment and escalation

B. CONTEXT AND ACTORS IN THE ATTACK RESPONSE PROCESS

Once an incident is assessed to be an attack, the analysis shifts to consider appropriate response. This is accomplished at two levels: the immediate and local effects, and the long-term and global impacts. The Law of Armed Conflict (LOAC) establishes the international norms that define how the use of force in responses should be planned and implemented. Fanelli and Conti [9] explore cyber operations effects in terms of their severity and persistence. Examining longer term and global impacts may require the methodical exploration of the dynamic context of cyber attack to assess policy options for using either continuous or discrete automation. This evaluation should consider possible consequences that build upon previous outcomes and thus intensify global tensions. Such a framework is the Kahn [10] escalation ladder which codifies in 44 metaphorical rungs the range of nuclear-related conflict between nations from subcrisis maneuvering up through various manifestations of military and civilian central nuclear war.

Any response must consider the actor nations that will be targeted. Was the initial attack conducted by actors that were rational or irrational, or could it have been an accidental initiation? Does automated decision making take all these possibilities into consideration? Any actor in the process is capable of rational or irrational decisions and as Gerras [11] notes, "logically fallacious arguments can be psychologically compelling." Such critical thinking flaws may influence the design or operation of automated systems by propagating biases into the beliefs-desires-intentions (BDI) reflective properties of automated agents [12].

If dealing with rational actors, automated response may enhance cyber deterrence by punishment [13] or perhaps even enable cyber coercion [14]. However, even with rational state actors, there is a range of state responsibility for the cyber activity that occurs within their sovereign borders [15]. But is there really a legitimate concern that nations may not apply critical thinking to decision making for the use of strategic weapons? Before exploring the implications for cyber attack situations, let us first look at how automated defenses may have affected recent historical events not directly related to cyberspace.

4. LESSONS FROM RECENT HISTORY

The benefit of hindsight allows us to examine how errors and shortfalls in critical thinking almost led to catastrophic effects in three cases studies that occurred over the last three decades.

A. *ABLE ARCHER (1983)*

In November 1983, NATO initiated a command-post exercise to test the procedures and communications necessary for theater nuclear war in Europe. Many historians assert that this exercise culminated a series of events that accidently led the world to the brink of nuclear exchange akin to the Cuban Missile Crisis of 1962 [16]. As facts surrounding this case continue to come forward, it is still not clear how this eventually resolved itself as a fortuitous "non-event." Perhaps its origin and closure are best thought of as "normal accidents" [17]—that is, there was no single clear cause or effect.

In this case, the key concern to clarify was for the U.S.S.R. to determine if NATO was going to launch a pre-emptive nuclear attack using the *Able Archer* exercise as a cover for the preparation and initiation. The Soviet point of view included an aging leadership that was biased to view U.S. actions as part of a conspiracy to eliminate their country. U.S. President Ronald Reagan adopted a tough stance that included stationing intermediate range nuclear missiles in Europe coupled with the

new AirLand battle doctrine, perhaps due in part to Soviet deployment of SS-20 nuclear missiles. Both sides assumed the worst of the other's actions, setting in motion a vicious cycle of escalating mistrust and misinterpretation of events. The U.S. added Soviet political and military command structure to its nuclear targeting, inferring that it would induce caution in Soviet leadership. The U.S.S.R. inferred that they could prognosticate U.S. nuclear intentions based on the model of their Operation RYAN, which used extensive and diverse information gathering and indication-based decision making. Unfortunately, the model's design had inherent egocentric myopia and blindness which encouraged the reporting of potential crises [18]. Reagan later came to recognize his own misunderstanding of Soviet intentions that were also fueled by ethnocentrism. Fortunately, based on advice from his advisors, he agreed not to have himself or other principals in Washington participate in the exercise [19]. Hampered by biases, both sides appeared to be able to discern the others' *capabilities* but not *intentions*. Some historians contend that the role of a KGB agent turned by British intelligence provided the critical insight that prevented *Able Archer* from escalating to catastrophe [20]. Regardless, it appears that fortuitous circumstances rather than critical thinking prevailed.

B. NORWEGIAN RESEARCH ROCKET (1995)

Almost twelve years after *Able Archer* came another nuclear close-call between NATO and Russia. The routine launch of a research rocket on 25 January 1995 was mischaracterized as a possible prelude to nuclear attack on Russia [21]. The situation occurred during a time of increased tension between Russia and Norway (and perhaps the world in general) that caused failures in the critical thinking of tactical and strategic intelligence as well as communication systems.

The concern to clarify for Russian leadership was simple—was the Norwegian rocket the first step in a NATO nuclear attack? After the collapse of the Soviet Union, a Russian government was formed in 1991 with much of its military structure—the Strategic Rockets Forces, specifically—mostly intact, but declining in capability [22]. From their point of view, the fledgling Moscow leadership was struggling with governing crises, lingering Warsaw Pact issues, and a war brewing in Chechnya, while U.S. and allied efforts in Desert Storm were being hailed as successful examples of next-generation warfare. Russia assumed that the world was hostile to their new place on the global stage and that the well-publicized eastward expansion of NATO might be an existential threat. Also, Norway was pressing an old claim dispute for over 150,000 square miles of territorial waters that were rich in resources, further fueling speculation that it was becoming the preferred springboard for rapid deployment of Western forces into Russia.

Norway informed the Russian embassy in Oslo of their scientific rocket launch plans on 21 December 1994 and again on 16 January 1995; based on past experience, they inferred that this was sufficient to reduce risk between the countries. Unfortunately, the Norwegians also inferred that the launch would be monitored and assessed by the Russians in the same way as previous such launches (over 600 since 1962). But the new *Black Brant XII* was almost twice the size of any previous rocket, with specifications similar to a Pershing II nuclear missile; they did not consider how its longer range and higher trajectory might be viewed by Russian early warning assets. The immediate evaluation of the launch data was that the flight profile fit that of an electromagnetic pulse attack—the anticipated prelude to knock out Russian command and control systems before a nuclear strike. Unfortunately, the Norwegian launch notification did not get passed internally by the Russians to the proper military or civilian authorities and so the Russian nuclear launch briefcases were activated by President Yeltsin and General Kolesnikov as a precaution. While the exact details are still coming forth, it appears that these leaders waited for almost seven tense minutes until it was clear that the rocket was not headed toward Russia [23]. Fortunately for all, the Russian release of nuclear weapons still required deliberate initiation by its civilian leader.

C. CHINESE ANTI-SATELLITE WEAPON TEST (2007)

Going forward twelve years after the *Black Brant XII* launch there was another rocket flight with significant international implications. On 11 January 2007, China conducted its first kinetic-kill anti-satellite (ASAT) test, destroying its own Fengyun-1c weather satellite and causing extensive collateral damage of spacecraft debris that poses collision hazards for operational satellites. China miscalculated both the magnitude of the damage they would cause as well as the negative international ramifications [24].

The concern to clarify is to determine the purpose for China to conduct such a destructive test with long-term negative effects on the commons of space. From the Chinese point of view, this test was simply part of a larger ASAT program that included electronic jamming and laser dazzling of satellite systems. They may have assumed that it was an acceptable operation since the U.S. and U.S.S.R. both conducted similar destructive tests in the 1970s and 1980s with little residual effects [25]. China inferred incorrectly that the test would not cause long-term damage, despite the fact that it occurred at an orbital altitude significantly higher than other ASAT tests. It is unclear if the evaluation of the operation went beyond military leadership; China gave no advance warning of the test and did not issue a public statement until twelve days later [26]. The implications of this test are still significant six years later as other nations' satellites must contend with a more hazardous space environment; although China has less than 4 percent of the world's

active payloads on orbit, it accounts for almost 28 percent of the on-orbit debris, the majority of which was generated by this one event [27].

D. IMPLICATIONS FOR AUTOMATED CYBER ATTACK

Table I summarizes the key elements of the historical cases. In any of these vignettes, one must consider how the outcomes may have changed if the leaders' responses had been automated (by either side). These cases were selected to illustrate where lapses and fallacies in critical thinking leading up to the crises were actually contributing factors to the development of the actual crises. To examine how this might apply to situations where automated cyber attack may be considered, let us look at the critical thinking factors from three possible perspectives summarized in Table II. These theoretical analyses are illustrative, not comprehensive.

Table I. Summary of Historical Cases With Strategic Attack Issues

Critical Thinking Factors	**Examples of Strategic Attack Concerns from Recent History**		
	Able Archer (1983)	**Norway Research Rocket (1995)**	**Chinese Anti-Satellite (2007)**
Clarify Concern	NATO nuclear attack on USSR?	NATO nuclear attack on Russia?	Purpose of Chinese destruction of satellite?
Point of View	- USSR deploy SS-20s. - US tougher stance with nuclear weapons. - AirLand doctrine.	- NATO expansion. - Chechnya war. - Tensions between Norway & Russia.	- PRC: logical progression of military space development.
Assumptions	- US & USSR doctrines more aggressive. - Aging USSR leadership more offensive-minded.	- Hostile world opinion toward Moscow. - NATO making Norway a springboard for attack on Russia.	- PRC: no long-term damage expected? - ASAT development similar to that of US and USSR programs.
Inferences	- US nuclear targeting of USSR leadership. - USSR Operation RYAN use of indicators.	- Routine research rocket notification and launch. - Issues with new rocket size and trajectory.	- Failure at technical level (to predict collateral effects). - Failure at decision level to consider implications.
Evaluation of Information	Still debated. Reagan made right call not to have principals play. Possible intervention by Soviet spy.	Launch assessed as possible pre-emptive strike on Russian communication. Yeltsin made right call not to respond.	Wrong call by PRC to destroy satellite. Unclear if military leadership had permission of civilian leaders.
Implications	**Fortuitous Non-Event** as part of a vicious cycle of mistrust; escalated near to point of nuclear exchange.	**Benign Event** misinterpreted by military—almost to point of nuclear exchange.	**Serious Event** that polluted space environment and increased risks for all space-faring countries.

The first perspective is the U.S. internal view to clarify whether automated cyber attack is necessary for its existential defense. This could be framed by a point of view of cyberspace as a domain where attacks may occur at network speeds and may cause devastating surprise attacks (e.g., "cyber Pearl Harbor"). Assumptions may include current defenses being too slow and dispersed, and that their automation and centralization will increase their effectiveness. The inference is that the use of pre-determined indicators and automated cyberspace agents that can attack threat systems is sufficient and appropriate. If such a system is deployed, it may be difficult to determine when decision makers will know that an attack and response have occurred as well as what their role will be during the hostilities. The implications are that the value of the automated attack system must be viewed not only regarding their effects on tactical threats, but also on how they shape the strategic defense and deterrence posture.

The second perspective is that of U.S. allies view to clarify if automated cyber attack responses are suitable for collective or cooperative defense. A logical point of view is one where cyber attacks on one partner nation may affect all nations and that pooled resources for cyber defense will enhance the security of all. Allies may assume that automated responses may limit the extent of effects from adversarial attacks. They may also assume that design criteria and implementation methods can be shared to help ensure unity of effort. The inference is that the use of pre-determined indicators and automated cyberspace agents requires significant cooperation and coordination among allies. Evaluation of this inference raises issues regarding how the roles and responsibilities are assigned for the development, maintenance, and application of the automated capability. The implication is that, if properly implemented, the use of automated attack responses can improve collective cyber defense and deterrence.

Table II. Critical Analyses of Possible Automated Cyber Attack

Critical Thinking Factors	**Possible Perspectives of Automated Cyber Attack**		
	U.S. Internal	**U.S. Allies**	**Other Countries**
Clarify Concern	Necessary for existential defense of US?	Suitable for collective/ cooperative defense?	Level of threat posed by primary and collateral effects?
Point of View	- Attacks may occur at network speeds. - Devastating surprise attacks possible.	- Attacks on one partner may affect all. - Pooled resources will enhance security.	- US and allied attacks primarily for their own interests.
Assumptions	- Current defenses too slow and dispersed. - Centralized and automated defenses better.	- Automated responses can limit extent of attacks. - Design criteria and implementation can be shared.	- Automated responses have no direct control. - No warning provided in advance of their use.

Inferences	Use of pre-determined indicators and cyber-space agents is sufficient and appropriate.	Use of pre-determined indicators and cyber-space agents requires coordination among allies.	Do any of the US responses inadvertently violate national sovereignty?
Evaluation of Information	When do decision makers know an attack and response have occurred?	Who is responsible for the coordinated development and maintenance of automated response systems?	Can countries receiving collateral damage respond?
Implications	Cyber national deterrence and defense enhanced?	Collective cyber deterrence and defense enhanced?	Potential escalation by automated means?

The third perspective is that of other countries that may be concerned about the level of threat posed by primary (intentional) and collateral (unintentional) effects caused by the automated attack systems. They may have the point of view that the systems are designed to support interests other than their own, and assume that the automated responses have no deliberate control and thus will issue no advance warning of their use. The inference is that the automated attack response of others may inadvertently violate their own national sovereignty, thus giving cause to evaluation if they can respond in kind to any collateral damage absorbed. The implication is that such responses to automated responses may lead to a cycle of escalation largely driven by mechanisms detached from deliberate decision making.

E. *RECENT ACTIVITY REGARDING MILITARY CYBER RESPONSE*

General Keith Alexander's March 2013 testimony to the U.S. Senate [28] outlined recent activity of U.S. Cyber Command worthy of critical analysis. The concern to clarify is how the U.S. military will respond to activities perceived as cyber attack. Alexander stated, "the Department of Defense and U.S. Cyber Command are being integrated in the machinery for National Event responses so that a cyber incident of national significance can elicit a fast and effective response to include pre-designated authorities and self-defense actions where necessary and appropriate." The point of view with regard to "fast and effective responses" in unclear, but Alexander mentioned that the inter-agency and international exercise CYBER FLAG "introduced new capabilities to enable dynamic and interactive force-on-force maneuvers at net-speed." From this perspective, can "pre-designated authorities and self-defense actions" include automated responses? If so, who determines if they are "necessary and appropriate," and what criteria do they utilize?

Two implicit assumptions in the testimony are that traditional organizational

structures can handle the challenges in cyberspace and that negative events in cyberspace are threat-based. The inferences lead to traditional military approaches such as establishing three main levels of forces: a Cyber National Mission Force, a Cyber Combat Mission Force (supporting Combatant Commands), and a Cyber Protection Force (for DoD systems). These forces are pursuing normalized cyber operations for "a more reliable and predictable capability to be employed." Following such ethnocentric approaches may open vectors for manipulation by other actors. The evaluation of information includes the drive for increased operational awareness by such means as "a weekly Cyber Operating Directive (CyOD) across the DoD cyber enterprise…so that all 'friendlies' can understand what is happening in cyberspace." However, such useful measures may unknowingly foster ethnocentrism akin to the Operation RYAN activities surrounding *Able Archer*.

The implications are that U.S. cyber forces may be leaning toward a threat-based viewpoint of cyberspace that encourages the rapid identification and response to perceived aggressive action with little account for the broader dynamics of the information environment. But is this a realistic concern? The U.S. Department of State Legal Advisor, Harold Koh [29], stated the U.S. may legally respond to cyberspace activities "that amount to an armed attack or imminent threat thereof." Regarding capability and intent, he notes that the "United States has impressive cyber-capabilities" and "that adherence to established principles of law does not prevent us from using those capabilities to achieve important ends." Koh's views on the international legal aspects regarding the use of such capabilities is largely congruent with *Tallinn Manual* principles [30], and he stresses that the preferred use of such capabilities considers multilateral and regional issues.

5. POLICY RECOMMENDATIONS

Many other questions and implications can be examined using the critical-thinking framework. This section provides recommendations for cyberspace-related policy summarized from the historical and hypothetical cases as well as current trends examined above.

A. *ROLE OF RESILIENCE*

Although automated cyber response measures may provide added security and deterrence, they also risk interacting with other mechanisms and indicators that may create reactive and escalatory vicious cycles such as those in case studies. Perhaps, instead the focus should be on fostering resilience, such as that proposed in the U.S Department of Homeland Security's healthy cyber ecosystem model, specifically calling for cyberspace resilience in critical infrastructure as well as

business, social, and civic process [31]. The current NATO Policy for Cyber Defence [32] also lists resilience as an overarching principle (with prevention and non-duplication). Having sufficient resilience measures in place can provide strategic leaders with adequate time for critical thinking in their decision making. This can include evaluating information and options with the goal of keeping responses from becoming escalatory. Balancing the combination of resilience and automated responses should be evaluated in the context of a dynamic cyberspace environment where the success of a nation's strategy depends on the strategy of other nations, and their interaction and behavior will change the environment [33].

B. ETHICAL IMPLICATIONS

Most of the debate among nations regarding cyber attack in general—let alone when such attack is automated—focuses on protecting their fundamental national purpose and interests. Thus, expanding the decision making to include international repercussions may only be done through the lens of *realpolitik* pragmatism. However, to nurture a more open and cooperative cyberspace environment, nations should also contemplate an ethical-based framework, possibly adopting first principles for Just Cyber Warfare proposed by Taddeo [34]. These principles state that Just Cyber Warfare should only be waged "against entities that endanger or disrupt the well-being of the Infosphere" and that it seeks to preserve, but not necessarily promote, the well-being of the Infosphere.

Leder and others [35] examine the struggle between what is technically feasible for the application of automated responses and the concern that they "may interfere with law or current ethical beliefs depending on their invasiveness and impact on third-parties." They examine ethics issues related to automated and proactive botnets that target control servers, traffic, or infected systems. Other researchers cite legal opinions that conclude that applying automated methods, such as "white worms" which enter systems to disinfect them from malicious software may be illegal if they operate without express consent of the owners [36].

C. SIGNALLING BETWEEN NATIONS

If automated response options are being considered or are in place, this fact can be communicated to other countries as a sign of commitment as well as deterrence against escalation. Such clear signalling of intent among nations can help mitigate tension; as noted in the *Able Archer* case, knowledge of intent is more difficult to discern than knowledge of capability. Also, too much secrecy may work against clear deterrence and signalling whereas simple declaratory statements may enhance effectiveness [37]. For example, the announcement of the establishment of U.S.

Cyber Command by Secretary of Defense Robert Gates in June 2009 caught much of the world by surprise; it may have been more effective if it was coordinated with the State Department's diplomatic connections.

For like-minded states, communication could be enhanced by establishing terms of reference, such those explored by Prescott [38], regarding participation in cyber hostilities. Such communication should include factors regarding the nature of the diffusion and interdependence of cyber attacks across global regions [39]. It may also be useful to develop hypothetical escalation models such as a cyber form of the Kahn nuclear ladder; signalling may include publically stating where a nation has its automated responses enabled based on the details of such a construct. Such standards of nation signalling may help to form the basis to facilitate agreements that could eventually lead to formal cyber weapon treaties [40].

6. SUMMARY

A very dangerous event would be an accidental incident in cyberspace that was interpreted as an attack during a period of heightened tensions between two world powers--adding automated attack systems could make a bad situation into a catastrophic one. Since such incidents have occurred in the physical domains of warfare during the last three decades, it is reasonable to assume they will happen in cyberspace. The application of critical thinking can help mitigate risk, but only if time is available for leaders to reflect. Adopting a policy that emphasizes cyber resilience may help provide time for decision makers to thoughtfully consider the situation and weigh alternatives. If automated attack responses are deemed necessary they should be implemented in a graduated manner that is signalled to potentially hostile nations. Adopting ethical principles of just cyber war may provide overarching guidance for the development and deployment of automated cyber attack responses that strive to preserve the overall well-being of cyberspace while protecting nation purposes and interests.

REFERENCES

[1] T.Rid and P. McBurney. "Cyber-Weapons." RUSI Journal, vol. 157, no. 1, pp. 6-13, February/March 2012.

[2] S. Gerras. "Thinking Critically about Critical Thinking: A Fundamental Guide for Strategic Leaders." Carlisle, Pennsylvania: U.S. Army War College, August 2008.

[3] K. Alexander. Statement before the Senate Committee on Armed Services, Washington, D.C., 12 March 2013, pp. 3.

[4] J. Michel et al. “Measured Responses to Cyber Attacks Using Schmitt Analysis: A Case Study of Attack Scenarios for a Software-Intensive System.” *Proc. of Twenty-seventh Annual International Software and Applications Conference, IEEE*, 2003.

[5] M. Schmitt. “Attack as a Term of Art in International Law: The Cyber Operations Context.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 283-293.

[6] *Tallinn Manual on the International Law Applicable to Cyber Warfare*. General editor M. Schmitt. New York: Cambridge University Press, 2013.

[7] J. Caton. “Cyberspace and Cyberspace Operations” in *Information Operations Primer: Fundamentals of Information Operations*, AY12 Edition. Carlisle, Pennsylvania: U.S. Army War College, November 2011, pp. 19-32.

[8] K. Ziolkowski. “*Ius ad bellum* in Cyberspace-Some Thoughts on the “Schmitt-Criteria” for Use of Force.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 295-309.

[9] R. Fanelli and G. Conti. “A Methodology for Cyber Operations Targeting and Control of Collateral Damage in the Context of Lawful Armed Conflict.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 319-331.

[10] H. Kahn. *On Escalation: Metaphors and Scenarios*. New York: Praeger, 1965.

[11] S. Gerras. “Thinking Critically about Critical Thinking: A Fundamental Guide for Strategic Leaders.” Carlisle, Pennsylvania: U.S. Army War College, August 2008, pp. 9.

[12] E. Tyugu. “Command and Control of Cyber Weapons.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 333-343.

[13] K. Geers. Strategic Cyber Security. Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence, 2012, pp. 119.

[14] F. Hare. “The Significance of Attribution to Cyberspace Coercion: A Political Perspective.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 125-139.

[15] J. Healey. *Beyond Attribution: Seeking National Responsibility for Cyber Attacks*. Washington D.C: The Atlantic Council, 2011.

[16] A. Manchanda. “When Truth is Stranger Than Fiction: The *Able Archer* Incident.” *Cold War History*, vol. 9, no. 1, pp. 111-133, February 2009.

[17] T. Czerwinski. *Coping with the Bounds: Speculation on Nonlinearity in Military Affairs*. Washington, D.C.: National Defense University, 1998, pp. 98-99.

[18] A. Manchanda. “When Truth is Stranger Than Fiction: The *Able Archer* Incident.” Cold War History, vol. 9, no. 1, pp. 117-118, February 2009.

[19] D. Hoffman. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. New York: Doubleday, 2009, pp. 94-96.

[20] B. Fisher. “Anglo-American Intelligence and the Soviet War Scare: The Untold Story.” *Intelligence and National Security*, vol. 27, no. 1, pp. 75-92, February 2012.

[21] P. Pry. *War Scare: Russia and America on the Nuclear Brink*. Westport, Connecticut: Praeger, 1999, pp. 214-241.

[22] D. Hoffman. The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy. New York: Doubleday, 2009, pp. 399-400.

[23] P. Pry. *War Scare: Russia and America on the Nuclear Brink*. Westport, Connecticut: Praeger, 1999, pp. 214-241.

[24] S. Kan. *China's Anti-Satellite Weapon Test*. CRS Report for Congress RS22652, Washington D.C.: Library of Congress, 23 April 2007.

[25] D. Ball. *Assessing China's ASAT Program*. Austral Special Report 07-14S. Melbourne: RMIT University, 14 June 2007.

[26] S. Kan. China's Anti-Satellite Weapon Test. CRS Report for Congress RS22652, Washington D.C.: Library of Congress, 23 April 2007.

[27] "An update of the FY-1C, Iridium 33 and Cosmos 2251 Fragments." *Orbital Debris Quarterly News*, vol. 17, no. 1, pp. 4-7, January 2013.

[28] K. Alexander. Statement before the Senate Committee on Armed Services, Washington, D.C., 12 March 2013.

[29] H. Koh. "Remarks as Prepared for Delivery By Harold Hongju Koh to the USCYBERCOM Inter-Agency Legal Conference, Ft. Meade, MD, Sept 18, 2012." *Harvard International Law Journal (Online)*, vol. 54, pp. 1-12, December 2012.

[30] M. Schmitt. "International Law in Cyberspace: The Koh Speech and Tallinn Manual Juxtaposed." *Harvard International Law Journal (Online)*, vol. 54, pp. 13-37, December 2012.

[31] "Enabling Distributed Security in Cyberspace: Building a Healthy and Resilient Cyber Ecosystem with Automated Collective Action." Washington, DC: Depart. of Homeland Security, Mar. 23, 2011, pp. 8, 26.

[32] "Defending the Networks: The NATO Policy on Cyber Defence." Brussels, Belgium: NATO Public Diplomacy Division, Jun. 2011.

[33] R. Jervis. "From Complex Systems: The Role of Interactions." in *Coping with the Bounds: Speculations on Nonlinearity in Military Affairs*, Washington, D.C.: National Defense University, 1998, pp. 259-277.

[34] M. Taddeo. "An Analysis For A Just Cyber Warfare." *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 209-218.

[35] F. Leder et.al. "Proactive Botnet Countermeasures: An Offensive Approach." *Proc. of the Conference on Cyber Warfare*, 2009, pp. 12-14.

[36] L. Vihul et al. *Legal Implications of Countering Botnets*, Joint report from the NATO Cooperative Cyber Defence Centre of Excellence and the European Network and Information Security Agency (ENISA). Tallinn: CCDCOE, 2012, pp. 45-46.

[37] D. Alperovitch. "Toward Establishment of Cyberspace Deterrence Strategy." *Proc. 3rd International Conference on Cyber Conflict*, 2011, pp. 87-94.

[38] J. Prescott. “Direct Participation in Cyber Hostilities: Terms of Reference for Like-Minded States?” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 251-266.

[39] Kim et al. “Comparative Study of Cyberattacks.” *Communications of the ACM*, vol. 55, no. 3, pp. 66-73, March 2012.

[40] L. Arimatsu. “A Treaty for Governing Cyber-Weapons: Potential Benefits and Practical Limitations.” *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 91-109.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

The Dawn of Kinetic Cyber

Scott D. Applegate
Center for Secure Information Systems
George Mason University
Fairfax, Virginia 22030
sapplega@gmu.edu

Abstract: Cyber attacks are often called non-violent or non-kinetic attacks, but the simple truth is that there is a credible capability to use cyber attacks to achieve kinetic effects. Kinetic Cyber refers to a class of cyber attacks that can cause direct or indirect physical damage, injury or death solely though the exploitation of vulnerable information systems and processes. Kinetic cyber attacks are a real and growing threat that is generally being ignored as unrealistic or alarmist. These types of attacks have been validated experimentally in the laboratory environment, they have been used operationally in the context of espionage and sabotage, and they have been used criminally in a number of attacks throughout the world. While these types of attacks have thus far been statistically insignificant, the rapid growth and integration of cyber physical systems into everything from automobiles to SCADA systems implies a significant kinetic cyber threat in the near future. It is imperative that the security community begin to take these types of threats seriously and address vulnerabilities associated with cyber physical systems and other devices that could be utilized to cause kinetic effects through cyber attacks.

Keywords: *kinetic cyber, cyber attacks, cyber conflict, cyber warfare*

1. INTRODUCTION

In the box office hit, *Live Free or Die Hard*, actor Bruce Willis takes on a group of cyber terrorists who begin systematically shutting down the United States by conducting cyber attacks and exploitation of critical infrastructure systems. In the midst of the movie, the main antagonist uses cyber attacks to inflict massive physical damage, injuries and death. While this kind of cyber inflicted mayhem currently remains in the realm of screenwriters and science fiction authors, the concept of inflicting physical damage, injury or death through *Kinetic Cyber* is no longer just a fictional construct of creative minds. Kinetic Cyber refers to a class of cyber attacks that can cause direct or indirect physical damage, injury or death solely though the exploitation of vulnerable information systems and processes. There have been a number of cyber attacks and laboratory experiments over the course of the last decade that foreshadow the dawn of kinetic cyber as the logical evolution of cyber warfare.

Kinetic cyber attacks are a real and growing threat that is generally being ignored as unrealistic or alarmist. Regardless of the views of the doubters and naysayers, there is a growing body of evidence that shows kinetic cyber to be a valid and growing threat. These types of attacks have been validated experimentally in the laboratory environment, they have been used operationally in the context of espionage and sabotage, and they have been used criminally in a number of attacks throughout the world. It is imperative that the security community begin to take these types of threats seriously and address vulnerabilities associated with cyber physical systems and other devices that could be utilized to cause kinetic effects through cyber attacks.

2. CYBER PHYSICAL SYSTEMS

Generally, the main targets for kinetic cyber attacks are cyber physical systems (CPS). CPS refers to the tight conjoining of and coordination between computational and physical resources. CPS is the integration of computer systems with physical processes and its applications have the potential to dwarf the information technology revolution of the last few decades [1]. “The economic and societal potential of such systems is vastly greater than what has been realized, and major investments are being made worldwide to develop the technology” [1]. CPS technologies are being integrated across a broad spectrum of industry sectors. These systems can be found in medical devices, traffic control and safety, advanced automotive systems, process control, energy conservation, environmental control, avionics, instrumentation, critical infrastructure control (electric power, water resources, and communications systems for example), distributed robotics, defense systems, manufacturing, and smart structures [1].

Unfortunately, like other information technologies, most were originally designed with little or no security, or security has been added after the fact. Many of these systems rely on the security-through-obscurity concept rather than building security into the design process. For example, of the 40 plus position papers presented at the National Science Foundation's Workshop on Cyber Physical Systems in 2006, only two actually focused on security aspects of CPS and these were more concerned with the networks that support these systems rather than the actual systems themselves [2], [3]. Furthermore, none of the presentations or working groups directly addressed the security requirements of these systems.

CPS technologies are designed to have kinetic effects. They are designed to monitor and control physical processes through the use of computers and information technology. To a hacker or to someone who thinks outside-the-box, the mere fact of their existence and their interconnection to cyberspace implies that they could be manipulated and used for purposes other than those they were intended for. That is exactly what is happening. Hackers and security researchers are exploring the limits of these technologies and, as will be shown below, manipulating them to cause kinetic cyber effects both in the laboratory and in the real-world.

3. VALIDATION OF KINETIC CYBER

Cyber attacks are often called non-violent or non-kinetic attacks, but the simple truth is that there is a credible capability to use cyber attacks to achieve kinetic effects. Kinetic cyber attacks have been around for at least a decade and the ability to conduct these types of attacks has been validated in the laboratory environment through experimentation; in the operational environment to sabotage physical devices; and in the wild by hackers, hacktivists and other malicious actors.

A. EXPERIMENTAL VALIDATION

Security researchers love to find new and interesting ways to manipulate technology and are very good at thinking outside-the-box. For example, during an experiment to see if they could hack the firmware on a laser printer, it occurred to security researchers Salvatore Stolfo and Ang Cui that they might be able to manipulate the printer in such a way as to start a fire [4]. While they were unable to accomplish this due to thermal safety switches built into the printer's hardware, the mere fact that they thought to attempt this feat is very demonstrative of the types of experiments that are happening in laboratories and research facilities throughout the world. Whether it is trying to see if you can use a printer to start a fire, or determining what systems on a modern automobile can be hacked and controlled remotely, simple curiosity often drives security researchers to see how they can exploit vulnerable

technology to do things it was never intended to do, often with very dangerous consequences.

1) Project Aurora

The Department of Homeland Security (DHS) conducted an experiment in 2007 in which security researchers hacked into a replica of a power plant's control system to see if they could shut down a large generator. The Experiment, dubbed Project Aurora, was conducted at the Department of Energy's Idaho laboratory and its dramatic results were released on video showing a generator spewing smoke and shaking itself to death over the course of about 30 seconds [5]. Researchers conducting the experiment changed the operating cycle of the generator which sent it out of control and resulted in catastrophic damage [5]. This type of attack could cause enormous damage if it were used to attack an actual operating electrical power plant. Beyond the immediate damage to the generator itself, the time and cost to replace these large, industrial turbines is immense and it could take months for a power plant to come back online if a successful attack resulted in this type of damage. Such an attack could have enormous economic consequences for the region served by a targeted power plant if it were successful. There has never been a publically acknowledged, successful cyber attack against a power plant, but the result of this experiment did alarm officials both in the energy sector and in government. The power industry has long been aware of the potential threat that cyber attacks might pose and has voluntarily adopted higher information security standards than most other sectors. Additionally, some vulnerabilities associated with this experiment have since been addressed according to Robert Jamison, then acting undersecretary of DHS's National Protection and Programs Directorate [5].

2) Hacking Medical Implants

In 2008, security researchers at the Harvard Medical School's Beth Israel Deaconess Medical Center in Boston, the University of Massachusetts Amherst and the University of Washington in Seattle raised alarms that implantable cardioverter defibrillators (ICD) or other medical implants could be vulnerable to hacking with devastating consequences [6]. These researchers cautioned that ICDs and pacemakers could be maliciously reprogrammed to fail "to shock a speeding heart or, conversely, jolts one that is beating normally" [6]. These devices could be remotely accessed using wireless technology and a laptop computer and most used only an unencrypted username and password to secure access. In many cases, the password was simply the device's serial number. These researchers also showed that you could easily intercept data wirelessly from these devices including the patient's name, date of birth, medical ID number, patient history, the name and phone number of the treating physician, the date of ICD implantation, the model, and the serial number of the ICD [7]. Researchers from this same study published

a series of recommended security measures to make implantable medical devices more secure, yet four years later, these devices were still demonstrably hackable [7], [8].

Security Researcher Barnaby Jack recently presented positive proof at the 2012 Breakpoint security conference that ICDs and pacemakers were still highly vulnerable to exploitation. Unlike the previous study, Jack actually demonstrated the ability to deliver a deadly 830-volt jolt to a pacemaker by logging into it remotely after hacking it [8].

> [Mr Jack] found the secret command doctors use to send a "raw packet" of data over the airwaves to find any cardioverter-defibrillator or pacemaker in range and have it respond with its model number and serial number. This information allows them to authenticate a medical device to receive telemetry data and perform commands or software updates [8].

A malicious actor could issue commands to an IDC to jolt the heart, as Mr Jack showed in his demonstration, or to not respond to a failing heart in an emergency. Worse, Mr Jack stated that "it would be possible to write a worm for one particular brand of pacemaker and defibrillator, then have it spread to other devices within range, from person to person" [8]. The only thing preventing these types of attacks, especially for a sophisticated actor such as a nation-state, is the will and motivation to do them. Mr Jack's research showed that medical implant technology is designed to be easily accessible, does not use encryption, is remotely accessible from a distance of up to 12 meters and can have life-threatening implications if abused.

3) CarShark

In 2010, security researchers from the University of Washington, Seattle and the University of California, San Diego conducted two studies on modern automobiles to see what systems could be hacked and exploited [9]. The research was conducted in three phases using bench testing, stationary vehicles and road tests to validate each attempted exploit. The study demonstrated "the ability to adversarially control a wide range of automotive functions and completely ignore driver input - including disabling the brakes, selectively braking individual wheels on demand, stopping the engine, and so on" [9]. To facilitate their experiment, the researchers wrote a custom tool designed to act as a bus analyzer and packet injector on Controllable Area Networks. This tool was called CarShark [9]. While the initial experiment was very successful and researchers were able to control dozens of functions in the car from locking and unlocking doors to disabling brakes at high speeds, the initial design of the experiment involved only direct physical access to the car. Researchers had to hook a laptop directly to the on-board diagnostics port in order to exploit the various automotive functions [9]. Researchers received so many questions on whether these

exploits could be accomplished remotely that they conducted a follow-on study to validate the ability to do just that.

In their follow-on study, researchers examined the potential attack surfaces of a modern automobile and determined that "remote exploitation is feasible via a broad range of attack vectors (including mechanics tools, CD players, Bluetooth and cellular radio)" [10]. They further showed that all of the exploits demonstrated in their initial study could be exploited by means of any of these attack vectors and "that wireless communications channels allowed long distance vehicle control, location tracking, in-cabin audio exfiltration and theft" [10]. There is little doubt that using the techniques demonstrated in these two studies it would be possible to seriously injure or kill the occupants of a vehicle. Turning off the headlights and disabling the brakes on a vehicle driving at highway speeds at night could easily result in a life threatening accident. The ability to do this remotely combined with the ability to set the malware to self-delete after an accident would make it very difficult for investigators to discover this type of attack, especially if they were not actually looking for it in the first place.

While there has been a great deal of work done by researchers in laboratory settings, the use of kinetic cyber is not limited solely to experimentation. Kinetic cyber attacks have been used by curious teenagers, hackers, criminals, and disgruntled employees in the real-world and many of these activities actually precede the more formal work done in labs.

B. REAL-WORLD VALIDATION

Activists, terrorists or criminals are always looking for new and innovative techniques to accomplish their goals and this is just as true in cyberspace as it is in the physical domain. There have been a number of criminal cyber attacks over the last decade that have directly resulted in kinetic effects. Many of these kinetic cyber attacks predate the experiments discussed above. The idea of causing physical damage using cyber attacks is not new; it has simply been relegated to obscurity as an outlier or an aberration. The incidents discussed below demonstrate that kinetic cyber capabilities do exist and are being used by hackers ranging from curious teenagers to disgruntled employees.

1) Maroochy Water Services, Queensland Australia

Starting in February of 2000, Vivek Boden, a 49 year old disgruntled utility worker, waged a three-month long hacking campaign against Maroochy Water Services and the Maroochy Shire Council in Queensland, Australia [11]. Boden was a former employee of Hunter Watertech, an Australian firm that installed supervisory

control and data acquisition (SCADA) systems and he had been a member of the team that had designed and implemented the SCADA systems for Maroochy Water Services. After leaving Hunter Watertech on poor terms, Boden had applied for and been denied a job by the Maroochy Shire Council. In an act of revenge for being denied the job, Boden began hacking the very SCADA systems he had helped install and released over 264,000 liters of raw sewage at a variety of locations over the course of the next three months [12]. This attack led to damage of the local environment and unhealthy conditions for the local residents. "Marine life died, the creek water turned black and the stench was unbearable for residents," said Janelle Bryant of the Australian Environmental Protection Agency [13]. Boden was eventually caught, charged, convicted and sentenced to two years in jail. Boden's series of attacks is one of the first to have caused physical damage solely through the use of information systems.

2) Los Angeles Traffic Management Center, Los Angeles, California

Over two days in late August 2006, striking traffic engineers from the Engineers and Architects Association picketed the Los Angeles City Hall demanding a better pay raise than the city was offering them over the next three years [14]. City officials, fearing that the striking workers would cause chaos with the city's traffic system, took steps to block access for the striking engineers. Two traffic engineers, Gabriel Murillo and Kartik Patel, managed to bypass this effort and hacked into the system causing gridlock at four key intersections in the city over the next several days [15]. Although access had been blocked for the striking engineers, access remained in place for top managers and one of the engineers was able to illicitly log into the system using one of his managers' credentials. Murillo and Patel then targeted four key intersections and extended the timing of red lights for the most congested approaches to these intersections causing traffic to come to a virtual standstill [16]. "Cars backed up at Los Angeles International Airport, at a key intersection in Studio City, at access onto the clogged Glendale Freeway and throughout the streets of Little Tokyo and the L.A. Civic Center area" [17]. Although there were no accidents attributed to this incident and therefore no physical damage or injuries, it is not a far stretch of the imagination to see that hacking into traffic control systems could easily result in kinetic effects. There is a large body of knowledge available on the Internet in regards to hacking traffic lights, and while this incident involved an insider threat, traffic lights and traffic management control systems are a popular target among hackers. Murillo and Patel were caught, charged with seven felonies between them and eventually sentenced to serve 240 hours of community service and fines amounting to $6000 dollars [17].

3) Tramways, Lodz, Poland

In January of 2008, a 14-year-old Polish teenager rewired a television remote control to interact with the wireless switch junctions on the Lodz city tram system. The teenager then used the remote control to reroute trams and essentially turned the tram system into his own personal train set [18]. The problem was discovered when a driver attempting to steer his vehicle to the right was involuntarily taken to the left. As a result the rear wagon of the train jumped the rails and collided with another passing tram. "The rear wagon then swung off the rails and crashed into another passing tram, hurling screaming passengers to the floor" [19]. The teen's actions caused the derailment of four vehicles and resulted in minor injuries to more than a dozen passengers. Lodz "transport employees were reported as saying that they knew immediately that someone outside their staff had caused the accident" [19]. This attack, although only done as a prank, is significant in that it was the first cyber attack to directly cause injuries.

C. OPERATIONAL VALIDATION

Kinetic cyber attacks have the potential to become very dangerous or even game-changing technologies in cyber warfare and other aspects of cyber conflict. The CPS that kinetic cyber generally targets are highly lucrative in terms of strategic value, and the ability to degrade, damage or destroy such systems represents a valuable weapon to a nation-state's arsenal. While only one such kinetic cyber attack is publically known to have been used at the present time, it would be dangerously short-sighted to believe that more such weapons are not currently in development. The Stuxnet attack against Iran in 2010 serves as operational example of the use of kinetic cyber-weapons and its success, however limited, has ushered in a new arms race among nation-state developing cyber warfare programs.

1) Stuxnet

In 2010, stories began to emerge in the media of a new worm that was described as the first cyber-weapon – a piece of targeted malware designed specifically to find and destroy specific physical devices. The Stuxnet worm was more complex than any previously discovered piece of malware. It contained four Windows zero-day exploits and was able to propagate itself through USB flash drives, network shares, a remote procedure call (RPC) vulnerability or a print spooler vulnerability [20]. Stuxnet was also the first piece of malware ever identified to include a programmable logic controller (PLC) root kit. Stuxnet spread itself via Microsoft Windows but appeared to target a specific PLC, the Siemens S7-300 system, and only if that PLC was attached to two specific types of variable-frequency drives which had to be spinning between 807 to 1210 Hz [20]. Once these and other specified conditions

had been met, the Stuxnet worm would periodically modify the frequency of the variable-frequency drives to 1410 Hz and then to 2 Hz and then to 1064 Hz while simultaneously masking these changes from attached monitoring systems [20].

The Stuxnet virus is known to have infected at least 120,000 Microsoft Windows systems worldwide, however, it is only known to have damaged systems in the Fuel Enrichment Plant in Natanz, Iran. This has led to popular speculation that the Stuxnet worm was designed to specifically target this facility. Although exact numbers have not been released by Iran, it is believed that Stuxnet damaged more than 1000 centrifuges used in Iran's nuclear fuel enrichment program [21]. While Stuxnet remains the only kinetic cyber-weapon that has thus far been seen in the wild, its discovery legitimizes the use of kinetic cyber in an operational context. The use of Stuxnet will have long-term implications in cyber warfare. As retired General Michael Hayden put it, "We have entered into a new phase of conflict in which we use a cyber-weapon to create physical destruction, and in this case, physical destruction in someone else's critical infrastructure" [22]. In essence, Stuxnet has opened Pandora's Box when it comes to the militarization of kinetic cyber technologies, and now that it is open, there is no going back. Nation-states around the world will look at this event as legitimizing the use of kinetic cyber in the international arena and will begin integrating these technologies into their own cyber warfare programs.

The above examples illustrate that kinetic cyber is a valid and credible threat. Security researchers are finding new ways to exploit vulnerable CPS to achieve kinetic effects beyond those intended by design. Hackers, cyber-criminals and hacktivists are actively exploring information systems with cyber physical connections and attempting to cause kinetic effects. This leads to the question of how these types of attacks may evolve in the future.

4. THE FUTURE OF KINETIC CYBER

Major investments, development and research are currently being conducted in the area of CPS and these types of systems are becoming more pervasive in industrialized states. The growth of CPS implies that the probability of seeing more kinetic cyber attacks targeting these types of systems is going to grow. Taking into account the types of attacks and research that has already occurred, it is not difficult to extrapolate the direction that kinetic cyber could take. The most dangerous avenue of growth would appear to be in the areas of SCADA, implantable medical devices, and automotive technologies although there are certainly other areas that are ripe for exploitation.

From the perspective of a nation-state, the ability to do serious damage to a rival state's critical infrastructure represents a strategic advantage. If an attack were able to successful damage a significant number of large electrical power plants in a manner similar to the Project Aurora experiment, the consequences could be economically destabilizing to the target state. Replacing the electrical generators in these types of plants can take months and cost millions of dollars per generator. In the meantime, the customers served by these plants would remain without power. Economist Scott Borg noted that if an attacker managed to knockout power to a third of the United States for a period of three months, the economy cost would be upwards of 700 billion dollars which is the economic equivalent of 40 to 50 large hurricanes hitting at the same time [5]. This type of attack would be economically devastating and would have significant long-term consequences. While it is unlikely that a state would engage in this type of large-scale attack outside the bounds of an openly declared war, it would also be short-sighted to assume that only states will have access to these types of attacks.

Looking at the subversion of implantable medical devices or automobile control systems, these technologies could easily be exploited to injure or kill individuals or even groups of people. Such a use of kinetic cyber could be employed for murder or assassination of key figures. What makes this approach particularly insidious is that investigators would probably not realize there was a cyber-component to these actions. Given the number of car accidents in a typical year, it is not beyond reason to assume that investigators would simply accept that a mechanical failure had caused a fatal accident rather that some form of cyber attack. This is especially true if the exploit leaves little or no residue of itself in the system after the fact. Since there have been no known incidents of cyber attacks causing car accidents, why would an investigator even suspect that this might be the case? The same is true of implantable medical devices. A recent article in Fire Engineering magazine points out that there is a possibility that arsonists may find a way in the near future to start fires using cyber attacks and that arson investigators would be highly unlikely to look for this as an underlying cause of a fire [23]. These types of incidents could be going on today and there is very little chance that they would be discovered.

The potential use of kinetic cyber by criminals or as a means of engaging in cyber warfare is only limited by the ability of hackers and researchers to approach these technologies from an unconventional and innovative direction. These systems already have the capability to produce physical effects; it is therefore possible to subvert their functionality to do new and potentially dangerous tasks. Given the pervasive nature of network technology and the convergence of networked systems with cyber physical devices, these types of attacks are going to become far more common in the near future and the security community needs to begin addressing this problem now.

5. ADDRESSING THE GROWING THREAT

One of the first steps that should be taken in addressing the threat of kinetic cyber is to begin hardening CPS since these systems are often the main target of this type of attack. Security in CPS has followed the same trend that has been seen throughout the information technology industry. CPS devices were originally designed with little or no security. As a credible threat has emerged against CPS devices, designers and security researchers have begun to look at better ways to protect these vital systems. In 2012, the National Institute of Standards and Technology (NIST) held their first workshop on Cyber Physical Systems Security in Gaithersburg, Maryland. This was a two-day event with presentations and working groups focusing on a variety of industry areas such as smart power grids, SCADA, implantable medical devices and modern automobiles.

During the course of the NIST conference, a number of consistent themes emerged across all sectors of CPS. First and foremost was the need to create digitally signed and trusted instruction sets for cyber physical devices. Currently most CPS devices will accept instruction sets from any source so long as they have the correct format and syntax. This leaves devices highly vulnerable to exploitation through man-in-the-middle attacks and attacks which leverage packet injection such as those used in the CarShark experiment. Another suggested avenue of research involves the development of intrusion detection systems and reputation management systems for specific types of SCADA infrastructure such as smart power grids [24]. These types of security systems are vital in an environment where not all data that is received by a CPS device can be trusted.

The above recommendations could be added to existing CPS technologies, however, that is not an ideal solution. Manufacturers and developers of these technologies must strive to build robust security into cyber physical devices throughout all stages of their development lifecycle. Security that is baked in throughout the systems development lifecycle is generally more effective than security that is bolted on after the fact. This is true for both the software that runs these systems, and the hardware platforms and devices that CPS run on. Developing hardware level security for CPS can act as a final safety barrier against compromise and exploitation [25]. Another important aspect of CPS that requires attention is sensor data. CPS devices base many of their functions on real-time feedback from sensors. Researchers should focus development efforts on specific controls to ensure sensor and monitor data is protected in terms of integrity and availability [26]. As demonstrated in the Stuxnet attack, the ability to corrupt sensor and monitor data can blind operators to a problem in the midst of an attack and allow greater damage to occur before a compromise is discovered.

Implantable medical devices (IMD) represent another growing segment of CPS technologies which, due to the very restrictive environment they operate in and the critical nature of their functions, will need very specialized security protocols. Restrictions on size, power consumption and processing power preclude many traditional security applications, but developers must take security into account when designing these devices. These devices are regulated in the United States by the Food and Drug Administration (FDA). While the FDA does do some testing to ensure IMDs perform in accordance with written specifications in a safe and effective manner, they do not do security testing of these devices in the context of information security and assurance. Inclusion of security and resilience testing in the testing guidelines for implantable medical devices should be a top priority for security researchers in the medical community [27]. Additionally, a review of authentication and access control protocols for IMDs should be conducted to ensure they balance adequate protection with the need for emergency access by medical personnel [27]. As noted in the study by Barnaby Jack, many of these devices currently have access controls that are trivial to bypass. One area that could assist in efforts to strengthen authentication and access control is the development of suitable encryption technologies. Development of appropriate cryptographic techniques that could be applied where necessary in the restricted operation environment of IMDs would make it much more difficult for a malicious actor to wirelessly eavesdrop and steal credentials for these devices. Security for IMDs will require a delicate balance between confidentiality and availability since too much security on these types of devices could hinder doctors in an emergency situation, endangering the patient's life. However, as Mr Jack showed in his experiment, a lack of proper security could be equally dangerous.

Moving beyond technical solutions, it is important for policy makers, standards bodies, and governments to create reasonable and effective regulatory schemes to address security requirements in CPS. These devices are used in many sectors considered to be critical infrastructure. Industry has traditionally been resistant to new regulations and that will probably be the case with the CPS industries as well. That having been said, industry has the opportunity to take the initiative and voluntarily establish industry standards for security of CPS [25]. Doing so can serve to stave off overly restrictive efforts by government regulators and will allow the industry to shape the standards as they move forward. In additional to new regulatory schemes, governments and international bodies need to begin addressing kinetic cyber through diplomatic and legal efforts. Honest and open dialogue is needed in the international community to codify the definition of kinetic cyber and to establish thresholds for when these types of activities qualify as a use of force. Thus far, the international community has mostly avoided addressing cyber warfare and cyber conflict under the laws of armed conflict; however, the growing threat

of kinetic cyber should spur new efforts to address these issues in a meaningful and thoughtful manner. It would be better to tackle this issue now, before a major kinetic cyber event happens, rather than trying to address the issue in the passion and turmoil that often follows such events.

These recommendations merely represent a good starting point for addressing the threat of kinetic cyber. There is a great deal of additional research that needs to be done to develop and implement technical solutions to address threats to CPS. In addition to technical solutions, policy makers, both domestically and in the international community, need to create common sense regulations for the CPS industry and begin to explore legal frameworks for codifying and addressing kinetic cyber.

6. CONCLUSIONS

Kinetic cyber is a real and growing threat. Numerous experiments have shown that it is possible to subvert CPS to cause damage, injury or even death under the right circumstances. Real-world incidents over the course of the last decade have validated this concept as curious hackers and disgruntled employees have exploited vulnerabilities in CPS devices to cause physical damage and injuries. Stuxnet has operationally validated this concept as well in its use of kinetic cyber attacks to damage more than a thousand centrifuges at the Natanz fuel enrichment facility in Iran.

Kinetic cyber mainly exploits vulnerabilities in CPS. Designers and manufacturers of these technologies need to incorporate better security controls into these systems beginning at the requirements and design stage of the systems development lifecycle and proceeding through the entire process to retirement. Beyond technical solutions, policy- and lawmakers should begin to address this issue through new industry standards and regulations. The international community must also act to codify cyber warfare and cyber conflict under international agreements and the laws of armed conflict. While many would discount the idea of kinetic cyber as unrealistic, the events that have occurred thus far represent the beginning of these tactics and foreshadow more dangerous attacks ahead. It is important to tackle the problem of kinetic cyber now, in its infancy, before development of these technologies leads to more serious and deadly outcomes.

REFERENCES

[1] E. A. Lee, “Cyber Physical Systems: Design Challenges,” 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC), pp. 363–369, May 2008.

[2] D. Kazakos, “Position Paper : Robust Communications Networks with Imbedded Security,” in National Science Foundation on Cyber Physical Systems, 2006.

[3] J. C. (Steve) Liu, “Secure plug and play architectures for cyber-physical systems A Position paper for the NSF workshop on cyber-physical systems,” in National Science Foundation on Cyber Physical Systems, 2006.

[4] A. Cui, M. Costello, and S. J. Stolfo, “When Firmware Modifications Attack : A Case Study of Embedded Exploitation,” in 20th Annual Network & Distributed System Security Symposium, 2013.

[5] J. Meserve, “US Sources : Staged cyber attack reveals vulnerability in power grid,” Cable News Network, 26-Sep-2007. [Online]. Available: http://www.cnn.com/2007/US/09/26/power.at.risk/index.html. [Accessed: 30-Oct-2012].

[6] L. Greenemeier, “Heart-Stopper : Could Hackers Hit Pacemakers , Other Medical Implants?,” Scientific American, 14-Mar-2008.

[7] D. Halperin, T. S. Heydt-Benjamin, K. Fu, T. Kohno, and W. H. Maisel, “Security and Privacy for Implantable Medical Devices,” IEEE Pervasive Computing, vol. 7, no. 1, pp. 30–39, Jan. 2008.

[8] B. Grubb, “Fatal risk at heart of lax security,” The Sydney Morning Herald, Sydney, Australia, 06-Nov-2012.

[9] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage, “Experimental Security Analysis of a Modern Automobile,” in 2010 IEEE Symposium on Security and Privacy, 2010, pp. 447–462.

[10] S. Checkoway and D. McCoy, “Comprehensive experimental analyses of automotive attack surfaces,” in 20th USENIX Security Symposium, 2011.

[11] M. Crawford, “Utility hack led to security overhaul,” Computerworld, vol. 2006, pp. 1–2, 2006.

[12] M. Abrams and J. Weiss, “Malicious Control System Cyber Security Attack Case Study – Maroochy Water Services, Australia,” in NIST Industrial Process Control System Workshop, 2008.

[13] T. Smith, “Hacker jailed for revenge sewage attacks Job rejection caused a bit of a stink,” The Register, 31-Oct-2001.

[14] S. Hymon, “Engineers, Architects Strike Out on Picket Lines,” Los Angeles Times, Los Angeles, California, 11-Sep-2006.

[15] S. Bernstein and A. Blankstein, “Key signals targeted, officials say,” Los Angeles Times, Los Angeles, California, 09-Jan-2007.

[16] M. Krasnowski, “2 men accused of hacking into traffic system,” The San Diego Union-Tribune, San Diego, CA, 21-Jan-2007.

[17] S. Grad, “Engineers who hacked into L.A. traffic signal computer, jamming streets, sentenced,” Los Angeles Times, Los Angeles, California, 01-Dec-2009.

[18] J. Leyden, “Polish teen derails tram after hacking train network,” The Register, 11-Jan-2008.

[19] G. Baker, “Schoolboy hacks into city’s tram system,” The Telegraph, 11-Jan-2008.

[20] A. Matrosov, E. Rodionov, D. Harley, and J. Malcho, “Stuxnet under the microscope,” 2010.

[21] D. Albright, P. Brannan, and C. Walrond, “Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant ?,” Washington D.C., 2010.

[22] A. Bloom, “60 Minutes - Stuxnet: Computer worm opens new era of warfare,” CBS News, 2012.

[23] K. Coleman, “Arson by Cyber Attack,” Fire Engineering, 12-Dec-2012. [Online]. Available: http://www.fireengineering.com/articles/2012/12/arson-by-cyber attack.html. [Accessed: 18-Dec-2012].

[24] R. Moreno, “Cyber-Physical Systems Security for the Smart Grid,” in Cybersecurity in Cyber-Physical Systems Workshop, 2012.

[25] A. Weimerskirch, “Safety-Critical Automotive and Industrial Data Security (Extended Abstract),” in Cybersecurity in Cyber-Physical Systems Workshop, 2012.

[26] M. Ben Salem, “Security Challenges and Requirements for Control Systems in the Semiconductor Manufacturing Sector (Extended Abstract),” in Cybersecurity in Cyber-Physical Systems Workshop, 2012, pp. 1–3.

[27] S. Gupta, “Implantable Medical Devices - Cyber Risks and Mitigation Approaches,” in Cybersecurity in Cyber-Physical Systems Workshop, 2012.

Chapter 3.

Cyber Attack Threat Assessment and Impact Propagation

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

A Control Measure Framework to Limit Collateral Damage and Propagation of Cyber Weapons

David Raymond
United States Military Academy
West Point, New York, USA

Gregory Conti
United States Military Academy
West Point, New York, USA

Tom Cross
Lancope Inc.
Alpharetta, Georgia, USA

Robert Fanelli
United States Cyber Command
Fort Meade, Maryland, USA

Abstract: With the recognition of cyberspace as a warfighting domain by the U.S. Department of Defense, we anticipate increased use of malicious software as weapons during hostilities between nation-states. Such conflict could occur solely on computer networks, but increasingly will be used in conjunction with traditional kinetic attack, or even to eliminate the need for kinetic attack. In either context, precise targeting and effective limiting of collateral damage from cyber weaponry are desired goals of any nation seeking to comply with the law of war. Since at least the Morris Worm, malicious software found in the wild has frequently contained mechanisms to target effectively, limit propagation, allow self-destruction, and minimize consumption of host resources to prevent detection and damage. This paper surveys major variants of malicious software from 1982 to present and synthesizes the control measures they contain that might limit collateral damage in future cyber weapons. As part of this work, we provide a framework for critical analysis of such measures. Our results indicate that a compelling framework for critical analysis emerges by studying these measures allowing classification of new forms of malware and providing insight into future novel technical mechanisms for limiting collateral damage.

Keywords: *cyber operations, malware controls, collateral damage, law of armed conflict*

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, US Cyber Command, or the United States Government.

1. INTRODUCTION

As the world becomes more reliant on computers and networks, it is only natural that they will become targets during geopolitical conflict. Computer networks underlie our public and private utility infrastructures, banking and financial systems, and military command and control systems, all potentially lucrative targets. Targeting these networks requires an offensive cybersecurity capability and the addition of cyberspace components to the U.S. and other countries' military organizations highlights the potential for future conflict in the cyber domain.

Centuries of armed conflict have informed an ethical and legal framework for warfare, which includes a responsibility to limit collateral damage. Current Law of Armed Conflict (LOAC) addresses traditional armed conflict, but is not well-defined in the cyber domain. The complex interaction and highly interconnected nature of systems in cyber and physical space make the application of these laws more challenging. However, this ambiguity cannot be an excuse to act without regard to ethical considerations in cyberspace. Offensive cyber weapons created for use in interstate conflict could cause serious collateral damage to physical and informational assets. For example, malware designed to shut down industrial control systems in an adversary's munitions manufacturing facility might accidentally shut down a hospital's power control system. We must carefully follow the LOAC in the development of our cyber weapons if we are to justify their use to ourselves and to the international community.

This research examines ways cyber weapons can be controlled to limit collateral damage. Even the earliest malicious software included controls to limit infection and restrict spread to certain systems. The first well-known computer worm, the Morris Worm, could only infect DEC VAX computers running specific operating systems and checked whether a machine was already infected to limit resource consumption [1]. Notably, these checks failed to function properly and the worm degraded much of the ARPANET.

This paper provides a framework of controls that cyber weapons developers can use to more carefully control their software and avoid unwanted collateral effects. The value of our framework is that it demonstrates how malware can be controlled to severely reduce the threat of collateral damage, and it provides a template against which malware can be evaluated to determine how well it conforms to the LOAC. The framework does not consider third-party control of malicious software released by other individuals or organizations, such as the FBI's response to the DNSChanger malware [2]. Furthermore, the framework is of little use to malicious actors who create malware without regard for ethical considerations.

Section 2 of this work provides background and discusses related malware research. Section 3 provides a representative sampling of malware that has employed control mechanisms to limit its spread and Section 4 proposes a framework for controlling malware to allow specific targeting while limiting collateral damage. Section 5 presents analysis and our conclusions.

2. BACKGROUND AND RELATED WORK

A. BACKGROUND

In military terminology, *targeting* refers to the process of selecting appropriate capabilities to achieve a commander's desired effects. Capabilities can either be kinetic (bombs and bullets) or non-kinetic (leaflets and press releases). Non-kinetic capabilities are usually preferred because they minimize loss of life. For example, the Stuxnet virus seems to have been designed to sabotage centrifuges at Iran's Natanz nuclear enrichment facility [3]. A bombing campaign might have achieved the same purpose, but with potentially high casualties.

Military leaders have recently recognized the potential for cyber weapons to produce effects that meet the commander's intent, either in conjunction with or in lieu of kinetic operations. Furthermore, cyber weapons are difficult to attribute to a specific individual, organization, or nation-state, carry minimal risk to friendly and enemy forces, and limit collateral damage in the traditional sense.

The cyber attacks launched against Georgia during hostilities with Russia in August 2008 exemplify the potential of this new form of warfare. The Russian invasion of Georgia was preceded by cyber attacks consisting of website defacements and distributed denial-of-service (DDoS) attacks targeting government, news media, and financial websites [4]. These attacks limited the Georgian government's ability to coordinate a response to the Russians and prevented Georgia from getting their story to rest of the world. Whether these cyber attacks were coordinated by the Russian government or not, they were of benefit to Russia's subsequent invasion [5].

As cyber weapons become more attractive as a component of military action, legal and ethical questions arise with regard to the LOAC and its application to cyberspace. In this work, we are concerned primarily with *jus in bello*, or the ethics of conduct during warfare, and specifically the ethics concerning the use of cyber weapons and the potential for collateral damage [6]. The principle of *distinction* requires that non-combatants be avoided in attacks because they are not legitimate military targets. According to this tenet, military leaders should also avoid collateral effects on non-combatants. The principle of *proportionality*

dictates that the defense against an aggressor must be proportional to the attack. While completely avoiding collateral damage is not always possible, proportionality dictates that collateral effects be minimized [6].

Several countries have recently increased their capabilities to conduct cyber operations. U.S. Cyber Command was established in May 2010 and is "responsible for planning, coordinating, integrating, synchronizing, and directing activities to operate and defend the Department of Defense information networks and when directed, conducts full-spectrum military cyberspace operations." China began forming a cyber force as early as 1997, and in July 2010 announced the establishment of an 'Information Protection Base' within the People's Liberation Army (PLA) to defend their networks [7]. Russia and Iran have well-defined military objectives in cyberspace [8] [9]. These are just a few examples of world leaders formalizing their cyber security efforts and placing them at least partially under the control of their militaries. As cyber operations increasingly become the purview of military leaders and are used as a component of military operations, it is important that we define the boundaries of moral-ethical behavior for the deployment of cyber weapons. Some countries will choose not to employ control measures in their cyber weapons. Those countries that choose not to follow established standards of behavior with their cyber weapons should be treated by the international community like countries that ignore the LOAC in other areas.

B. RELATED MALWARE RESEARCH

Cohen conducted extensive virus experiments starting in the 1980s, first coding them and then developing virus defenses. One of his earliest papers provided pseudocode for a generic virus that included one of the basic malware controls, checking to see whether a file was already infected before modifying it [10]. Cohen studied the potential for identifying malware on a system and proved that no single algorithm can positively detect all computer viruses. He also made a case for the benevolent use of computer viruses [10].

Some of the earliest published research on computer malware is in Ludwig's *Little Black Book of Computer Viruses* [11] and his follow-on, *Giant Black Book of Computer Viruses* [12]. These seminal books describe the development of self-replicating malware and discuss methods for hiding malicious code and avoiding antivirus software. Ludwig even envisioned the potential for military applications of malicious software back in 1990 [11], an idea that has only recently been acknowledged by government and military leaders.

Research by Fanelli explored a methodology for targeting and controlling collateral damage in cyber operations [13]. He argued that the LOAC mandates that countries

seek to avoid collateral damage in cyber operations and shows that, despite the complex nature of these operations, it is possible to affect specific targets while minimizing effects on non-target systems and organizations.

Importantly, much of the most significant related research comes from academic and industry analyses of each malware family and variant. We include key references later in the paper.

3. MALWARE CONTROL EXAMPLES

Some consider malware such as viruses and worms to be uncontrolled once released, however, from the earliest examples of malicious software, controls were used to limit propagation and restrict behavior. Recent malware examples use sophisticated controls that even seem to specifically target organizations or facilities. The most basic control measure, observed in 1987 with the discovery of the Stoned virus, is for malware to check to see whether the target system is already infected [14]. Stoned alters the master boot record (MBR) on floppy disks and moves the original MBR to another location on the disk. Once resident in memory, Stoned checks whether disks inserted into the computer are already infected and, if so, does not alter them. For Stoned, this pre-infection check prevents the original MBR from being overwritten. In other malware it might be done to prevent unnecessary resource consumption or for other reasons. Control measures have grown in sophistication and the evolution of controls is summarized in Table I.

One of the first large-scale cases of malware infection was the Morris Worm, released in November 1988 by Robert Tappan Morris [1]. Morris took advantage of vulnerabilities in the *fingerd* and *sendmail* daemons in some versions of Berkeley Software Distribution (BSD) UNIX. The worm was written to affect Sun Microsystems Sun-3 systems and VAX systems running 4 BSD, however, it did not affect systems running the Sun-4 operating system (OS) even though Morris pointed out the flaw in its *fingerd* daemon to staff at Carnegie-Mellon University a year before the worm was released [1]. This oversight may have been designed to draw attention away from the worm's author, but shows that malware can be written to exploit not only a specific OS, but particular versions of that OS.

The Morris Worm checked to see whether a target was already infected and if so, would not re-infect it, thus limiting propagation and reducing resource consumption on affected systems. The worm was programmed to probabilistically skip this check one in seven infections to make it harder to eradicate [1]. Morris' lack of understanding of the potential propagation rate and incomplete testing caused the worm to replicate much faster than anticipated.

Table I. Mapping of Control Mechanisms to Representative Malware since 1982.

	Check previous infection	Check OS and version	Check application software	Self propagation	Check date/time	Propagation counters	Checks language settings	Check for reverse engineering	Contact C2	Other controls
Elk Cloner (1982) [29]		Apple II		Infects boot sector						Payload executes every 50th boot
Stoned (1987) [14]	Check master boot record			Infects floppy						
Morris Worm (1988) [1]	Attempt port connect	Sun-3 & Vax BSD		using sendmail & fingerd						
Jerusalem (1989) [15]	Checks files for infection			Infects boot sector	Payload executes on Friday 13th					
Frodo (1989) [28]				Infects boot sector	Propagates on Sept 22					
Michelangelo (1991) [16]				Infects boot sector	March 6th					
Concept (1995) [27]	Checks for macro		MS Word 6	Infects Word docs		Only sends to 50 contacts				
Melissa (1999) [26]	Registry key		MS Word 97/2000	Sends itself to addr book entries						
Code Red (2001) [25]		Windows NT/2000		IIS exploit in HTTP GET	Behavior based on day of month		English Windows			Will not infect if file c:\ notworm is present
Code Red II (2001) [24]	Atom	Windows NT/2000		IIS exploit in HTTP GET	Termination on Oct. 1, 2001		Chinese Windows			Aware of IP address to avoid reinfection
Blaster (2002) [23]	Mutex	Windows2000/XP		DCOM RPC(135) and tftp(4444)	Executes DoS on certain dates					
Welchia (2003) [22]	Mutex	Windows 2000/XP	IIS and msblast.exe	DCOM RPC(135) and http (80)	Terminates on Jan. 1, 2004					Attempts to disinfect Blaster and patch host
MyDoom (2004) [21]	Mutex			Email or file sharing (Kazaa)	DDos on 2/1/04; stop spread 2/12					Avoids propagation to certain domains (.gov/.mil)
Storm Worm (2007) [20]	Registry key		Shuts down A/V					Checks for VM or dbg	Overnet P2P	Geolocation via network address
Conficker (2008) [19]	Mutex	Windows variants		Network and USB	Contact C2 starting on 26 Nov 2008		Ukranian settings	Check for VM	HTTP and P2P	
IXESHE (2009) [18]	Registry key								HTTP	Primary vector is targeted email
Stuxnet (2010) [3]	Mutex Registry key	Windows variants		Network and USB	Several date checks	Delete USB at 3 infections			P2P/RPC	Checks for specific attached hardware
Flame (2012) [17]	Mutex	Windows variants		USB & Win update					HTTPS	

Another control is to deliver a payload on a specific date. The Jerusalem virus, discovered in 1989, triggered on any Friday the 13th, and the Michelangelo virus (1991), deleted important data on March 6th [15] [16]. Targeting an organization on a specific date might help to coordinate a large scale cyber attack or to coordinate cyber-based effects with a kinetic operation.

The Concept virus (1995) was the first known to take advantage of macros in the Microsoft Office suite [27], infecting computers with Microsoft Word installed. In 1999, the much more sophisticated Melissa virus took advantage of the Microsoft Office macro framework, using Word macros to replicate via emails using Microsoft Outlook. Melissa had other controls, running only once per session to limit propagation and displaying a special message if the minute of the hour matched the day of the month at the time of infection [26]. Melissa limited propagation by sending itself to only 50 entries on the victims' Microsoft Outlook address book. Other email worms, such as MyDoom (2004) limit email propagation by avoiding certain domains [21]. One could imagine extending this functionality to limit propagation to address book entries with specific email domains, telephone prefixes, surnames, or mailing addresses.

A recent trend in malware is to terminate to prevent analysis when running in a virtual machine or debugger. An example is the Storm Worm mentioned above [20]. Another is one of the most ubiquitous worms ever deployed, Conficker, identified in November 2008 [19]. Conficker has used a variety of control mechanisms over several revisions. It self-replicates, trying to connect to other computers on a local network by exploiting a Windows service vulnerability [19]. Later variants replicated via removable media and using a peer-to-peer mechanism. Conficker checks the OS version to determine which exploit to trigger, checks network connectivity, and attempts to subvert firewalls. Version A would not infect systems whose keyboard language layout was set to Ukrainian or that had a Ukrainian IP address. Starting with version B, Conficker attempted to shut down antivirus products on the target. After infection, Conficker checks the date and beginning on 26 Nov 2008, attempted connections to command and control (C2) servers to download more code. It also encrypted its payload and employed anti-debugging logic to self-destruct if it sensed an attempt at forensic analysis [19].

From our analysis, the rise of the Advanced Persistent Threat (APT) in the late 2000s has seen more carefully targeted infection attempts, often in the form of direct emails that contain links to web sites or files that take advantage of application vulnerabilities to plant malicious code on the recipient's computer. After initial infection, many of these tools contact a C2 server for additional instructions or to download new modules. Removing self-propagation allows attackers to target an individual or organization with more precision. It is less reliable, however, because targeted emails can draw suspicion and they typically require user action, such

as opening a file or clicking on a link. An example of this type of threat is the IXESHE (2009) APT campaign [18].

One of the most tightly-controlled pieces of malware ever discovered is Stuxnet [3]. The consensus among several security firms, including Symantec, Kaspersky Labs, and others, is that Stuxnet was designed to cause subtle failures in industrial equipment. Before installing itself, Stuxnet ensures a certain system configuration is present. It first checks the operating system and version, choosing to only target specific Windows systems. Stuxnet then checks a registry key to determine whether the host is already infected and checks the system date, exiting if the date is after 24 June 2012.

Once installed, Stuxnet only affects specific types of Programmable Logic Controllers (PLCs) supervised by the Siemens Company's Step 7 software and connected to frequency converter devices manufactured by the Fararo Paya company in Iran or the Vacon company in Finland. Specifically, Stuxnet would only infect S7-315 PLCs attached to arrays of 33 or more frequency converter devices or S7-417 PLCs attached to 6 groups of 164 frequency converter drives.

Many suspect that Stuxnet was designed to target Iranian centrifuges at their Natanz uranium enrichment plant [3]. It did, however, propagate beyond Natanz, both through infected machines that left the network and joined another, and via USB flash drives. Analysis of the code indicates that Stuxnet should delete itself from infected USB drives after three infections and that it should have deleted itself after 21 days, however these controls were non-functional [30]. Stuxnet was discovered in 2010 after this error allowed it to infect systems in several countries [31]. While additional machines were infected and there was a cost associated with eradication, Stuxnet was inert on devices that did not meet the above configurations. Collateral damage therefore was minimized by the specific controls included in Stuxnet.

In May 2012, researchers at Kaspersky Lab identified a new piece of malware, dubbed Flame, whose primary propagation mechanism is infected USB drives. Flame is also the first known instance of malware to subvert Windows Update [17]. Infected machines can masquerade as Windows Proxy Auto-Discovery (WPAD) servers and hijack requests for Windows Updates within their local network to provide malicious patches. The authors of Flame used forged certificates that allowed them to make their illegitimate Windows updates appear to be signed by Microsoft. Infected hosts contact a C2 server for modules and instructions. C2 servers can send a kill module that causes the malware to be wiped from the system.

4. FRAMEWORK FOR MALWARE CONTROLS

Our malware control framework builds upon the cyberspace planes suggested by Fanelli (see Figure 1) [13]. The geographic plane includes the physical location of the target and includes the implications imposed by geographic boundaries, as well as physical aspects of the location of a specific target system such as power infrastructure and building location. The physical plane includes a device's physical hardware and protocols that allow for communication. This plane encompasses the physical layer (layer 1) of the Open Systems Interconnection (OSI) model, in addition to other features of a device's hardware such as serial numbers and types of attached peripheral devices. We subdivide the logical plane into the top six layers (layers 2 - 7) of the OSI model, which provides logical abstraction layers for communications systems (detailed in Table II). The cyber persona plane resides above the logical plane and includes individual virtual identities in the cyber domain. Finally, the supervisory plane includes persons and systems that provide the command and control necessary to start, stop, or redirect cyber weapons. Our framework for malware controls, discussed below, maps the cyberspace planes to cyber weapon control measures.

The following paragraphs differentiate between active and passive cyber weapon control measures, then map different types of control measures to the cyberspace planes in Figure 1.

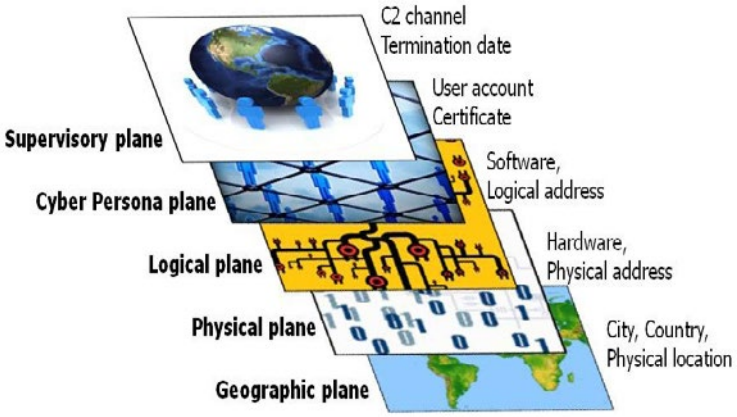

Figure 1. Cyberspace planes. The four planes described in [13] are expanded to include the Geographic plane and the OSI model

A. CONTROL MEASURE CLASSIFICATIONS

We classify malware control mechanisms as active, passive, or hybrid. Passive (or autonomous) control has a cyber weapon observing its environment and acting on those observations, based solely on internal logic resident in its code. Observations can include a variety of system characteristics like those mentioned in Section 3, such as inspecting the current date, checking for existing copies of itself, examining registry keys, reviewing installed operating system and application software, or checking for attached hardware.

Active control measures allow an external decision-maker to decide what actions to take either by directly issuing commands or via code updates. Examples include malware that contacts a C2 server and receives instructions or a virus that opens a port and sends notification that the machine is ready to accept network connections.

Hybrid control is a combination of active and passive control measures. An example is malware that checks to see if a system meets certain configuration specifications and based on those observations, decides whether to contact a C2 node.

B. CONTROL MEASURES BY CYBER PLANE

To maximize the reliability of malware targeting, the possible control measures at each cyber plane and OSI layer must be considered. These controls are summarized in Table II and are examined in the following paragraphs.

1) Geographic Plane. Here we consider control measures that require an agent deploying malware to be in the same geographic area as the targeted system. Someone might drop a USB "thumb" drive in a parking lot, hoping an employee will connect it to a computer that has network access to a target system. Someone might even be able to gain physical access to a target device to connect removable media or bypass login procedures to install software. These are all considered active control measures.

2) Physical Plane/Physical Layer. The physical plane of cyberspace maps almost directly to the physical layer of the OSI model. Different from the geographic plane, the physical plane includes components of a computer system and other hardware attached to it. Control measures at this layer are primarily passive ones, like checking serial numbers of peripheral devices or determining which physical layer protocols are being used. If reconnaissance of a target is sufficiently detailed, several such controls could be incorporated at the physical layer to provide very specific targeting, such as Stuxnet's checking for specific types of attached PLCs [30].

Table II. Malware C2 Methods Mapped to the Five Cyberspace Planes from Figure 1.

Cyber Plane/OSI Layer		Command and Control Mechanisms
Supervisory plane		(Active) Control malware via active C2 architecture (Active) Human decision maker (Hybrid) Develop targeting code update and push to malware on system
Cyber persona plane		(Passive) Check for specific identity – user ID, email address, social network identity, etc. (Active) Collect and report identify information to a controller
Logical Plane	**7. Application layer**	(Passive) Check OS or application software and versions (Passive) Check hostname, domain of target systems (Passive) Check for presence (or absence) of VM host (Passive) Check for evidence of debugger (Passive) Check local date and/or time (Active) Propagation counter - limit automatic propagation to fixed number
	6. Presentation layer	(Passive) Check for specific encryption or encoding techniques used to translate data between network and application formats (Passive) Check language/character set translations
	5. Session layer	(Passive) Check application layer protocol fields (e.g. fields in ICCP or ELCOM protocol messages can identify specific SCADA systems)
	4. Transport layer	(Passive) Check for specific transport-layer protocols used by target (e.g. COTP or TPTK protocols can indicate SCADA systems)
	3. Network layer	(Passive) Check for network address of target area or organization
	2. Data link layer	(Passive) Check link-layer protocol used in network (Passive) Check medium access control (MAC) address or organizationally unique identifier (OUI)
Phy Plane	**1. Physical layer**	(Passive) Check physically connected devices or device serial numbers (Passive) Check for RS-485 physical layer, used by many SCADA control systems (Active) Restrict propagation to specific removable media
Geographic Plane		(Active) Insider/physical access (Active) Drop thumb drive in parking lot

3) Logical Plane. This plane consists of the operating system, application software and software settings on a device. We further subdivide this layer into the upper six layers of the OSI model.

a) Data Link Layer. Here we are concerned with the data link layer protocols and addresses. Reconnaissance may tell us that an organization uses network adapters from a certain manufacturer. Malware could be programmed take action only on network interfaces with certain Organizationally Unique Identifiers (OUI). Data link layer protocols that might be examined include Ethernet, WiFi, Bluetooth, ZigBee, or others.

b) Network Layer. At the network layer, network addresses could be compared to a target entity's assigned network address space. While Network Address Translation (NAT) limits the reliability of inspecting network addresses, a thorough examination of a device's network environment, including routers and gateway addresses, might allow a better picture of the physical location of a device to be developed.

c) *Transport Layer.* While the majority of Internet connected systems use Transport Control Protocol (TCP) or User Datagram Protocol (UDP), many supervisory control and data acquisition (SCADA) systems and industrial control systems (ICS) use transport protocols tailored to those systems. Examples include the Connection Oriented Transport Protocol (COTP) and TPTK, which are sometimes used in place of, or in conjunction with, TCP [32]. Differentiating between SCADA or ICS systems and other networked devices might be important in a cyber campaign when a country's electrical infrastructure or power generation capability are to be targeted.

d) *Session Layer.* This layer, and the following presentation layer, are treated as part of the application layer in some network models. Here we list these layers separately as they have specific purposes with which cyber weapon control measures might be associated. The session layer provides the ability to establish a semi-permanent connection between two end points. At this layer, the Inter-Control Center Protocol (ICCP) and IEC 60870-6 (ELCOM) protocols are used for communication between utility control centers in SCADA and emergency management systems (EMS). Again, the ability to positively identify specific SCADA systems might be advantageous during cyber weapon employment.

e) *Presentation Layer.* This layer is used to convert data encoding or encryption formats used for network transfer to and from formats that can be used by the application layer. Passive control measures at this layer might include checking for specific data encoding or encryption techniques known to be used by the target entity.

f) *Application Layer.* There are a variety of passive checks that can be made at the application layer. Cyber weapons might be programmed to check for specific operating system software or versions, certain application software or versions, hostname, username, domain name, environment data such as date, time, or location settings, or the presence or absence of a virtual machine environment.

4*) Cyber Persona Plane.* As defined by Fanelli [13], this plane identifies identities in the cyber domain, which might have many-to-one or one-to-many, or many-to-many relationships with individuals in the physical world. The presence or absence of specific identities may be used to validate targets or limit application of effects.

Controls on the cyber persona plane consider indications of use or ownership by a specific person, group, corporation or government. Examples include account credentials, certificates, cookies, licensed software, biometric data, and observations of network activity such as logging into accounts correlated with a persona

5) *Supervisory Plane*. At the top of the hierarchy is the supervisory plane, which provides oversight and the authority to start, stop, modify, or redirect a cyber weapon or cyber campaign, within the limits of the weapon's capabilities and C2 infrastructure. At this level, operational decisions are made about the prosecution of a cyber campaign.

5. ANALYSIS AND CONCLUSIONS

One approach to analyzing malware controls is to specify undesired effects that cyber weapons might create, actions that can be taken to mitigate those effects, and corresponding controls that can provide mitigation. Table III provides one such analysis.

Based on our framework and the large variety of controls that can be used at varying levels of specificity and effectiveness, we believe that cyber weapons can be very carefully crafted and targeted to affect only specific systems and organizations, greatly reducing undesired collateral effects. As with kinetic attacks, more detailed intelligence allows for better targeting and weapon development. Decision-makers must weigh the value of a target against the potential for collateral damage and may have to assume risk. The difficulty in attributing a cyber attack to a specific entity might reduce the risk of being held accountable for collateral damage, but it does not alleviate the moral responsibility to limit it. Furthermore, very fine-grained controls used to ensure that cyber weapons will only affect specific targets might provide clues to the origin of those weapons.

Another risk that cyber weapon authors must consider is the potential for controls included in their software to identify their intentions. Had Stuxnet been analyzed before centrifuges were damaged, Iran might have suspected that those centrifuges were the target, causing them to tighten defenses. One novel approach to prevent such analysis is to encrypt the malware payload and use data gathered from the infected system, such as registry entries, portions of the physical or network address, or device serial numbers, to generate a decryption key. This technique is used in the Gauss malware (2012), which gathers information about the system path and installed software, then calculates an MD5 hash and attempts to use it as a key to decrypt the payload [33]. As of this writing, security researchers have not been able to decrypt and analyze the Gauss payload.

Table III. Malware Collateral Effects, Mitigating Actions, and Representative Controls

Undesired collateral effects	Mitigating actions	Representative Controls
Unintended infection	Limit propagation to specific targets	Disallow self-replication Infect systems only via spear-phishing with malicious attachment or link to download or through previously infected systems
Unintended payload execution causing loss of: Confidentiality (data exposure) Availability (loss of data, denial of service, consumption of network resources) Integrity (data modification)	Prevent payload execution on non-target systems	Use only active control measures to activate payload Use detailed reconnaissance to determine triggers for passive or hybrid control Trigger malware based on known target configuration
Vulnerability disclosure to unintended individuals or general public	Prevent reverse engineering and subvert forensic investigation	Encryption Tamper protection Temporary payloads that delete themselves from memory
Attribution of attack or source of the malware	Eliminate evidence of authors	Encryption Tamper protection Use widely used languages, libraries, and coding techniques Temporary payloads

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. Just as kinetic weapons should not be used without sufficient intelligence regarding the target, cyber weapons should not be used unless intelligence is available to adequately limit potential damage to non-target systems.

As nations increasingly recognize the potential for cyber weapons as tools of warfare, it is important to find ways to ensure that they are used responsibly in a way that conforms with the LOAC and minimizes unwanted collateral effects. Since the introduction of malicious software, techniques have been used to control it, either actively or passively, to target specific systems or otherwise shape its effects. In this work we have established the potential to better control the behavior of cyber weapons and summarized previously used techniques. We go on to develop a framework for malware controls, mapping them using our cyber planes model and categories of propagation techniques. This framework can be used to incorporate effective controls during the development of cyber weapons. Of particular value is the ability to analyze malware in the context of this framework to determine whether it conforms to internationally recognized standards of ethical behavior during design and planning, while in use, and during post-use analysis by the aggressor, the target entity, or third-parties seeking to verify appropriate behavior.

REFERENCES

[1] E. Spafford, «The Internet Worm Program: An Analysis,» Purdue University Technical Report CSD-TR-823, 1988.

[2] Federal Bureau of Investigations, «DNSChanger Malware,» September 2012. [Online]. Available: http://www.fbi.gov. [Accessed 4 September 2012].

[3] Falliere, L. Murchu and E. Chien, «W32.Stuxnet Dossier, v1.4,» Feb 2011. [Online]. Available: http://www.symantec.com. [Accessed September 2012].

[4] E. Tikk, K. Kaska, K. Runnimeri, M. Kert, A. Taliharm and L. Vihul, «Cyber Attacks Against Georgia: Legal Lessons Identified,» Tallin, Estonia, November 2008.

[5] M. Johnson, «Georgian Websites Under Attack - Don't Believe the Hype,» August 2008. [Online]. Available: www.shadowserver.org. [Accessed 4 September 2012].

[6] M. Walzer, *Just and Unjust Wars*, New York: Basic Books, 1977.

[7] D. Ball, «China's Cyber Warfare Capabilities,» *Security Challenges,* vol. 7, no. 2, pp. 81 - 103, Winter 2011.

[8] D. J. Smith, «How Russia Harnesses Cyber War,» *Defense Dossier,* vol. 4, pp. 7 - 11, August 2012.

[9] I. Berman, «Cyberwar and Iranian Strategy,» *Defense Dossier,* vol. 4, pp. 12 - 15, August 2012.

[10] F. Cohen, «Computer Viruses, Theory and Experiments,» *Computers and Security,* vol. 6, no. 1, 1987.

[11] M. Ludwig, *The Little Black Book of Computer Viruses*, Show Low, AZ: American Eagle Publishing, 1990.

[12] M. Ludwig, *The Big Black Book of Computer Viruses*, Show Low, AZ: American Eagle Publishing, 1995.

[13] R. Fanelli and G. Conti, «A Methodology for Cyber Operations Targeting and Control of Collateral Damage in the Context of Lawful Armed Conflict,» in *International Conference on Cyber Conflict (CyCon)*, Tallinn, Estonia, June 2012.

[14] «Virus:Boot/Stoned,» F-Secure, [Online]. Available: http://www.f-secure.com. [Accessed 1 October 2012].

[15] «Jerusalem,» F-Secure, [Online]. Available: http://www.f-secure.com. [Accessed 30 September 2012].

[16] «Michelangelo (computer virus),» [Online]. Available: http://en.wikipedia.org. [Accessed 12 October 2012].

[17] M. Lee, «Flame Used Windows Update to Spread,» 5 June 2012. [Online]. Available: http://www.zdnet.com. [Accessed 4 August 2012].

[18] D. Sancho, J. D. Torre, M. Bakuei, N. Villeneuve and R. McArdle, «IXESHE: An APT Campaign,» Trend Micro Research Paper, September 2012.

[19] N. Fitzgibbon and M. Wood, «Conficker.C: A Technical Analysis,» 2009.

[20] X. Chen, J. Andersen, Z. M. Mao, M. Bailey and J. Nazario, «Towards an Understanding of Anti-virtualization and Anti-debugging Behavior in Modern Malware,» in *IEEE International Conference on Dependable Systems and Networks (DNS)*, Ann Arbor, MI, June 2008.

[21] «Worm:W32/Mydoom,» [Online]. Available: http://www.f-secure.com. [Accessed 18 December 2012].

[22] «Worm:W32/Welchi,» F-Secure, [Online]. Available: http://f-secure.com/. [Accessed 30 December 2012].

[23] «Virus alert about the Blaster worm and its variants,» Microsoft, [Online]. Available: http://support.microsoft.com. [Accessed 18 December 2012].

[24] D. Moore, C. Shannon and J. Brown, «Code-Red: a case study on the spread and victims of an Internet Worm,» in *2nd ACM SIGCOMM Workshop on Internet Measurement*, Marseille, France, November 2002.

[25] J. C. Dolak, «The Code Red Worm,» SANS Institute Reading Room, 2001.

[26] Carnegie-Mellon University Software Engineering Institute, «CERT Advisory CA-1999-04 Melissa Macro Virus,» March 1999. [Online]. Available: http://www.cert.org. [Accessed 12 August 2012].

[27] «Virus:W32/Concept,» F-Secure, [Online]. Available: http://f-secure.com. [Accessed 12 October 2012].

[28] «Frodo,» [Online]. Available: http://virus.wikia.com. [Accessed 14 December 2012].

[29] «Elk Cloner,» [Online]. Available: http://virus.wikia.com. [Accessed 12 December 2012].

[30] B. Schneier, «Stuxnet,» 7 Oct 2007. [Online]. Available: http://www.schneier.com. [Accessed 19 December 2012].

[31] D. Sanger, «Obama Order Sped Up Wave of Cyberattacks Against Iran,» 1 July 2012. [Online]. Available: http://www.nytimes.com. [Accessed 25 July 2012].

[32] IETF, «RFC 1006 - ISO Transport Service on top of the TCP,» May 1987. [Online]. Available: http://tools.ietf.org/. [Accessed 1 December 2012].

[33] Kaspersky Labs, «The Mystery of the Encrypted Gauss Payload,» 14 August 2012. [Online]. Available: http://www.securelist.com. [Accessed 20 Dec. 2012].

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

A Baseline Study of Potentially Malicious Activity Across Five Network Telescopes

Barry Irwin
Security and Networks Research Group, Department of Computer Science
Rhodes University
Grahamstown, South Africa
b.irwin@ru.ac.za

Abstract: This paper explores the Internet Background Radiation (IBR) observed across five distinct network telescopes over a 15 month period. These network telescopes consisting of a /24 netblock each and are deployed in IP space administered by TENET, the tertiary education network in South Africa covering three numerically distant /8 network blocks. The differences and similarities in the observed network traffic are explored. Two anecdotal case studies are presented relating to the MS08-067 and MS12-020 vulnerabilities in the Microsoft Windows platforms. The first of these is related to the Conficker worm outbreak in 2008, and traffic targeting 445/tcp remains one of the top constituents of IBR as observed on the telescopes. The case of MS12-020 is of interest, as a long period of scanning activity targeting 3389/tcp, used by the Microsoft RDP service, was observed, with a significant drop on activity relating to the release of the security advisory and patch. Other areas of interest are highlighted, particularly where correlation in scanning activity was observed across the sensors. The paper concludes with some discussion on the application of network telescopes as part of a cyber-defence solution.

Keywords: *network telescope, darknet, internet radiations, scanning*

1. INTRODUCTION

This paper explores the Internet Background Radiation (IBR) [1], [2], [3] observed across five distinct network telescopes over a fifteen month period. These five network sensors each monitored a block of 256 IP version 4 (IPv4) addresses, with a combined size of 1 280 addresses. No live services or hosts existed in this address space, and as such one would expect relatively little traffic to have been observed. In contrast nearly 100 million errant, unsolicited datagrams were observed across the sensors, recorded from over 14 million sources. Of particular interest is the degree in the similarity of traffic observed across portions of the observed traffic, despite the monitored address blocks being numerically distant in terms of the IPv4 addressing scheme.

An advantage using smaller blocks is that one can attain a wider view of what trends are occurring with IBR, than one would with the same address space in a contiguous block. Greynets [4], [5] are a related implementation using smaller slices of address space than have traditionally been used for the operation of network telescopes, and may be of increasing value in the future. The case studies presented serve to illustrate some of the value in running distributed network sensors, as traffic can be correlated for an extended period, and responses to events such as security advisories observed in the collected data.

While a detailed analysis of all aspects of this observed traffic is beyond the scope of this paper, several interesting observations are presented, and analysed. Conventional wisdom relating to the sizing of network telescopes [1], [2], [6] has agreed that a large address space, such as that utilised by CAIDA[1] is needed in order to obtain meaningful data but, as shown in the following sections, viable results have been obtained using a significantly smaller aggregate sensor size than the /8 used by CAIDA, or /16 typically used by other researchers [3]. This work is also novel in terms of the correlation of activity and observed hosts across different network telescopes over a fairly lengthy period.

A. STRUCTURE

The remainder of the paper is structured as follows. Section 2 provides a brief introduction to the use and history of network telescopes. The data sets used in this paper are disclosed in Section 3 along with the collection methods used. A high level analysis of observed traffic is explored in Section 4. Two case studies are presented in Section 5, considering the application of a network telescope toward

[1] Cooperative Association for Internet Data Analysis http://www.caida.org/

the monitoring and identification of network threats. These are presented with a focus on the similarities and differences in traffic targeting these ports across the different monitored network address blocks. Section 6 concludes this paper, providing a discussion on the findings presented, their potential application, and future research relating to this area of study.

2. NETWORK TELESCOPES

Network telescopes are a class of network security sensors, which have been used by security researchers in recent years. The basis of a network telescope is to monitor portions of unused IP address space [7], [8]. Specifically a network telescope makes use of unallocated IP addresses which are not being used for running services. Based on this premise, any incoming traffic observed as destined for the monitored IP address range can be viewed as unsolicited, as no clients or servers are operating using these addresses. This allows researchers to focus on the traffic commonly termed Internet Background Radiation (IBR) [1], [8] without having to worry about distinguishing it from potentially legitimate traffic to servers or client systems. From a research perspective, no legitimate traffic should be arriving at the sensor, which can ease privacy concerns relating to traffic capture.

Care is taken to filter traffic so as to ensure that no response traffic is sent so as to appear to remote hosts as indistinguishable from an unallocated address. A more detailed discussion of the varying modes of operation for network telescopes and related analysis methods can be found in [9].

What is important to bear in mind when analysing the data collected using the passive means of collection such as that used in this research, is that one of the shortcomings of this type of network telescope setup is that only the first packet of the potential TCP 3-way handshake is actually captured. Since the handshake, by design, cannot complete due to filtering of any return traffic, no data payload can be captured. Due to this limitation it can only be inferred, albeit with a high level of certainty that traffic targeting a given port is related to particular protocols or malware such as the Conficker [10], [11] and Morto [12] worms considered in the case studies.

In essence a network telescope is a passive sensor system that collects incoming traffic or ‘radiation’ from the Internet. This radiation is constituted from multiple source systems and traffic types. The analysis of this collected data can provide useful insight into the operation of the Internet, or even particular events such as worms or distributed denial of service (DDoS) attacks. Over the last few years researchers have focussed on using telescopes for DDoS analysis as discussed

in Moore et al. [7], [13]. Data collected has been successfully utilised for worm analysis, particularly that of Code Red [14] (the first worm observed on a Network Telescope) [15], [16]; Witty [17], [18], [19]; SQL Slammer [20] as well as more generically [21], [22]. In analysis reported by Kumar et al. [19], the researchers were able to perform detailed analysis of the Witty worm based on the traffic observed, to the extent of evaluating the number of physical drives present in infected systems and the probable identification of 'patient zero'. This was achieved through the analysis of data collected by a network telescope.

While all traffic received at the network telescope monitoring node can be seen to be unsolicited, the collected backscatter can be further classified under a number of categories. Strictly speaking backscatter can be regarded as traffic that is passive, and as such distinct from the active traffic recorded on the sensor. The term is, however, often misused in the sense of referring to all traffic that is not directly associated with communications of hosts on a network. This section builds on the view that traffic can be divided into the two broad classes of active and passive. Further discrimination is performed within these categories. Passive traffic can be defined as traffic from which no legitimate response can be expected from a system's TCP/IP networking stack when received [9]. As such it is unlikely that a potential attacker or instance of malware will be able to determine anything about the target system. The backscatter traffic observed can be seen to be the result of activities which result in the reflection of traffic from the originating machines to the telescope sensor. This requires that the source address of datagrams be spoofed to be within the IP address range monitored by a telescope sensor. Active traffic is defined as traffic which is expected to elicit a response of some kind when processed by a target system's TCP/IP stack. This differentiation is only easily possible for TCP and ICMP traffic where a clear structure is present as to what are a 'request' and a corresponding 'response'. UDP traffic is near impossible to classify as active and passive without doing payload analysis. To this end some sensors such as CAIDA's backscatter datasets filter it out [15].

3. DATA SOURCES

The data collected for this research was sampled from five network telescopes over a continuous period of 464 days from February 10th 2011 to May 20th 2012. The telescope's sensors each consisted of a /24 netblock (comprising 256 individual addresses) routed to a collection server. Packets were logged to disk on the sensor systems using libpcap with appropriate filters, to only log traffic destined to a specific netblock. The collection system host firewalls were configured to prevent any response to incoming traffic being generated, so as to avoid potentially disclosing their presence. The collector systems were located at two points on the TENET

network. A sample setup of a collection system is shown in Figure 1Capture files were subsequently collated, and processed on a separate system using an analysis platform as described in [9].

The blocks of IPv4 address space being monitored were distributed across three distinct top-level IP version 4 network address blocks (netblocks) – 146/8, 155/8 and 196/8. These networks are all contained within the TENET[2] (AS2018) network. Three /24 blocks of addresses were contained in 196/8 block, in two separate /16 netblocks. Datasets are referred to by the /8 netblock in which they reside. A summary of the data sets collected is shown in Table I, along with the naming of the datasets. These have been sampled from the much larger datasets collected from these sensors, some of which have been running since 2005. Combined, the datasets used in this study comprise 99 007 576 packets.

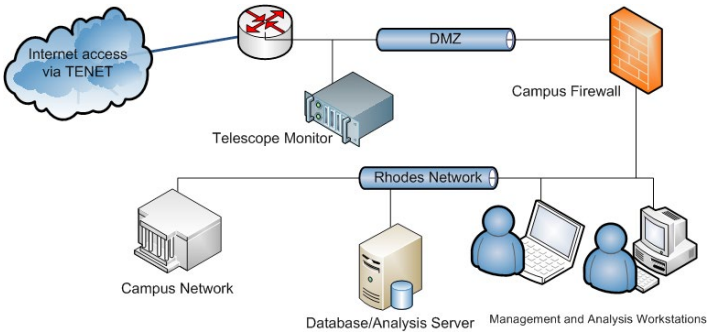

Figure 1. Sample Network Telescope Setup

Table I. Dataset overview

Dataset Name	Total packets	Protocol %			Sources /32
		TCP	UDP	ICMP	
146	4 768 524	62.11	25.26	12.62	467 419
155	7 547 605	77.18	14.45	8.36	498 134
196-1	20 071 795	91.42	6.25	2.30	3 304 445
196-2	32 479 388	92.87	4.87	2.25	5 036 684
196-3	34 140 264	90.63	7.43	1.92	5 103 316

[2] TENET is the Tertiary Education Network in South Africa http://www.tenet.ac.za/

The sensor collecting traffic for dataset 196-1 experienced a significant 171 day outage from May 20th to November 9th 2011 due to a failed network card[3]. The other sensors recorded traffic on all days in the period of study. Factoring in the outage, based on the recorded data, it's reasonable to expect that similar traffic levels would have been observed for this dataset as with the other two sensors in 196/8. What is particularly interesting is that despite this outage, the composition of traffic on 196-1 is very similar to the other two netblocks being monitored in 196/8 as shown in Table I.

4. ANALYSIS

The focus of this paper is to consider the similarity of the activity observed across the five network telescope systems. A summary of the traffic composition across TCP, UDP and ICMP for each dataset is shown in 0 TCP is seen to be the predominant protocol observed, particularly for the three blocks in 196/8 where it accounts for more than 90% of packets. Over 99.97% of traffic observed was accounted for as being one of TCP, UDP or ICMP. Traffic falling outside of these protocols is not being considered as part of this research, and consists predominantly of what appear to be damaged or corrupted, or otherwise nonsensical datagrams. In all cases TCP was the dominant protocol observed followed by UDP and ICMP. The higher proportion of UDP traffic present in 146 and 155, is most likely due to the decreased dilution caused by lower volumes of traffic destined to 445/tcp as compared to the sensor blocks in 196/8. This is discussed further in Section 5A. Datasets 146 and 155 also experienced significantly less traffic than the sensors in 196/8, with both host counts and packet counts being an order of magnitude less.

The remainder of this section comprises a closer look at TCP and UDP traffic observed, and the origins of some of traffic from a network perspective. A summary of the top ten ports for TCP and UDP as observed for the monitored address blocks are shown in Tables II and III respectively. The final area of analysis is a brief discussion considering the hosts which have been observed across multiple sensors.

[3] At this stage three sensors were operated by an external party with traffic only periodically analysed. Post outage the author took over the data collection and management of these sensors.

Table II. Top 10 TCP ports (SYN flag set)

	146		155		196-1		196-2		196-3	
Rank	Port	%TCP	Port	%TCP	Port	%TCP	Port	%TCP	Port	%TCP
1	**445**	21.92	**3389**	8.02	**445**	68.68	**445**	68.82	**445**	69.21
2	**3389**	15.51	**1433**	7.81	**22**	2.33	**22**	2.20	**22**	2.05
3	**1433**	11.95	**445**	6.76	**80**	1.84	**1433**	1.74	**1433**	1.90
4	**80**	10.89	**80**	5.95	**1433**	1.66	**80**	1.64	**80**	1.49
5	**22**	6.12	57471	5.86	**3389**	1.46	**3389**	1.54	**3389**	1.22
6	**8080**	5.12	**22**	4.34	**23**	1.13	49787	1.15	10300	1.13
7	139	4.26	**8080**	2.63	135	0.98	**23**	1.08	135	1.01
8	**23**	3.84	**23**	2.10	39459	0.84	135	0.89	**23**	0.87
9	135	3.59	3072	1.56	**8080**	0.81	**8080**	0.74	**8080**	0.76
10	3306	1.57	135	1.54	25	0.48	5900	0.48	5900	0.46
$\sum_{Top10}$		84.78		46.62		80.22		80.28		80.11

A. TCP

Traffic reported on in this section related only to those TCP datagrams received that had the SYN flag set. As such these packets were deemed to be 'active' in the sense that they would likely generate a response, and potentially establish a TCP session with a target host. Specifically excluded is traffic not matching this criterion which is determined to be backscatter. Only active traffic was considered, as it is felt that this provides a better indication of potentially malicious activity targeting hosts. Backscatter traffic can arise from a number of situations, such as the monitored address space being used in spoofed packets generated as part of a decoy scan or denial of service attack. These typically present as packets arriving with the ACK flag set if ports are open or RST otherwise.

TCP traffic observed across the five datasets is fairly consistent, being dominated by traffic targeting 445/tcp. Seven ports, 22/tcp (ssh), 23/tcp (telnet), 80/tcp (http), 445/tcp (microsoft-ds), 1433/tcp (ms-sql-s), 3389/tcp (rdp) and 8080/tcp (http/proxy), were present in the top ten actively probed ports across all sensors. These ports have been highlighted in bold in Table II. Dataset 155 is somewhat of an anomaly with the top ten ports representing only 46.62% of the TCP packets received, in contrast to the others which are over 80%. In this case the top twenty ports only accounted for 56% of the TCP data received. Other commonly targeted ports observed are 25/tcp (smtp), 135/tcp and 139/tcp which are used by older Microsoft RPC and file sharing implementations, 3306/tcp (mysql) and 5900/tcp (vnc). Scanning for hosts with open ports on common services such as these, particularly at his kind of volume and scale, is a typical pre-cursor to future possible exploitation attempts.

The port 3072/tcp and the ‘high value’ ports of 10300/tcp, 39459/tcp, 49787/tcp and 57471/tcp, are not commonly used by established protocols and couple possibly be scans for backdoors. Without TCP payloads this is difficult to determine with any certainty they do however warrant further exploration beyond the scope of this paper.

Table III. Top 10 UDP ports

	146		155		196-1		196-2		196-3	
Rank	**Port**	**%UDP**	**Port**	**%UDP**	**Port**	**%UDP**	**Port**	**%UDP**	**Port**	**%UDP**
1	**5060**	20.27	**5060**	22.27	**5060**	30.82	**5060**	36.11	**5060**	21.87
2	24003	12.66	**1434**	6.30	21566	8.32	19416	5.62	22549	2.86
3	**1434**	5.57	**137**	2.11	53	6.88	**1434**	4.32	**1434**	2.69
4	**137**	2.14	6257	2.02	**1434**	4.50	**137**	2.65	41560	2.20
5	5159	2.12	32737	1.71	**137**	2.25	6257	1.38	**137**	1.54
6	6257	1.75	53	1.71	6257	1.09	473	1.36	41559	1.21
7	41511	1.64	6568	1.48	9115	0.77	31683	1.20	6257	0.84
8	18261	1.62	60505	1.32	17762	0.63	38834	1.19	53	0.82
9	30989	1.55	43815	1.02	1046	0.63	53	0.87	15401	0.76
10	4375	1.54	39455	0.90	48170	0.57	6655	0.74	64578	0.71
$\sum_{Top10}$		50.86		40.85		56.46		55.44		35.49

B. *UDP*

Observed traffic destined to ports using the UDP protocol on the sensor networks was found to be much more diverse than the case with TCP previously discussed. Only three ports 137/udp (netbios-ns), 1434/udp (ms-sql-m) and 5060/udp (sip) are common across the top ten on all sensors. These ports have been highlighted in bold in Table III. Common exploits exist for the 1434/udp service, in many cases scanning activity is attributable to the SQL Slammer worm, which has been around since January 2003. In this case the attribution can be performed with certainty for a given packet as the payload is present, and can be matched against known samples. The top ten ports in each sensor accounted for more than half the observed traffic in the cases on 145,196-1, and 196-2 and a significant portion in excess of a third in the case of the other two sensors. Port 53/udp is used by DNS and present in four of the sensors and rank in 15th place in dataset 146. Traffic commonly consists of spurious requests, often in the quest for open resolvers, which can be utilised maliciously in a number of ways. As with TCP, there are a significant number of ‘high’ ports, including a number from the ephemeral range (>49152), although Microsoft Windows systems typically use ports in the range 1025-5000 for this purpose.

Exploitation of Voice over IP (VoIP) services including those using SIP as a transport has become popular in recent years. Compromised systems are monetised by on selling calls. This is evidenced by it being the top ranked port in each sensor and accounting for more than 20% of UDP traffic in each case. As seen in the following subsection, significant proportions (19.27%) of the hosts targeting this port have been seen across all sensors. The relatively low host counts would indicate a well-co-ordinated scanning network, rather than random independent systems or attackers. Further investigation would serve as another interesting area to explore in the future.

Table IV. Observed sources across sensors

Target	3389/tcp		445/tcp		5060/udp	
Sensor Count	**Hosts**	**%Sources**	**Hosts**	**%Sources**	**Hosts**	**%Sources**
1	145 609	65.52	7 046 086	76.45	1 475	24.58
2	35 850	16.13	1 611 847	17.48	1 075	17.92
3	18 649	8.39	555 218	6.03	1 401	23.35
4	11 903	5.35	1 905	0.02	893	14.88
5	10 210	4.59	1 527	0.02	1 156	19.27
	222 221		9 216 583		6 000	

C. *OBSERVED CROSS SENSOR HOSTS*

Table IV contains a summary for three of the most popular ports observed in the datasets; 445/tcp, 3389/tcp and 5060/udp. In each case an analysis has been performed looking at the number of hosts observed targeting a port for each sensor. These lists were then combined, and common hosts enumerated, and a classification done based on the number of sensors on which a remote IP was recorded. Of these, 5060/udp was found to be the most interesting, despite the small total number of hosts, due to the fairly high proportion having been observed on multiple sensors. In the case of 3389/tcp more than ten thousand hosts were seen across all five sensors. Further exploration of this mode of analysis will explore this prevalence across sensors over shorter temporal periods.

5. CASE STUDIES

Two specific case studies have been chosen from the datasets. These are an analysis of the active (scanning) traffic destined for ports 445/tcp and 3389/tcp respectively. In both cases the similarity in observed trends across datasets is found to be significant. These also were two of the top TCP ports by packet count.

A. RPC/DCOM (445/TCP)

Port 445/tcp is used by the Microsoft Windows family of operating systems for providing distributed RPC services. This service has a long history of vulnerabilities and associated exploitation. One of the earliest of these was detailed in MS03-026 [23] and later in MS03-039 [24] − and subsequently exploited by the Blaster and Welchia worms in August 2003 [25]. A further vulnerability in the RPC stack was exploited by Sasser in April 2004, taking advantage of the vulnerability disclosed in MS04-011 some seventeen days previously [26]. The problems with the RPC/DCOM stack as implemented in the Microsoft Windows Family operating systems continued and MS06-040 was released in September 2006 [27]. This was further widely exploited following MS08-067 [28], most notably by the Conficker worm. Work has previously been published relating to the observation of Conficker on single network telescopes [29], [30].

What is of interest in this research is the grouping of the datasets into two sets, based on the activity observed relating to this. As discussed in Section 4, only 'active' TCP traffic was analysed, as this was seen to be a more reliable indicator of actual malware activity, and associated scanning for the presence of the vulnerability, either by the malware itself or other sources. The observed counts of packets and distinct sources observed by each target address in the datasets are plotted in Figures 2 and 3 respectively.

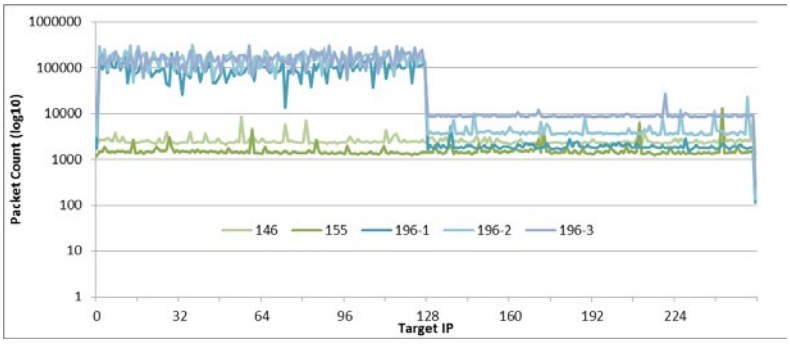

Figure 2. 445/tcp Packet count by target address

In both cases, two trends are notable. The first of these is that the three datasets in 196/8 exhibit fairly similar behaviour. Observations in the 146 and 155 datasets show just fewer than two orders of magnitude less traffic. The second trend is the substantial differentiation between the upper and lower /25 portions of each netblock for the 196/8 datasets. In all cases the highest monitored address x.x.x.255

received substantially less traffic. This reduced traffic volume is most likely due to it generally being considered as a broadcast address for a /24 subnet, and therefore unlikely to be utilised by a host. The sharp change in observed traffic levels occurs at x.x.x.128. This is due to a flaw [31], [32] in the propagation generation algorithm used by Conficker. The net effect of this is that the 2nd and 4th octets of the generated IPv4 address are limited to be in the range 0-127. The 146 and 155 datasets fall outside these ranges, whereas all of those in 196/8 are included. While this finding is not novel it is useful as a means of confirming the function of the data sources. While the majority of hosts scanning the lower /25 range could be regarded as being infected with the Conficker malware, this does not account for all traffic, as evidenced by the sustained scanning of addresses in the upper /25 of the monitored block. This is an important consideration when one considers that traffic targeting 445/tcp is the top destination by packet volume in all datasets with the exception of 155 where it is ranked 3rd.

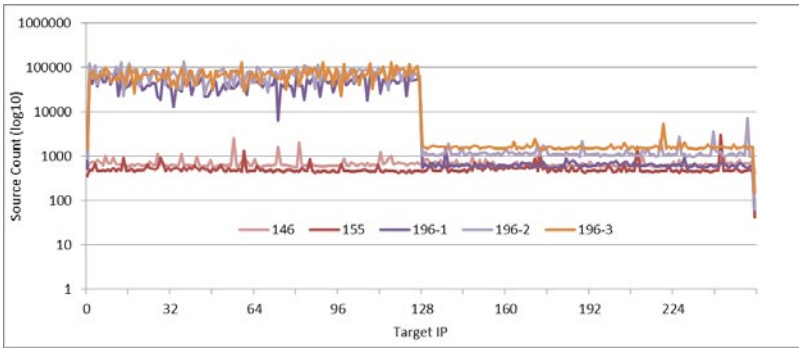

Figure 3. 445/tcp Distinct count by target address

Over the observation period 9 216 583 source IP addresses were observed as emitting datagrams targeting 445/tcp on the monitored networks. Of these 1.6 million (7.48%) were seen on two sensors and 3 432 were observed in four or more datasets. A summary of this can be seen in Table IV. When processing data, care needs to be taken to ensure that the volumes of traffic generated by hosts targeting 445/tcp don't obscure other interesting datasets, with significantly smaller volumes. This is of particular concern in datasets such as those in 196/8 where it accounts for more than 68% of TCP traffic and greater than 62% of the total.

B. *RDP (3389/TCP)*

The Microsoft Remote Desktop Protocol (RDP) runs over port 3389/tcp. This case study was chosen for two reasons; the first being that there has traditionally been

relatively little traffic observed targeting this port. In the first six months of the observation, the average number of packets observed was 324 per day, with 11 sources. This changed significantly from early August 2011, from which time traffic volumes increased substantially. The second reason is that the Microsoft Windows RDP service had a vulnerability disclosed, as detailed in MS12-020 [33]. The Morto worm also targeted this, gaining access to systems with this service exposed by guessing passwords. This was found in the wild and reported by Antivirus vendors on August 28th 2011 [12]. A detailed analysis of the worm and in particular its brute-force password technique can be found at [34].

An overview of the observed traffic can be seen in Figure 4, with Figure 5 containing an enlarged version of the area of interest from January to May 2012. The letters indicating areas of interest correspond in these two figures. The sharp spike in scanning activity can be seen starting in August 2011 (A), reaching a peak on August 24th 2011 (B), with 1 712 sources observed across all sensors. This trend of almost identical activity across all sensors continues until February 22nd 2012 (D) at which point there is a sudden divergence. The synchronised pattern observed in the source count during this period only varies by a few hosts, rarely differing by more than 120, and generally by less than 40 hosts between datasets. The cause for this departure from the trend is unknown. MS12-020 was published on March 13th 2012 and resulted in an almost instantaneous decrease (E) in the levels of scanning activity observed, reaching local minimums by March 20th 2012. The remaining period of observation (F) shows a steady increase in the volume of scanning activity observed. Points G and H are the result of connectivity outages for the two blocks being monitored at one physical site, although in the case of H, sensor 155 is also affected, possibly due to routing problems with the International peering link to the TENET network which were experienced around this time.

The traffic observed on the different network telescopes diverges from February 22nd 2012 (D). From this point, activity for 196-3 and 196-1 drops substantially, but these two remain at very similar levels for the remainder of the observed period, which is of interest given they are in different /16 netblocks, which are not adjacent. The data from 196-2 (which is in the same /16 network as 196-1) experiences an increase in traffic along with 146 and 155, attaining a local peak on February 28th. Traffic levels remain high through to March 18th when a substantial decrease is observed, possibly in response to the release of the Microsoft Security Advisory. From March 23rd, levels stabilise and start increasing again, with 146, 155 and 196-2 following similar paths. The dip in traffic (H) on April 30th 2012, marks a second split in the traffic similarity, with 196-2 experiencing relatively smooth continual growth, in contrast to 146 and 155. These two sensors then display a rapid growth, and maintain high levels of traffic until near the end of the observation period.

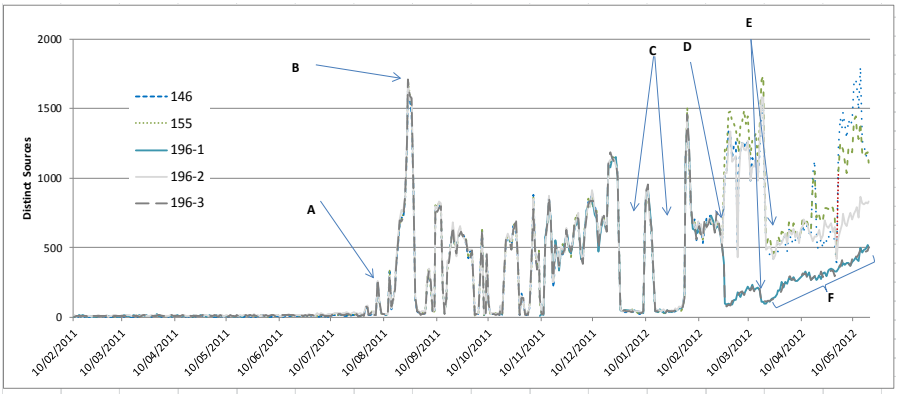

Figure 4. Distinct sources targetting 3389/tcp

While the reasons for the dips in activity as indicated by C are unknown, it is worth noting that the first trough started on December 25th 2011, through to January 8th 2012, followed by a week of traffic and then a lull for two weeks. A further exploration of the events around D and the changing composition of hosts at this point will serve as an area for further research, particularly comparing against other data sets.

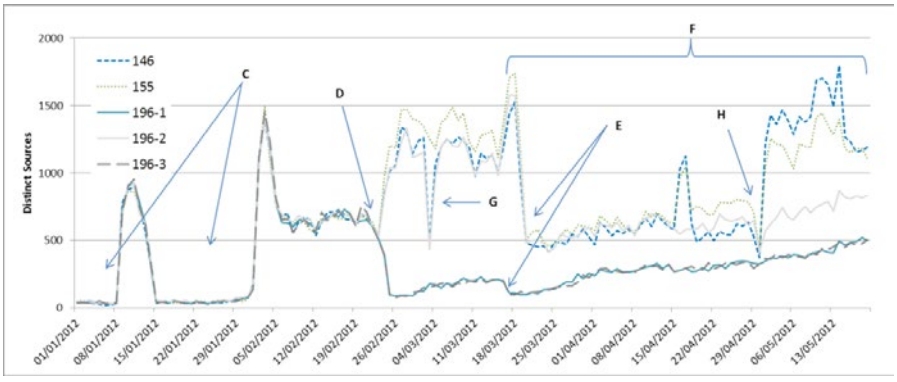

Figure 5. Distinct sources targetting 3389/tcp (1/1/2012 to 20/5/2012)

The most likely reason for the similarity of scanning across five distinct netblocks for much of the period of study is due to a series of co-ordinated scans being run.

The spike in traffic observed in August 2011, prior to the announcement of the Morto worm, can be seen as an early warning of unusual activity being observed on a wide scale, which may warrant further investigation. However as mentioned,

without being able to capture traffic payloads, though tools such as honeypots, it is impossible to definitively state what was responsible for the scanning, although it is interesting that scanning almost ceased for a period following the announcements posted relating to the presence of Morto on August 28th 2011.

6. CONCLUSION

This paper has provided an introduction to the use of network telescopes in a coordinated manner across diverse IPv4 address space. Along with the general characterisation of the observed traffic presented in Section 4, and the highlighting of hosts being seen across multiple sensors, two specific case studies have been presented, illustrating two specific areas of interest within the dataset. These were chosen so as to be able to provide examples of the continued monitoring of an existing threat (in Conficker and similar malicious activity targeting 445/tcp) and the observation of two new and emerging threats targeting 3389/tcp.

A. CHALLENGES AND APPLICATION

The continued use of network telescopes faces a significant challenge. By their nature, they consume address space, which is becoming all the more valuable with IPv4. Work, such as this, demonstrates the effective use of smaller address blocks than have traditionally been used for conducting similar research. The introduction of IPv6 also brings challenges. In the researchers' experience, unsolicited traffic was not observed in the /48 IPv6 sensor that was previously operated. This may be a measure of the lack of general deployment of IPv6, combined with the general infeasibility of scanning such large swathes of address space.

An identified weakness of this collection technique, as previously identified, is the lack of payload for TCP connections, due to the lack of 3-way handshake. This could be mitigated to some extent by using honeypot systems interspersed with the addresses used by the telescope sensors.

The information produced by a network telescope can be used in conjunction with existing network security technologies to allow for a means of shunning or otherwise managing potentially hostile hosts, and protecting clients inside a network. This could be achieved through a variety of means, as appropriate for an organisation, ranging from route black holing to blacklist population. The observed issues, as exhibited by the problems with Conficker's propagation algorithm, highlight the importance in considering the diversity of placement of network telescope sensors in the future. Where possible ranges should be spread across a /16 blocks, and in both halves of a /24 – particularly for researchers with relatively small ($\leq$ /25)

address space being utilised. The viability of a range of address blocks has been demonstrated in terms of the diverse behaviour seen across them. Some of this may be due to numeric locality (malware tends have a preference for 'close' address space), rather than poor implementations as in the case of Conficker.

B. *FUTURE WORK*

The datasets used in the research can still be further analysed, particularly from the point of an extended geopolitical and topological analysis, such as that performed in [9]. Further exploration of these datasets and other subsequently collected datasets may provide better insight into the spread of malware and related malicious activity on a global scale, as well as how to better monitor and defend against these threats.

A goal of the researcher is to foster improved information sharing within the security research community. One challenge around this is the confidentiality of datasets (particularly relating to the ranges monitored), and the size of the raw captures. This can to some extent be mitigated though the publishing of metrics relating to observed data, rather than the data itself. A discussion of the specifics behind this, and some recommended metrics which can be used can be found in [9] and [35]. A significant step towards this would be the completion of an extended data processing framework what was prototyped for this analysis, which could publish reports on a regular basis.

REFERENCES

[1] R. Pang, V. Yegneswaran, P. Barford, V. Paxson and L. Peterson, «Characteristics of internet background radiation,» in *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, 2004.

[2] E. Wustrow, M. Karir, M. Bailey, F. Jahanian and G. Huston, «Internet background radiation revisited,» in *Proceedings of the 10th annual conference on Internet measurement*, New York, NY, USA, 2010.

[3] D. S. Pemberton, «An Empirical Study of Internet Background Radiation Arrival Density and Network Telescope Sampling Strategies,» 2007.

[4] F. Baker, W. Harrop and G. Armitage, «RFC6018 IPv4 and IPv6 Greynets,» IETF, September 2010. [Online]. Available: http://www.ietf.org/rfc/rfc6018.txt.

[5] W. Harrop and G. Armitage, «Greynets: a definition and evaluation of sparsely populated darknets,» in *MineNet '05: Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, New York, NY, USA, 2005.

[6] J. Goebel, T. Holz and C. Willems, «Measurement and Analysis of Autonomous Spreading Malware in a University Environment,» in *Detection of Intrusions and Malware, and Vulnerability Assessment*, Springer Berlin / Heidelberg, 2007, pp. 109-128.

[7] D. Moore, «Network Telescopes: Observing Small or Distant Security Events,» August 2002. [Online]. Available: http://www.caida.org/publications/presentations/2002/usenix_sec/.

[8] D. Moore, C. Shannon, G. . M. Voelker and S. Savage, «Network Telescopes: Technical Report,» 2004. [Online]. Available: http://www.caida.org/outreach/papers/2004/tr-2004-04/tr-2004-04.pdf.

[9] B. Irwin, «A framework for the application of network telescope sensors in a global IP network,» Grahamstown, 2011.

[10] Microsoft, «Virus alert about the Win32/Conficker worm (KB962007),» August 18 2008. [Online]. Available: http://support.microsoft.com/kb/826234.

[11] Microsoft, Win32/Conficker, Jan *8* 2009. Updated: Nov 10, 2010. [Online]. Available: http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Name=Win32/Conficker

[12] Microsoft, «Worm:Win32/Morto.A,» 28 August 2011. [Online]. Available: http://www.microsoft.com/security/portal/threat/encyclopedia/entry.aspx?name=Worm%3AWin32%2FMorto.A.

[13] D. Moore, G. Voelker and S. Savage, «Inferring Internet Denial-of-Service Activity,» in *In Proceedings of the 10th Usenix Security Symposium*, 2001.

[14] CERT, *CERT Advisory CA-2001-19 "Code Red" Worm Exploiting Buffer Overflow In IIS Indexing Service DLL,* 2001.

[15] D. Moore and C. Shannon, «The CAIDA Dataset on the Code-Red Worms - July and August 2001, (collection),» August 2001. [Online]. Available: http://www.caida.org/data/passive/codered_worms_dataset.xml.

[16] D. Moore, C. Shannon and K. Claffy, «Code-Red: a case study on the spread and victims of an internet worm,» in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, New York, NY, USA, 2002.

[17] C. Shannon and D. Moore, «The Spread of the Witty Worm,» *IEEE Security and Privacy,* vol. 2, pp. 46-50, July 2004.

[18] C. Shannon and D. Moore, «The CAIDA Dataset on the Witty Worm - March 19-24, 2004, (collection),» March 2004. [Online]. Available: http://www.caida.org/data/passive/witty_worm_dataset.xml.

[19] A. Kumar, V. Paxson and N. Weaver, «Exploiting underlying structure for detailed reconstruction of an internet-scale event,» in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, Berkeley, CA, USA, 2005.

[20] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford and N. Weaver, «Inside the Slammer Worm,» *IEEE Security and Privacy,* vol. 1, no. 4, pp. 33-39, 2003.

[21] U. Harder, M. W. Johnson, J. T. Bradley and W. J. Knottenbelt, «Observing Internet Worm and Virus Attacks with a Small Network Telescope,» *Electronic Notes in Theoretical Computer Science,* vol. 151, no. 3, pp. 47-59, #jun# 2006.

[22] C. Zou, N. Duffield, D. Towsley and W. Gong, «Adaptive defense against various network attacks,» *IEEE Journal on Selected Areas in Communications,* vol. 24, no. 10, pp. 1877-1887, 2006.

[23] Microsoft, «MS03-026 : Buffer Overrun In RPC Interface Could Allow Code Execution (KB823980),» July 16 2003. [Online]. Available: http://www.microsoft.com/technet/security/Bulletin/MS03-026.mspx.

[24] Microsoft, «MS03-039 : Buffer Overrun In RPCSS Service Could Allow Code Execution (KB824146),» September 10 2003. [Online]. Available: http://www.microsoft.com/technet/security/Bulletin/MS03-039.mspx.

[25] Microsoft, «Virus alert about the Nachi worm (KB826234),» August 18 2003. [Online]. Available: http://support.microsoft.com/kb/826234.

[26] Microsoft, «MS04-011: Security Update for Microsoft Windows (KB835732),» April 13 2004. [Online]. Available: http://www.microsoft.com/technet/security/bulletin/MS04-011.mspx.

[27] Microsoft, «MS06-040 : Vulnerability in Server Service Could Allow Remote Code Execution (KB921883),» September 12 2006. [Online]. Available: http://www.microsoft.com/technet/security/bulletin/ms06-040.mspx.

[28] Microsoft, «MS08-067 : Vulnerability in Server Service Could Allow Remote Code Execution (KB958644),» Oct 23 2008. [Online]. Available: http://www.microsoft.com/technet/security/Bulletin/MS08-067.mspx.

[29] E. Aben, «Conficker/Conflicker/Downadup as seen from the UCSD Network Telescope,» February 2009. [Online]. Available: http://www.caida.org/research/security/ms08-067/conficker.xml.

[30] B. Irwin, «A network telescope perspective of the Conficker outbreak,» in *Proceedings of Information Security South Africa (ISSA)*, Sandton, South Africa, 2012.

[31] M. Richard and M. Ligh, «{Making fun of your malware},» August 2009. [Online]. Available: https://www.defcon.org/images/defcon-17/dc-17-presentations/defcon-17-michael_ligh-matt_richard-making_fun_of_malware.pdf.

[32] Carnivore.IT, «Conficker does not like me?,» 3 November 2009. [Online]. Available: http://carnivore.it/2009/11/03/conficker_does_not_like_me.

[33] Microsoft, «Microsoft Security Bulletin MS12-020 Vulnerabilities in Remote Desktop Could Allow Remote Code Execution (2671387),» 13 March 2012. [Online]. Available: http://technet.microsoft.com/en-us/security/bulletin/ms12-020.

[34] T. Bitton, «Morto Post Mortem: Dissecting a Worm,» September 2011. [Online]. Available: http://blog.imperva.com/2011/09/morto-post-mortem-a-worm-deep-dive.html.

[35] B. Irwin, «Network Telescope Metrics,» in *Southern African Telecommunications and Applications Conference (SATNAC)*, George, South Africa, 2012.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Illicit Network Structures in Cyberspace

Kaarel Kalm
Department of Security and Crime Science
University College London
kaarel.kalm.12@ucl.ac.uk

Abstract: Different types of covert illicit networks in cyberspace hold the potential to become actors in cyber conflicts. Current literature on structures of covert networks in cyberspace is often constrained by the lack of quantitative data and researchers mostly focus on networks operating outside the cyberspace. The purpose of this paper is to review the current state of research into illicit networks in cyberspace and to apply the terminology and concepts of Social Network Analysis on criminal organisations operating online. Social Network Analysis is a quantitative data analysis method, which can identify hierarchies, subgroups, individuals and their relative importance in covert illicit networks, by using data from multiple sources (academic research, law enforcement, black market trading, semantic web analysis etc.). Here I explore how Social Network Analysis offers methods to discover hidden structures of covert networks in cyberspace.

Keywords: *cybercrime, covert networks, illicit networks, social network analysis*

1. INTRODUCTION

Covert networks operating in cyberspace are involved in organized crime, espionage, terrorism, trafficking and a list of other illicit or destructive activities with a high impact on society. Current literature on structures of covert networks is often constrained by the lack of quantitative data and researchers mostly focus on illicit networks operating outside the cyberspace [1], [2]. Along with the advances of technology, the nature and activity of covert networks in cyberspace have changed. A sophisticated underground economy has emerged, along with ideology driven "dark webs", and state sponsored cybercrime groups. Covert networks are stateless, fluid and adaptable and function as the main facilitators of trafficking, proliferation and terrorism [3]. They present an asymmetric threat to nation states and have emerged as one of the main concerns of international political agenda [4]. A research commissioned by BAE Systems in 2012 found that 80 per cent of cybercrime can be attributed to organised groups including hybrid criminal groups which combine online and offline offending [5].

The need to have in-depth knowledge of covert networks will become increasingly acute as such networks develop towards holding a very high threat potential. As this trend is unlikely to reverse, the practical aspects of network identification are of importance to policy makers and law enforcement. Ability to describe and map the properties of such networks is a basis for developing effective prevention, detection and disruption mechanisms. Obtaining information about covert networks is made difficult by their very nature and purpose. However, covert networks have to constantly manage the trade-off between security and efficacy. To successfully function, information must be exchanged inside the network, and if necessary, between network members and outsiders. Through exchange of information, networks become exposed for detection and analysis [6].

This research looks at the problems of covert networks and the general threats they present in the context of cyber conflicts. I shall describe the typology of covert networks in cyberspace and then describe how Social Network Analysis (SNA) can identify hierarchies, subgroups, individuals and their relative importance in such networks. I shall review existing data about covert networks from multiple sources and suggest generic structures of four different type of networks. With each type of network, relevant background is given and data is examined in the social network context. This is followed by a short list of recommendations on how policy makers and law enforcement can counter the threat that arises from specific types of covert networks and what tactics might be useful for detection and disruption. Discussion of limitations of social network analysis methods follows along with some suggestions for further research. The questions of data availability and the cooperation between law enforcement and academia are also briefly addressed.

In this paper, covert and illicit networks refer to organised groups of individuals that are involved in criminal activities taking partly or entirely place in cyberspace. This includes activities associated with crimes for profit, terrorism, espionage, destruction and disruption of property, antisocial behaviour etc. As the definitions of cyber conflict and cybercrime are still very much up for debate [7], [8], I do not aim to make a distinction between them in this paper. Rather, I presuppose that any cyber conflict consists of criminal acts that are enabled by technology. As different criminal acts require different organisational structures, covert networks in cyberspace can take several forms.

There are at least four distinct types of covert illicit networks operating in cyberspace—traditional criminal organisations, cybercriminal organisations, ideologically motivated organisations, and state sponsored organisations [9], [10]. Those four types of networks have both unique and common properties. Unique properties derive primary from motivational and ideological factors. Traditional criminal networks and cybercriminals are mostly profit oriented and therefore more engaged with outside actors. Ideological and government sponsored groups are more closed but also more interconnected. Overlapping properties arise from the need to operate in secrecy, victimisation and structural resilience. They also use similar tactics and technology. The theory of crime-terror nexus also asserts that methods invented and successfully applied by one type of criminal organisation are likely to be appropriated by another type of organisation, regardless of underlying ideology [11]. As such, all four types of covert networks are considered to hold the potential of becoming participants in a cyber conflict and are therefore included in this analysis.

2. SOCIAL NETWORK ANALYSIS

Social network analysis (SNA) is defined as study of structural aspects of networks. Social network theory argues that “Any action of actors is not isolated but correlated. The relationship ties among them are transmission channels of information and resources and network relation structure decides their action opportunities and results” [12]. A social network is represented by nodes (actors) and links (relationships) between those nodes. In the context of this study, those nodes are either people or technical facilitator (e.g. websites, microblogs, forums etc.). SNA is not a precise analysis technique but incorporates a set of mathematical, graphical and theoretical tools to measure the location of network nodes and to identify relationships between them [13].

There has been a considerable increase of interest in networks analysis theories in the past decade. A lot of research on social networks has been done in biology,

geography, economics, information science and sociology [14]. SNA has also been successfully applied to security studies, although the lack of quantitative data has forced researchers to mostly focus on illicit networks operating outside the cyberspace. Social network analysis can be used to establish key members and structural weaknesses of criminal organisations [15], [16], evaluate relative influence and connectedness of a particular actor in a networks [1] and to identify hubs and bridges in illicit networks to study effective disruption tactics [17].

Based on their characteristics, networks can be classified as random, small-world or scale-free. *Random networks* have a small number of nodes and a small number of links between them. *Small-world networks* have a larger number of nodes and a small number of links between them. Unlike random networks, small-world networks often contain clustering of nodes. *Scale-free networks* have similar properties to small-world networks, with an important addition of power-law degree distribution. Power-law degree distribution implies that while most nodes still have a small number of connections, a very small number of nodes are highly connected [18]. In addition to general topological properties, SNA enables researchers to measure several descriptive metrics inside the networks. In relation to nodes—centrality, betweenness, clustering, and eigenvector values describe the relative influence and connectedness of a particular actor in the network. Note that connectedness and influence are separate descriptives as the most connected node might not be the most influential and vice versa.

Node *centrality* measures the location of a node in relation to the centre of the network. The more central a node is, the smaller is the number of links connecting it to other nodes. In human networks, a person with the highest degree of centrality can reach all other people in the network through smallest number of connections. This person is likely in a leadership position, binding the network (or a part of it) together. If network nodes represent technical facilitators, content severity is an additional indicator of influence. From law enforcement perspective, monitoring nodes with high centrality can provide information and removing them can break larger networks into smaller cells.

Betweenness of a node measures the number of shortest connections between two other nodes passing through that particular node. A person with a high measure of betweenness functions as a bridge for communications and should be a prime target for monitoring by law enforcement.

Eigenvector values identify highly connected nodes that are connected to other highly connected nodes. This is also known as the “rich club” effect or the “rich-get-richer” phenomenon [17], [19], where high degree connected nodes tend to become even more interconnected resulting in subgroup clustering. In human networks this

implies that important people are and will become connected to other important people. From law enforcement perspective, it would be meaningful to target such individuals simultaneously. This subgroup holds most information and removing just single individuals from it is less likely to disrupt the rest of the network.

Clustering or transitivity measures the likelihood that if a link exists between nodes A and B, and nodes A and C, it also exists between modes B and C. Link structures are basic indicators of clustering, as for example, described above by the "rich club" effect. In networks with low overall centrality, clustering may still occur in forms of small subgroups connected by central authority. Such groups may not hold information on the larger network, but they also have capabilities to act independently from it.

As both general topological properties and node descriptive metrics influence measures available for disruption, covert network analysis should follow three logical steps: (a) identify covert network type; (b) analyse network characteristics; and (c) evaluate key nodes in the network.

3. COVERT NETWORKS IN CYBERSPACE

As described above, networks can be differentiated between random, small-world and scale-free networks. Covert networks operating in cyberspace follow either financial or ideological motivations, making the existence of a random type of illicit network very unlikely. Yet researchers can use random networks as comparisons models. Most basic random networks are characterised by low average path lengths and low clustering measures [20]. In human networks this would mean that all members are closely connected, while no hierarchical structures and subgroups exist. Such networks are very robust and node removal would have little effect on their overall performance [21]. Concurrently, their overall performance would also be very low, resulting from absence of leadership and coordination. Disruption of covert networks can take a form of targeted attacks against key individuals, simultaneous attacks against a subgroup, progressive attacks or random removal of network members. While random networks might be robust against most forms of attack, small-world and scale-free networks have properties that make some attacks more effective than other. In scale-free networks, a small number of members are highly connected, making the networks robust against random removal of members but vulnerable to targeted attacks. In small-world networks, a larger number of members are well connected to each other, duplicating connections. Therefore such networks are more vulnerable to random attacks (compared to scale-free networks) but targeted attacks may not be sufficient to disable information flow inside the network. Following, I shall explore the network structures of four types of covert networks from the practical viewpoint of detection and disruption.

A. TRADITIONAL CRIMINAL ORGANISATIONS

Organised crime is mostly market-driven, even if the consumer need is created by the organisations themselves. This means that they provide services driven by financial rationality. Interconnected global economy and the spread of Internet has created opportunities and incentives for organised criminal groups to exploit competitive advantages cyberspace can offer. Key drivers of international economy like financial deregulation, technological development, interconnectedness of infrastructure and global labour markets have enabled a surge in trade of drugs, arms, illegal goods, people, and money [22]. A report from UK Serious and Organised Crime Agency suggests that advances in technology are increasingly exploited by members of organised crime groups. The internet provides criminals with tools to commit traditional crimes in a more sophisticated way along with opportunities for new types of crime [23].

Technology-enabled crimes carried out by traditional criminal organisations include network intrusions, identity frauds, online scams, malware distribution etc. Technology connects a geographically very distant demand and supply sides of the illicit market previously outside the sphere of interest of traditional organised crime. Several crime organisations have also established a strong online presence for propaganda and recruitment purposes, to issue threats and monitor the media [24].

Several empirical studies have used social network analysis on police arrest-data and court-data to identify criminal networks [1], [25]. There is also a significant amount of open source data available to enable social network analysis by non-law enforcement organisations [24], [26]. The main findings from those studies indicate that similar social network characteristics describe both offline and online traditional criminal groups. Such networks have small average path lengths and high clustering or transitivity metrics. Therefore they can be classified as small-world networks. This means that the covert network consists of a group of well-connected members who can reach each other easily. Connections with other networks are low as is expected from groups competing for resources. Traditional criminal networks demonstrate also and overall low link density, implying that network members interact mostly inside the network and with a certain set of other members [17]. This can be explained with the traditional structure of organised crime groups, where individual members are tasked with specific assignments and do not operate outside those limits.

The question of power-law distribution in traditional criminal networks can be dependent on their historical structure. Where the crime network operates in strict top-down hierarchical manner, scale-free properties can be not as apparent as

the power dynamics inside the network are more stable. Whereas in horizontally organised networks, the power-law distribution can be more apparent, along with the "rich club" effect [9].

B. CYBERCRIMINAL ORGANISATIONS

Cybercriminal organisations form and operate online and are engaged in technology-enabled crime. As such crimes require specific knowledge and experience, both individual members and the overall networks structure differs from traditional criminal organisations operating in cyberspace. Cybercriminal networks are characterised by technically capable members, anonymous (in relation to real identities) interaction, and opportunistic financial motivations [10].

Cybercriminal networks face a task of leveraging security with the need to interact with outside members willing to pay for their services. In comparison with other types of covert networks, they are most directly involved in what might be described as black market dynamics. Members often take part in direct price negotiations and sales, are influenced by competitors' offerings and customers' demands. As the online black market is increasing, such dynamics can lead to a fully functional marketplace with high utility and low participation risks [19]. Reports suggest disappearance of independent and small-group hackers and appearance of hierarchical cybercrime networks with role-based memberships [27], [28].

Research data about cybercriminal networks suggest that unlike traditional criminal organisations in cyberspace, cybercriminal networks are not scale free. This indicates that while network members form ties based on preference, they also form a substantial number of random links. As cybercriminals have to participate in market activities, there is a need to find orders for services and customers for products. Random link formation can be attributed to members seeking buyers for their services or looking for business opportunities through cooperation. Networks engaged in online black markets are also highly clustered with evidence of hierarchical structures in the networks [19]. This is a result of participating in market activity, where certain positions are established – administrator, escrow, seller, buyer, etc. The need to establish trust in the network requires some members to reveal their transactions to build-up trust and acquire more customers. As more active and more contributing members are likely to have an exponential increase of links to other members, a clustering formation appears [29]. As the overall network is not scale-free, it is more robust against targeted attacks as well as random removal of members. The network can also easily incorporate new members to replace those that are removed. However, gradual appearance of some very well connected members can provide sufficient grounds for targeted attacks that are likely to disrupt the networks but unlikely to disable it.

C. IDEOLOGICALLY MOTIVATED ORGANISATIONS

Ideologically and politically motivated organisations are increasingly taking to cyberspace. This follows a similar trend as observed for traditional criminal organisations. Same technological factors are facilitating this move, but in addition to that, an increasing nexus between ideologically motivated and financially motivated criminal organisations is becoming apparent [11]. As ideologically motivated organisations in cyberspace require increasingly more funding and are unlikely to establish any legitimate base for financial income, it is likely that they will also get increasingly more involved in online illicit economy. In addition to financing, such groups are using internet as a platform for communication and publicity. As their motivation for action is ideological, they actively seek widespread publicity and are largely indiscriminate in their use of force or violence. The internet serves as an effective facilitator of ideological propaganda and recruitment.

Findings from academic studies indicate that ideologically and politically motivated networks in cyberspace are described by very high subgroup clustering with long path lengths compared to traditional and cybercriminal networks. They are scale-free networks with evidence of the 'rich club' effect, where influential members are well interconnected. The power-law distribution is also evident in ideological networks [17] and is also supported by data from web forums analysis', where a small number of members are the most prolific communicators, followed with a sharp decay in number of postings by other members. [28]. High subgroup clustering can result from the overall trend of ideological networks becoming more fluid and horizontal in their structure as well as from recruitment practises, where new members are indoctrinated by a certain subgroup [4]. Members are characterised by a small number of in- and outwards connections, meaning that a member's knowledge about the larger network is limited. The member also has low impact on other members and the network has multiple leadership figures on different levels [12].

Multiple leadership positions make the network as a whole more resilient but the smaller subgroups vulnerable to targeted attacks. This represents a calculated risk on increasing secrecy while reducing operational capability. According to Drozdova and Samoilov [30] "In environments dominated by hostile opponents and where there is significant resource imbalance and incomplete information, the choice of fault-intolerant network organizations structure for clandestine mission networks helps protect the broader organization by minimizing its internal connectivity and allowing all parties plausible deniability of their relations".

D. STATE SPONSORED ORGANISATIONS

State sponsored cybercriminal organisations impose highest threats in the context of cyber conflict as they lack many properties that expose other type of covert networks. They are not directly financially motivated, opportunistic nor ideologically constrained. As state sponsored cybercrime mostly involves espionage and technical operations, a substantial amount of resources is required. While the direct cost of software development and deployment may not be that high in comparison to possible gains from all forms of cybercrime, technology development and operational secrecy requirements impose substantial demands on state sponsors [31]. State sponsored cybercrime also carries a high risk of conflict escalation through retaliation and confrontation, possibly leading to a direct cyber conflict or –war [32].

Alleged state sponsored cyber attacks are a common theme in media with regular reports claiming Russian hackers attacking USA, USA and Israeli hackers attacking Iran, Iranian hackers attacking China, Chinese hackers attacking USA and India, etc. The Director General of the UK Security Service [33] has called the extent of cybercrime "astonishing – with industrial-scale processes involving many thousands of people lying behind both state sponsored cyber espionage and organised cybercrime." There are also well-published incidents of cyber attacks against Estonia, Georgia and Azerbaijan. Yet the data on state sponsored cybercriminal organisations is sparse and academic access and analysis of it is almost non-existent.

Lack of empirical data on state sponsored cybercriminal groups can be explained by several factors. First of all, relevant data could be unavailable for academic research as organisations collecting it are unwilling to share it. Existing data could also be inconclusive, making the academic analysis meaningless and further discouraging its sharing. Secondly, relevant data might actually refer to regular cybercriminal groups that act on behalf of the state when necessary. Several hypotheses have been proposed by researchers on how state structures incorporate cybercriminals and "patriotic hackers" [7], [32], [34]. Based on those hypotheses, some state sponsored cybercriminal networks should display similar properties to regular cybercriminal networks – small-world and non scale-free metrics. While cybercriminal networks have to balance exposure risks with a need to interact with customers, state sponsored groups have no need to establish trust with possible buyers. This should reduce the number of random links and clustering in the network. A special case should be made for state structures are directly participating in cybercriminal activities. While empirical data on them is again non-existent, an argument could be made that their networks will reflect the bureaucratic structures of the state and secrecy oriented structures of traditional covert government organisations. This

would imply hierarchical structures, small size of the network and short average link paths with few very well connected members.

4. IMPLICATIONS AND DISCUSSIONS

Existing data on covert networks in cyberspace allows division of such networks into four groups—traditional criminal organisations, cybercriminal organisations, ideologically motivated organisations, and state sponsored organisations. As different criminal motivations require different organisational structures, covert networks in cyberspace take several forms. Traditional criminal networks are likely to have small-world and scale-free properties. They are resilient to random removal of members but vulnerable to targeted attacks. Cybercriminal networks are small-world and not scale-free. There are preferential links between members, but also a substantial amount of random links. This results from the need to engage with possible clients and establish trust on the market. Both random and targeted removal of members has limited effects as the network can easily incorporate new members. Ideological networks are small-world and scale-free with high sub clustering coefficients. Members have few connections and little influence in the network. High number of subgroups indicates a need for targeted attacks but the overall network is relatively robust to them. Empirical data on state sponsored groups is too scarce to draw meaningful conclusions. If a state has outsourced its cybercriminal activities, similar network properties should be apparent as in regular cybercriminal networks. If state structures are directly participating in cybercriminal activities, hierarchical bureaucratic structures should be expected.

In studying covert networks, this paper has largely ignored the social psychological aspects of networks formation. Arguably, some psychological factors are incorporated into members' link formation and clustering preferences but it would be unwise to assume, that all network dynamics can be described by link paths, clustering coefficients and leadership hierarchies. The human component of covert networks should not be ignored but rather attempts should be made to incorporate that into the analysis. There have been advances in studies of cybercriminal profiling that could be included into future research of covert networks in cyberspace. Focusing on social networks has also disregarded what is popularly known as 'lone wolf' offending. Lone actors are capable of inflicting serious damage in the cyberspace and should also be regarded as possible participants or initiators of cyber conflicts. As 'lone wolf' criminals by definition do not form co-offending groups, social network analysis cannot provide much insight into their activities. At the same time, the very nature of the internet is likely forcing 'lone wolf' offenders into participating in some kinds of social networks to acquire know-how and resources for attacking. Whether indications of 'lone wolf' offending can be found from analysing social networks is another topic for future research.

The question of data availability is a major factor in covert networks research. Sufficient data might be available to law enforcement but the lack of resources and need for operational secrecy hinder their analysis and distribution. This is not a criticism addressed at organisations investigating and countering covert networks but a recognition that law enforcement is always lacking resources and has to triage to prioritise their actions. A solution would be a deeper cooperation between law enforcement and academia. Understandably there are a lot of obstacles that would have to be overcome but the existing studies assert that actionable data and insight can be gleaned from such research. As covert networks in the cyberspace are increasingly developing towards holding very high treat potential, development of effective counter-measures requires active research of such networks, their structure and dynamics.

REFERENCES

[1] M. Tayebi, U. Glässer, and P. Brantingham, "Organized Crime Detection in Co-offending Networks," in IEEE- 9th International Conference Proceedings, 2011.

[2] M. Coscia and V. Rios, "How and where do criminals operate? Using Google to track Mexican drug trafficking organizations," Center for International Development at Harvard, Oct. 2012.

[3] R. Dietz, "Illicit Networks: Targeting the Nexus Between Terrorists, Proliferators, and Narcotraffickers," Naval Postgraduate School, Dec. 2010.

[4] M. Eilstrup-Sangiovanni and C. Jones, "Assessing the Dangers of Illicit Networks," International Security, vol. 32, no. 1, pp. 7–44, 2008.

[5] BAE Systems, "Organised crime in the digital age," 2012. Available at: www.baesystemsdetica.com/news/organised-crime-in-the-digital-age/

[6] R. Lindelauf, P. Borm, and H. Hamers, "The influence of secrecy on the communication structure of covert networks," Social Networks, vol. 31, no. 2, pp. 126–137, May 2009.

[7] J. Carr, Inside Cyber Warfare: Mapping the Cyber Underworld (Google eBook). O'Reilly Media, Inc., 2011, p. 314.

[8] R. A. Clarke and R. Knake, Cyber War: The Next Threat to National Security and What to Do About It (Google eBook). HarperCollins, 1082, p. 320.

[9] K.-K. R. Choo and R. G. Smith, "Criminal Exploitation of Online Systems by Organised Crime Groups," Asian Journal of Criminology, vol. 3, no. 1, pp. 37–59, Nov. 2007.

[10] K.-K. R. Choo, "Organised crime groups in cyberspace: a typology," Trends in Organized Crime, vol. 11, no. 3, pp. 270–295, Jul. 2008.

[11] [T. Makarenko, "The Crime-Terror Continuum: Tracing the Interplay between Transnational Organised Crime and Terrorism," Global Crime, vol. 6, no. 1, pp. 129–145, Feb. 2004.

[12] S. Duo-Yong and G. Shu-Quan, “Study on covert networks of terrorists based on interactive relationship hypothesis,” 2011 IEEE International Conference on Intelligence and Security Informatics, pp. 26–30, 2011.

[13] A. Reid, M. Tayebi, and R. Frank, “Will the Defendants Please Rise? A Social Network Analysis of Accused Individuals in the Criminal Court System,” Simon Fraser University, pp. 1–24.

[14] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, “Network analysis in the social sciences.,” Science, vol. 323, no. 5916, pp. 892–5, Feb. 2009.

[15] M. Sparrow, “The application of network analysis to criminal intelligence: An assessment of the prospects,” Social Networks, vol. 13, no. 3, pp. 251–274, Sep. 1991.

[16] M. Sparrow, “Mapping Networks of of Terrorist Terrorist Cells,” Connections, vol. 24, no. 3, pp. 43–52, 2002.

[17] J. Xu and H. Chen, “The topology of dark networks,” Communications of the ACM, vol. 51, no. 10, p. 58, Oct. 2008.

[18] R. Albert, H. Jeong, and A. Barabasi, “Error and attack tolerance of complex networks,” Nature, vol. 406, no. 6794, pp. 378–82, Jul. 2000.

[19] M. Yip, N. Shadbolt, and C. Webber, “Structural analysis of online criminal social networks,” 2012 IEEE International Conference on Intelligence and Security Informatics, pp. 60–65, Jun. 2012.

[20] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks.,” Nature, vol. 393, no. 6684, pp. 440–2, Jun. 1998.

[21] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, “Attack vulnerability of complex networks.,” Physical review. E, Statistical, nonlinear, and soft matter physics, vol. 65, no. 5 Pt 2, p. 056109, May 2002.

[22] M. Naím, “The Fourth Annual Grotius Lecture : Five Wars of Globalization,” American University International Law Review, vol. 18, no. 1, pp. 1–18, 2002.

[23] UK Home Office, “Extending our reach: a comprehensive approach to tackling serious organised crime,” 2009. Available at: www.official-documents.gov.uk/document/cm76/7665/7665.asp

[24] M. Coscia and V. Rios, “How and where do criminals operate? Using Google to track Mexican drug trafficking organizations,” Center for International Development at Harvard, no. 57. p. 23, 2012.

[25] U. Glässer and M. Tayebi, “Estimating Possible Criminal Organizations from Co-offending Data,” Public Safety Canada, 2012.

[26] T. J. Holt and E. Lampke, “Exploring stolen data markets online: products and market forces,” Criminal Justice Studies, vol. 23, no. 1, pp. 33–50, Mar. 2010.

[27] Y. Ben-Itzhak, “Organised cybercrime and payment cards,” Card Technology Today, vol. 21, no. 2, pp. 10–11, Feb. 2009.

[28] C. Lu and W. Jen, “Cybercrime & cybercriminals: An overview of the Taiwan experience,” Journal of Computers, vol. 1, no. 6, pp. 11–18, Sep. 2006.

[29] V. Benjamin and H. Chen, "Securing cyberspace: Identifying key actors in hacker communities," 2012 IEEE International Conference on Intelligence and Security Informatics, pp. 24–29, Jun. 2012.

[30] K. Drozdova and M. Samoilov, "Predictive analysis of concealed social network activities based on communication technology choices: early-warning detection of attack signals from terrorist organizations," Computational and Mathematical Organization Theory, vol. 16, no. 1, pp. 61–88, Aug. 2009.

[31] J. B. Sheldon, "Strategic State of the Art : Attackers and Targets in Cyberspace," Journal of Military and Strategic Studies, vol. 14, no. 2, pp. 1–19, 2012.

[32] H. Lin, "Escalation Dynamics and Conflict Termination in Cyberspace," Strategic Studies Quarterly, vol. 6, no. 3, pp. 46–70, 2012.

[33] J. Evans, "The Olympics and Beyond," in Lord Mayor's Annual Defence and Security Lecture, 2012. Available at: https://www.mi5.gov.uk/home /about-us/who-we-are/staff-and-management/director-general/speeches-by-the-director-general/the-olympics-and-beyond.html

[34] N. Kshetri, "Pattern of global cyber war and crime: A conceptual framework," Journal of International Management, vol. 11, no. 4, pp. 541–562, Dec. 2005.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Threat Implications of the Internet of Things

Michael J. Covington
Security Intelligence Operations
Cisco Systems, Inc.
San Francisco, California, USA
Michael.Covington@cisco.com

Rush Carskadden
Click Security
Austin, Texas, USA
Rush@clicksecurity.com

Abstract: There are currently more objects connected to the Internet than there are people in the world. This gap will continue to grow, as more objects gain the ability to directly interface with the Internet or become physical representations of data accessible via Internet systems. This trend toward greater independent object interaction in the Internet is collectively described as the Internet of Things (IoT).

As with previous global technology trends, such as widespread mobile adoption and datacentre consolidation, the changing operating environment associated with the Internet of Things represents considerable impact to the attack surface and threat environment of the Internet and Internet-connected systems.

The increase in Internet-connected systems and the accompanying, non-linear increase in Internet attack surface can be represented by several tiers of increased surface complexity. Users, or groups of users, are linked to a non-linear number of connected entities, which in turn are linked to a non-linear number of indirectly connected, trackable entities. At each tier of this model, the increasing population, complexity, heterogeneity, interoperability, mobility, and distribution of entities represents an expanding attack surface, measurable by additional channels, methods, and data items. Further, this expansion will necessarily increase the field of security stakeholders and introduce new manageability challenges.

This document provides a framework for measurement and analysis of the security implications inherent in an Internet that is dominated by non-user endpoints, content in the form of objects, and content that is generated by objects without direct user involvement.

Keywords: *Internet of Things, attack surface, threat evolution, security intelligence*

1. INTRODUCTION

There are currently more objects connected to the Internet than there are people in the world [1]. This gap will continue to grow, as more objects gain the ability to directly interface with the Internet or become physical representations of data accessible via Internet systems. This trend toward greater object interaction in the Internet is collectively described as the Internet of Things (IoT). As with previous global technology trends, such as widespread mobile adoption and datacentre consolidation, the changing information landscape associated with the Internet of Things represents considerable change to the attack surface and threat environment of the Internet and Internet-connected systems.

The precise definition of the Internet of Things is a subject of some debate, due to the influence of several contributing trends, as well as various interpretations of the phrase in everything from scientific research to marketing materials [2]. For purposes of attack surface and threat analysis, let us confine our discussion to two component trends within the larger IoT landscape, namely ubiquitous network-connected technologies, and object-embedded information produced and consumed by those pervasive technologies.

The past decade has seen staggering growth in the number of devices that humans use to directly produce and consume network information. As of 2010, there were over 12.5 billion such devices on the Internet, up from 500 million in 2003, and we estimate that there will be 50 billion by 2020 [1].

However, there are also an increasing number of technologies that do not require human interaction to produce and consume network information. In 2020, we estimate that there will be over a trillion such systems.

Further, the number of objects that do not directly connect to the Internet, yet contain embedded information, is also on the rise. Much focus in the context of the Internet of Things has been placed on RFID tags, of which over 15 billion have been produced [3]. However, objects may also contain embedded information in the form of barcodes (representing over 5 billion machine-object interactions per day [4]), serial numbers, and other forms of machine-consumable object symbology, which are present on the vast majority of objects involved in commerce.

The Internet of Things is defined as much by its interconnectivity as by its comprising entities. Early attempts at understanding the relationship between entities of the IoT were focused on their statistical relationships. Using this approach, one might project a world population of 7.6 billion in 2020, and each person matched up with 6 connected devices, over 130 sensors, and innumerable embedded information objects. Simple statistical relationships, however, do not

reflect the actual distribution of objects and technology, or the dynamic nature of the interactions between IoT entities. Usman Haque has suggested that we think of the IoT in terms of environments, as opposed to objects or sensors [5]. In order to assess the threat implications of the IoT, we will first discuss the relevant surface characteristics of these environments, and their dynamic nature. What systems and information are present in this environment at this time? What interactions are possible between them? Then, we will consider the agency of those characteristics in the frequency and effects of various cyber attacks.

2. RELEVANT SURFACE CHARACTERISTICS

Comprehensive enumeration of the Internet of Things' characteristics, even in comparison with previous eras of network evolution, is beyond the scope of this document. Rather, we seek to identify those characteristics most likely to have agency in cyber attacks.

Manadhata and Wing have provided an attack surface metric that is applicable to specific software systems, but when we apply it to dynamic networks, we must necessarily accept less granular definition [6]. It's not likely that we would be able to assess the attack surface of each entity comprising a specific environment in time (at the very least, we're unlikely to have access to all of the necessary source code). We can, however, abstract Manadhata and Wing's concepts of channels, methods, and data items (collectively, resources), and apply them, without weight, to generic IoT environments in comparison with previous network environments. Relevance is denoted by material change in the number of system resources. Admittedly, this would be a crude metric for measuring the absolute attack surface of a specific environment, but this approach allows us to assess the relevance of IoT surface characteristics in general terms.

A. *POPULATION*

The first concern associated with an IoT environment is the population of entities. As previously discussed, the population of entities is expected to grow rapidly, as users embrace more connected devices, more sensors are deployed, and more objects are embedded with information. Each entity, depending on its type, carries with it an associated set of channels, methods, and data items, each of which is subject to potential abuse. This increased population has the effect of creating an explosion in the total number of potential target resources across the Internet, as well as within any specific environment.

B. COMPLEXITY AND COST

Each new entity can be classified into one of three tiers, defined by its characteristics, see Table I. Each tier inherits the characteristics of the lower tiers.

Table I. Classification Tiers

Tier	2020 Population	Examples	Characteristics
3	50 billion	Desktop, Laptop, Smartphone	Entity has channel(s) and method(s) for interaction with users
2	1 trillion	Sensor, Controller	Entity has channel(s) and method(s) for interaction with other entities
			Entity has channel(s) and method(s) for interacting with its environment
1	No estimate	Barcode, RFID	Entity contains data item(s) that may be consumed by other entities via an automated method
			Entity has a unique identity

These tiers represent the level of complexity inherent in the entities, as defined by their resources. As this table indicates, the anticipated population of entities is greatly skewed towards lower complexity entities. In the context of attack surface analysis, entities with a comparatively low complexity also have a comparatively small attack surface. There simply aren't that many channels, methods, and data items to consider for each entity, which is good for any specific low-complexity entity. However, when you take into consideration the massive population of tier 1 and 2 entities, the aggregate number of attack vectors is still daunting. Even a single attack vector for each tier two system, compared with 14 attack vectors for a tier 3 Linux system (based on Manadhata and Wing's estimate), still results in tier 2 systems presenting over 42% more attack vectors in aggregate than tier 3 systems.

Population and complexity also imply cost, and hence available compute and storage resources, as well as quality of components and materials. As we will later see, the balancing act between cost and resources has an important impact on the resources available for system security, encryption methods, key size and distribution, and software updates.

C. HETEROGENEITY AND INTEROPERABILITY

The number of distinct tier 3 system types has been increasing as a result of their pervasiveness, but the explosion of tier 1 and 2 entities also represents increased heterogeneity across the Internet of Things. However, heterogeneity may not hold true within a specific environment. A dam that is embedded with a network of

sensors to measure its integrity would be a fairly homogenous environment. So, though a dramatic increase of tier 1 and 2 entities increases the heterogeneity of the IoT in aggregate, specific environments may still be highly homogenous.

Given this anticipated heterogeneity across the IoT, we are due some further consideration of interoperability between entities within an environment, and across the IoT at large. While some have advocated the need for, and made some early progress towards, universal interoperability and open standards in the IoT, the extent to which it's possible is largely dependent on how – and how rapidly – the IoT evolves [7]. The National Intelligence Council (NIC) outlines four possible scenarios for this evolution: Fast Burn, Slowly But Surely, Connected Niches, and Ambient Interaction [8]. Of all of the scenarios, Slowly But Surely, which predicts pervasiveness in 2035, is the only scenario that permits universal interoperability. However, our projections for entity population growth and the vertical nature of extant stakeholders are much more indicative of the Connected Niches scenario, in which interoperability is challenged by reluctance of industries to cooperate. Interoperability struggles present a challenge to accountability and manageability. As the number of system stakeholders increases, accountability for preventing, identifying, and resolving security issues will be more distributed. Similarly, the channels and methods for interaction will grow more voluminous and complex.

D. MOBILITY AND DISTRIBUTION

The increase in mobile tier 3 entities, such as laptops and mobile phones, coupled with the increase in tier 1 and 2 entities, will result in more dynamic operating environments. Systems and data items will shift rapidly between environments. This exacerbates the challenges of establishing appropriate access control, monitoring, and automated decision-making within limited domains of visibility and control. However, mobile entities that do not maintain connectivity to the broader Internet will have a smaller window of compromise in any one environment.

One of the chief advantages of the Internet of Things is that you can deploy systems and information where people are not. The utility of such sensors, along with mobility, will cause the population of IoT entities to be more broadly distributed in physical space than previous networks. As we continue to drive down the relative cost and complexity of entities, we will see a related increase in population in previously sparse geographies.

3. CYBER ATTACK IMPLICATIONS

Changes to the operating landscape affected by the Internet of Things will necessarily result in changes to the nature of cyber attacks. The weapon actions that comprise a cyber attack are defined by their objectives [9]. Applegate provides a useful perspective on these objectives by defining cyber maneuvers as "the application of force to capture, disrupt, deny, degrade, destroy, or manipulate computing and information resources" [10]. Privilege escalation, for instance, is defined by the objective of capturing positional advantage. By loosely grouping the objectives of cyber maneuver, we can establish a structure in which we can assess the threat implications of the IoT.

A. CAPTURE

Capture attacks take two primary forms, depending on the targeted resources. Some capture attacks are designed to gain control of physical or logical systems, while others are designed to gain access to information. Attempts to capture systems are intended to gain a positional advantage that can be leveraged in subsequent operations. Attempts to capture information are intended to gain an exploitative intelligence advantage [10].

1. Systems

Systems composing the Internet of Things are uniquely susceptible to capture, due to a number of their characteristics. Their ubiquity and physical distribution afford attackers with greater opportunity to gain physical or logical proximity to targets.

Increased mobility and interoperability amplify the threat to IoT systems, in that they complicate access control by enabling an attacker to introduce compromised systems into the environment or remove systems in order to compromise and reintroduce them without detection. They also provide opportunity for attackers with a foothold in the environment to compromise transient systems in order to spread compromise to other environments. However, mobility may also dampen the threat by narrowing the window of opportunity to attack transient systems.

The heterogeneity of IoT systems is another factor in capture. Heterogeneity can complicate update and patch procedures to the point of increasing the window of vulnerability to a specific attack, but it may also limit threat propagation by requiring different weapon actions to successfully capture different systems, provided the vulnerability isn't found in the common channels and methods of interoperability.

2. Information

Information in the Internet of Things is widely distributed throughout component systems, so that any successful capture of a system will likely result in capture of information to which that system has access. Wide distribution of systems may also necessitate a longer chain and / or a denser mesh of communications, affording attackers greater opportunity to intercept or intercede in information transmission within the environment.

System resource limitations, particularly in tier 2 entities, may limit systems' access to robust encryption, while necessitating frequent, small bursts of information in a standard format. The expected asymmetry between a tier 2 system's encryption resources and the resources of, for instance, an attacker with a multi-core analysis system, aids in the attackers ability to capture information. Further, the frequency of these transmissions affords greater opportunity, and the standard format may aid in cryptanalysis. However, small burst size, combined with frequent key exchange, limits the amount of information that an attacker can capture with a given solution.

B. DISRUPT, DEGRADE, DENY, DESTROY

Disrupt, degrade, deny, and destroy attacks (hereinafter collectively referred to as disrupt attacks) differ from capture attacks, in that they are intended to confer a competitive disadvantage on the target, as opposed to conferring an advantage upon the attacker. When considering the threat of disruption, we must evaluate attacker opportunity, as well as target resistance, resiliency, and assurance.

Attackers seeking to disrupt systems in the Internet of Things share the opportunity advantages of system capture attackers, in that opportunity to capture a system also affords attackers the opportunity to disrupt it. However, disrupt attacks against information are slightly different, as opportunity to capture information does not imply opportunity to disrupt it, unless the attacker has captured either a single point of failure, or all requisite points of failure, for information storage and / or transmission.

The relative low cost and complexity of tier 1 and 2 entities in the IoT are directly related to the entities' resistance to disruption. Unless they exist within a hardened environment, we may assume that these entities are susceptible to physical abuse and tampering. If they are mobile entities, they are also susceptible to displacement.

The combination of heterogeneity and interoperability in IoT entities is key to resiliency. Heterogeneity is generally assumed to result in higher survivability for the network as a whole [11]. In the event of disruption of one entity in the environment,

other entities may resist the attack, and be able to continue functioning. Provided that the participating entities are interconnected and able to route information using a standard set of protocols, the network gains greater transmission resiliency, as well. However, given the current Connected Niches mode of IoT evolution, it's unlikely that we'll have our cake and eat it too, with regards to heterogeneity and interoperability within any specific environment.

Assurance is the environment operators' ability to determine that a disruption has occurred and then perform incident management. The challenge is to verify confidentiality, integrity, and availability of all systems and data within the environment. Assurance in the IoT is significantly complicated by entity mobility and the number of stakeholders implied by interoperability challenges.

C. MANIPULATE

Manipulate attacks, as distinct from capture and disrupt attacks, are intended to influence opponents' decision cycles [10]. Using Boyd's OODA loop construct as a reference for general decision cycles, we can determine several different forms of manipulate attack within the context of the Internet of Things [12].

At the earliest point in the cycle, an attacker may manipulate the outside information itself. This involves intercession at the entry point in the information collection process, usually via physical means. Outside information manipulation may be something as simple as local environmental manipulation (e.g., heating the environment around a temperature sensor) and analog data manipulation (e.g., modifying a document prior to OCR), or it may be as complex as World War II's Operation Fortitude. Similarly, manipulate attacks may involve manipulating embedded data, whether by physically replacing or modifying tagging information, or infecting a portable data store, as in the events that lead to Operation Buckshot Yankee.

Further into the decision cycle, an attacker may directly manipulate sensors that gather information. As opposed to feeding a sensor manipulated information from its environment, the attack would, in this case, use a compromised sensor manipulate information available to other entities. This same approach applies to manipulation of controllers to change their actions, so that sensors observing the results of the controllers' actions would receive information that is not reflective of an undisturbed closed loop.

The last common form of manipulate attack is manipulation of the feed-forward mechanisms in the decision cycle, through employment of a man-in-the-middle or spoof attack. In this case, the attacker intercedes in the communications between entities, in order to exert control over information transmission.

It's clear that, as with the other types of attacks we've considered, the large population of entities in the IoT presents opportunity for a manipulate attacker, but this is even truer when we consider potential communications interoperability. Due to the network effect, each additional interoperable entity that is added to the network greatly increases the possible intercommunications, and affords greater opportunity for a man-in-the-middle attack. Mobility and distribution in the IoT also increase opportunity for attack, as they make it easier to manipulate entities without fear of detection. Manipulate attacks also present the same assurance challenges that disrupt attacks do, and in that sense, mobility and number of stakeholders also apply here.

4. PRIVACY CONCERNS IN THE INTERNET OF THINGS

The smart, connected objects that will densely populate the Internet of Things will interact with both humans and the human environment by providing, processing, and delivering all sorts of information or commands. These connected things will be able to communicate information about individuals and objects, their state, and their surroundings, and can be used remotely. All of this connectivity carries with it a risk to privacy and information leakage.

A significant body of work has explored privacy issues in ubiquitous computing systems and much of that research is applicable to the Internet of Things. Establishing meaningful identity, using trusted communication paths, and protecting contextual information is all very important to ensure the protection of user privacy in this environment. We will touch briefly on each of these issues as part of the exploration of threats within the Internet of Things.

Beresford and Stajano [13] have explored anonymous communication techniques and the use of pseudonyms to protect user privacy while also working on metrics to assess user anonymity. Their work takes a novel approach by hiding identity from the applications that utilize it in order to better protect the user consuming those services.

In their work on Decentralized Trust Management, Zhao et al [14] propose new technologies that enable the bootstrapping of trust, and subsequently, the calculation of trust metrics that are better suited to mobile, ad-hoc networks. In their model, every member of a community (users, devices, sensors, etc.) can serve as an authority to enroll and authenticate other entities for the community. Their model showcases the inherent problems with establishing trust in ad-hoc networks like those in the IoT where new sensors, services, and users are constantly introduced and asked to share data.

Finally, applications in the IoT, which will be enabled by a ubiquitous computing and communications infrastructure, will provide unobtrusive access to important contextual information as it pertains to users and their environment. Clearly, the successful deployment of such applications will depend on our ability to secure them and the contextual data that they share.

One example of sensitive contextual information is location. When location-aware systems track users automatically, an enormous amount of potentially sensitive information is generated and made available. Privacy of location information is about both controlling access to the information and providing the appropriate level of granularity to individual requestors. The Location Services Handbook [15] explores a variety of location-sensing technologies for cellular networks and the coverage quality and privacy protections that come with each.

5. CONCLUSIONS

The Internet of Things continues to march forward apace, and will accelerate over the coming years. We will see the Internet change in many important ways, and in the context of threat analysis, we will need to continue to explore the impact of these changes on the attack surface of the Internet as a whole, as well as specific environments.

Growth in network-capable and consumable entities is the largest potential concern with regards to potential attack surface, as we anticipate an explosive increase in both the breadth and density of the global information environment. Many of these new entities will be fairly unsophisticated in comparison to today's network-connected devices, as increased deployment of tier 1 and 2 devices outpaces miniaturization and cost reduction trends, resulting in entities with constrained security resources. They will be quite diverse in their designs and functions, and it's unlikely that they will broadly interoperate, creating some considerable monitoring and management challenges. They will be increasingly mobile and distributed, meaning that many contemporary security processes and tools that rely on information density will need to change considerably.

Attackers will find that the characteristics of the IoT in general embody an accelerating shift from the relatively controlled technology world that they know today to a world of increasing opportunities. Attackers seeking to capture systems and information will find a broad spectrum of targets from which to choose, and when their objectives require capture of any system, as opposed to a specific system, in an environment, they will likely have a broader set of tools to achieve their goals. Attackers seeking to disrupt IoT systems and environments will likewise

identify new opportunities and approaches to achieve their ends, with their only new concern being potential confluence of heterogeneity and interoperability – an unlikely result. Perhaps the greatest opportunity will be for attackers seeking to manipulate IoT entities, as they take advantage of a broad, dynamic network with exponential channels of communication.

The Internet of Things will bring many great new advances, including whole new ways of thinking about and interacting with our world. However, with those opportunities come many challenges in the world of information security, and we will need to continue to research and develop new approaches to ensuring our safety, security, and privacy.

REFERENCES

[1] Evans, Dave. «The Internet of Things How the Next Evolution of the Internet Is Changing Everything.» *CISCO white paper* (2011).

[2] Uckelmann, Dieter, Mark Harrison, and Florian Michahelles. «An architectural approach towards the future internet of things.» *Architecting the Internet of Things* (2011): 1-24.

[3] Harrop, P., and Raghu Das. «RFID Forecasts, Players and Opportunities 2012-2022.» *IDTechEx, Cambridge, UK* (2012).

[4] Varchaver, Nicholas. «Scanning the globe.» *Fortune Magazine, available at: http:// money. cnn. com/magazines/fortune/fortune_archive/2004-05/31/370719/index. htm* (2004).

[5] Tish Shute, *Pachube, Patching the Planet: Interview with Usman Haque*.: UgoTrade, 2009.

[6] Manadhata, Pratyusa K., and Jeannette M. Wing. «An attack surface metric.» *Software Engineering, IEEE Transactions on* 37.3 (2011): 371-386.

[7] INFSO EU, *Internet of Things in 2020: Roadmap for the Future*.: INFSO EU, 2008.

[8] National Intelligence Council (NIC), *Disruptive Civil Technologies: Six Technologies With Potential Impacts on US Interests Out to 2025*., Official US Government Document, Accession Number ADA519715 (2008).

[9] Cartwright, General James W. «Joint Terminology for Cyberspace Operations.» *Joint Chiefs of Staff (JCS) Memorandum, Nov* (2010).

[10] Scott D. Applegate, «The Principle of Maneuver in Cyber Operations,» in *2012 4th International Conference on Cyber Conflict*, vol. 4, Talinn, Estonia, 2012, p. 13.

[11] Zhang, Yongguang, Harrick Vin, Lorenzo Alvisi, Wenke Lee, and Son K. Dao, «Heterogeneous Networking: A New Survivability Paradigm.» *Proceedings of the 2001 workshop on New security paradigms*. ACM, 2001.

[12] Boyd, John R. «The essence of winning and losing.» *Unpublished lecture notes* (1996).

[13] Beresford, Alastair R., and Frank Stajano. «Location privacy in pervasive computing.» *Pervasive Computing, IEEE* 2.1 (2003): 46-55.

[14] Meiyuan Zhao, Hong Li, Rita Wouhaybi, Jesse Walker, Vic Lortz, and Michael J. Covington. Decentralized trust management for securing community networks. Intel Technology Journal, 13(2):148-169, 2009. Invited Article.

[15] Eladio Martin, Ling Liu, Michael Covington, Peter Pesti, and Matt Weber. Chapter 1: Positioning technology in location-based services. In Syed A. Ahson and Mohammad Ilyas, editors, Location Based Services Handbook: Applications, Technologies, and Security. CRC Press, July 2010.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Cyber Deception and Autonomous Attack – Is There a Legal Problem?

William Boothby
Royal Air Force (Ret.)
United Kingdom

Abstract: The publication of the Tallinn Manual on the Law of Cyber Warfare is a huge step forward and now States must decide whether to adopt, formally or otherwise, the rules and guidance it provides. A discussion of deception operations in the cyber age reveals some of the challenges we face in simply transposing existing law of armed conflict rules into cyber terms. Deception operations in warfare are nothing new; some are lawful, and some are not, but does a person have to be deceived for an act that otherwise breaches article 37(1) to be perfidy? How does the law address the improper use of protective indicators and, indeed, espionage in the cyber context? And then we have the crunch question. If cyber deception operations become pervasive so that little or no reliance can be placed, say, on targeting data, what implications does this have for the ability of combatants to comply with distinction, discrimination, proportionality and precautions rules, and does that matter?

Keywords: *Law of Armed Conflict, cyber deception, autonomous attack*

1. INTRODUCTION

The publication this year of the Tallinn Manual[1] has done much to clarify the law on the offensive and defensive use of cyber capabilities in periods of armed conflict. Many matters that were being extensively debated in the literature have been subjected to the critical analysis of the International Group of Experts assembled by the Cooperative Cyber Defence Centre of Excellence in Tallinn. The Experts included *jus in bello* issues in their deliberations and the Manual therefore addresses the rules that regulate the use of cyber force during both international and non-international armed conflicts. While it will be for states to decide whether the conclusions reached by the Experts should guide their warlike activities in the future, there is no doubt that the Manual will at the very least inform the views of States in that regard.

The Experts reached the general conclusion that the law of armed conflict does apply to military cyber operations during and in connection with armed conflicts. Specifically, they reached the clear consensus that the principles of distinction, discrimination, proportionality and the precautionary rules so apply.[2] They also concluded that the rules as to perfidy and ruses of war apply broadly speaking as written in API.[3]

Until a generality of State practice has made the position of States in general clear on particular issues, it will be premature to talk of clear customary law on these cyber warfare issues. Rather what the Manual is doing is to take legal rules that are clearly customary in nature and determine whether there is any apparent reason why they should not apply in cyberspace. The Rules that appear in the Manual are those which, by consensus, the International Group of Experts found to apply in cyberspace as a matter of customary law. The outstanding issue is therefore whether States agree with that interpretation of the Experts.

Until the position of States in that respect becomes clear through practice over coming years and decades, it is sensible to discuss particular issues relating to the conduct of military cyber operations by reference to the black letter Rules and associated Commentaries set forth in the Manual. Those Rules and Commentaries will of course be a valuable resource to States and will assist them to identify perceived gaps in the legal architecture and to determine whether new law is required and, if so, what form it should take. Nevertheless, where there was previously an absence

1 Tallinn Manual on the Law of Cyber Warfare, CUP, January 2013.

2 See for example Rules 31, 32, 37 and 49 to 59.

3 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, Geneva, 8 June 1977.

of conventional law specifically addressing these subjects, there is now a document asserting contemporary customary and thus universally applicable Rules, and that represents significant progress in this field.

One of the challenges in drafting the Manual was to consider the fundamental differences between cyber activities and more traditional methods of warfare.[4] A recurring issue was to determine whether the existing traditional rules make sense when applied to the cyber domain. Can the peculiarities of cyber operations be accommodated to otherwise well accepted legal rules? To illustrate, and examine the challenges associated with, this issue it is the purpose of the present article to look at one particular aspect of the law on the conduct of cyber hostilities and to consider in some detail the difficulties that the characteristics of cyber operations can be expected to pose.

2. THE ISSUE

Deception is an established and inherently lawful method of undertaking military operations. Perhaps, one of the best known, classical deception operations is Virgil's account in the Aeneid of the use of a wooden model of a horse to infiltrate a Greek unit into the city of Troy after ten years of siege. Another example was the World War Two Operation Mincemeat aimed at convincing the German High Command that the allies would attack Sardinia and Greece in 1943 instead of Sicily, when the means used was the planting of a dead body with false papers concealed upon it to that effect.

So if deception operations are nothing new, we should consider some of the different types of deception operation that have been undertaken in the past. Spaight refers to the use during World War One on occasion of false nationality marks on aircraft. "The inadmissibility of the use of such marks was established, first, by the accusations which the belligerents made against one another of resorting to the practice, secondly by their indignant denials of any complaints that they had done so themselves."[5] On the other hand merely simulating death to avoid being attacked and to permit later escape from a difficult tactical situation has long been seen as legitimate.[6] Spaight also reports the different but relevant and similarly legitimate case of Lieutenant L G Hawker who was seeking to attack a German airship shed at Gontrode in April 1915. It appears that he used "an occupied German

[4] Throughout the period of the project to produce the Manual, the author was a member of the Group of Experts.

[5] J M Spaight, Air Power and War Rights, 3rd Edn (1947) 85-6.

[6] Spaight, ibid at page 173.

captive balloon to shield him from fire whilst manoeuvring to drop the bombs".[7] Note also the use of dummy communications to mislead the enemy to believe that fighter aircraft are active when this is not in fact the case[8], also a legitimate practice. Interestingly, Spaight then discusses the legitimacy of tactics during World War One in which an aircraft would simulate landing signals of the enemy's military aircraft in order to enable it to get close to the enemy airfield before dropping its bombs. He concludes that such tactics were lawful because the machine must have been either friend or foe and, in either case, a combat aircraft.[9]

So as can be seen, these deception operations, although they use a considerable variety of techniques, were all aimed at causing the enemy to misunderstand for example the military posture, the identity, the intentions, the manpower capabilities, the resources or the ultimate objectives of the party to the conflict using the deception.

All the indications are that cyber military operations will employ deception to a very considerable degree. Some cyber operations will be so constructed as to appear not to have been undertaken by the State that was in fact responsible. Indeed, in some cases the State undertaking the cyber operation will make it appear that some other State is responsible. In other cases, the cyber operation may be so undertaken as to conceal the very fact of the operation from the enemy.[10] More routinely, damaging cyber packages can be initiated from one computer but may appear to have been sent from an entirely different computer. It may be made to appear that a particular person is the author of a cyber operation when in fact another person originated it. Even if the author of the operation can be identified, false information may be put out to the effect that the originator is acting on behalf of one State or organization when in fact he or she is acting on behalf of another.

Of course these are only examples of the sorts of deception that may be undertaken and it should be borne in mind that multiple deceptions may be used, with the probable intent either that it shall remain permanently unclear who was responsible for a particular event, or that it shall be clear and widely accepted that State or organization A was responsible for it when in fact entity B was in fact answerable. This immediately raises questions over the acceptability of automatic responses to cyber operations. The automatic response may target the computer, system or network where the initial operation appeared to have initiated when, in fact,

---

7 Spaight, ibid, at page 174 citing London Gazette, 8 May 1915.

8 Spaight, ibid, at pages 176-8 cites numerous examples of such ruses in both World Wars.

9 Spaight, ibid at page 179.

10 Consider for example the manner in which the Stuxnet weapon concealed the effect it was having on the Iranian centrifuges from those responsible for monitoring the relevant indicators, thus making it appear that everything was operating normally.

some other computer, system or network was in fact its source. Causing damage to unwilling conduits in this way is likely to prove unacceptable, and risks expanding the scope of conflicts and causing unwanted casualties and damage.

A further question to consider, and the central topic of this paper, is therefore whether this increasing prevalence of military deception that we can foresee as a feature of future military cyber operations is consistent with current interpretations of the law or whether it challenges those interpretations.

3. THE BACKGROUND TO THE CURRENT LAW

We should start our consideration of the law by referring to the Lieber Code. Dr Francis Lieber's text[11] does not have the status of a source of the law[12] but it does indicate what legal thinking was in the mid-nineteenth century on these and related issues. The Lieber Code stipulates that "[m]iliary necessity admits of ….. obstruction of the ways and channels of … communication…. And of such deception as does not involve the breaking of good faith either positively pledged, regarding agreements entered into during the war, or supposed by the modern law of war to exist. Men who take up arms against one another in public war do not cease on this account to be moral beings, responsible to one another and to God."[13] He further noted that military necessity "admits of deception, but disclaims acts of perfidy"[14], a distinction which, as we shall see, lies at the core of the current law.

Dr Lieber included in his text particular provision relating to another form of deception, namely the misuse of a flag of truce. He asserted that if such abuse takes place "for surreptitiously obtaining military knowledge, the bearer of the flag thus abusing his sacred character is deemed a spy" and he goes on to emphasise how necessary the sacred character of the flag of truce is and that its abuse is a heinous crime.[15]

---

[11] Instructions for the Government of Armies of the United States in the Field, 24 April 1863, prepared by Professor Francis Lieber.

[12] This is because the text as such is not one of the law's fundamental principles, nor is it customary law per se and it does not have treaty status; see Statute of the International Court of Justice, article 38.

[13] Lieber Code, article 15.

[14] Lieber Code, article 16.

[15] Lieber Code, article 114. See also Brussels Declaration, article 45, where a parlementaire loses his rights of inviolability if it is shown that he has taken advantage of "his privileged position to provoke or commit an act of treason."

The authors of the Brussels Declaration[16] found a principle of law that has since come to be accepted as one of its cornerstones, namely that the "laws of war do not recognize in belligerents an unlimited power in the adoption of means of injuring the enemy".[17] Applying this principle, they found to be especially forbidden "murder by treachery of individuals belonging to the hostile nation or army"[18] and "making improper use of a flag of truce, of the national flag or of the military insignia and uniform of the enemy, as well as the distinctive badges of the Geneva Convention".[19] The Declaration drew an important distinction between such activities, however, and lawful deception by providing that "ruses of war and the employment of measures necessary for obtaining information about the enemy and the country….are considered permissible".[20]

The Oxford Manual [21] also does not have the status of a source of the law. However, it was written in 1880 by acknowledged experts of the time and it is therefore useful to note that it contained some similar provisions to those in the Brussels Declaration of six years earlier. Article 4 repeats that the means of injuring the enemy are not unlimited, and specifically prohibits perfidious and unjust acts. This Manual requires that conventions, or agreements, between the parties during the conflict must be "scrupulously observed and respected"[22] and that it is forbidden "to make treacherous attempts upon the life of an enemy, as for example by keeping assassins in pay or by feigning to surrender", "to attack an enemy while concealing the distinctive signs of an armed force" or "to make improper use of the national flag, military insignia or uniform of the enemy, of the flag of truce and of the protective signs prescribed by the Geneva Convention"[23].

By the time of the Hague Peace Conferences of 1899 and 1907, thinking and terminology had clarified further. The negotiators included in their texts the general customary admonition that "the right of the belligerents to adopt means of injuring the enemy is not unlimited."[24] More specifically, the Hague Regulations,

---

16 Project of an International Declaration concerning the Laws and Customs of War, Brussels, 27 August 1874.

17 Brussels Declaration, article 12. For the modern formulation see API, article 35(1).

18 Brussels Declaration, article 13(b).

19 Brussels Declaration, article 13(f).

20 Brussels Declaration, article 14, which refers to an exception that the civilian population cannot be forced to take part in military operations against their own country.

21 The Laws of War on Land, Oxford, 9 September 1880

22 Oxford Manual, article 5.

23 Oxford Manual, article 8(b) to (d).

24 For example, Annex to Hague Convention IV Respecting the Laws and Customs of War on Land, art. 22, The Hague, 18 October 1907. Similarly but not identically expressed regulations had been annexed to the Hague Convention II of 1899 but the 1899 text was superseded by the 1907 text and it is therefore on the latter that we will rely.

which have both treaty law and customary status, especially prohibit "kill[ing] or wound[ing] treacherously individuals belonging to the hostile nation or army".[25] Equally importantly, the Regulations made specific provision as to ruses of war, so the distinction that we are discussing in the present article was already embedded in international law in 1899 and 1907. Thus, article 24 of the 1907 text states: "Ruses of war and the employment of measures necessary for obtaining information about the enemy and the country are considered permissible." So an important distinction was made from the outset between acts deceiving the enemy as to matters of protection and lawful ruses and espionage, the latter being accepted as measures in warfare that do not breach the law of war, or in more modern parlance, the law of armed conflict.

4. THE MODERN LAW OF PERFIDY AND RUSES IN API AND THE TALLINN MANUAL

The modern law is to be found in API, article 37. It should be explained at the outset that for the purposes of the present discussion the important distinction is between paragraphs (1) and (2) of that Article, which are as follows:

"(1) It is prohibited to kill, injure or capture an adversary by resort to perfidy. Acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence, shall constitute perfidy. The following acts are examples of perfidy:

(a) the feigning of an intent to negotiate under a flag of truce or of a surrender;

(b) the feigning of an incapacitation by wounds or sickness;

(c) the feigning of civilian, non-combatant status; and

(d) the feigning of protected status by the use of signs, emblems or uniforms of the United Nations or of neutral or other States not Parties to the conflict.

(2) Ruses of war are not prohibited. Such ruses are acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under the law. The following are examples of such ruses: the use of camouflage, decoys, mock operations and misinformation."

[25] Note that the word 'treachery' means for all practical purposes the same as the word 'perfidy' as used in the following discussion of the modern law; Tallinn Manual, paragraph 1 of the commentary accompanying rule 60.

Both of these paragraphs are, it must be appreciated, concerned with deception operations. However, some such operations are lawful under paragraph (2) whereas others are rendered unlawful by paragraph (1). It is therefore important to identify the critical points of difference between the two classes of operation. Both classes of operation invite the confidence of the enemy in relation to matters which are in fact untrue. Both classes of operation are aimed at persuading the enemy either to act or to refrain from acting on the basis of that induced false appreciation of the facts.

The critical point of difference is the nature of the false belief that is being induced in the enemy by the operation. In the second, lawful class of deception operations, the deception does not breach a rule of the law of armed conflict and is not inviting confidence "with respect to protection under the law". In the first, unlawful class of deception operation, the deception is directed at inducing the adverse party to believe that there is either a right to, or a duty to give, protection under the law. AMW notes that a "typical example of perfidy would be to open fire upon an unsuspecting enemy after having displayed the flag of truce, thereby inducing the enemy to lower his guard".[26]

The Tallinn Manual expresses the perfidy rule in terms that are very similar to article 37. The significant point of difference is that the Manual's rule[27] omits reference to capture as a result of the perfidy, simply because the customary rule does not extend to capture, unlike the rule in article 37. [28] It should be emphasized that an act of perfidy that does not result in death, injury or capture does not breach either rule. The Tallinn Manual makes the useful point that the person deceived need not necessarily be the same person as the one whose death or injury results provided that the person who is killed or injured is in fact the intended target of the attack.[29] However, the perfidy must be the proximate cause of the death or injury. While there may be a time delay between the two events, it is causal proximity that is relevant here.[30]

---

[26] AMW, commentary accompanying Rule 111(a), paragraph 8.

[27] Tallinn Manual, rule 60.

[28] See paragraph 2 of the Commentary associated with Rule 60, which notes that Hague Regulations article 23(b) makes no mention of capture. It is noted there that the corresponding war crime under the Rome Statute of the International Criminal Court, 1998, namely article 8(2)(b)(xi) in relation to international armed conflicts, also makes no mention of capture resulting from perfidy. The corresponding war crime in the Rome Statute that arises in relation to non-international armed conflicts is in article 8(2)(e)(ix) which refers to "[k]illing or wounding treacherously a combatant adversary". Notwithstanding the omission of capture from the Rome Statute war crimes, the ICRC contends that the customary law of armed conflict rule includes capture; ICRC Customary Humanitarian Law Study, Rule 65 and commentary at page 225 where it is suggested that the consequences of capture may not be grave enough to constitute the act of perfidy a war crime.

[29] Tallinn Manual, commentary accompanying Rule 60, paragraph 4.

[30] Tallinn Manual, commentary accompanying Rule 60, paragraph 6.

The Group of Experts then considered a situation which brings us rather closer to the topic of the present paper. They considered whether a person has to be deceived for the perfidy rule to be broken or whether the rule extends to deception of a machine. The example referred to in the Manual is that of a cyber deception operation that targets a pacemaker fitted, for example, to an enemy commander, causing the pacemaker to malfunction thus killing the commander. If the cyber operation betrays the confidence of the computer controlling the pacemaker, a majority of the Experts concluded that the perfidy rule is broken. The minority view was that for perfidy to be made out, the deception must operate on a human mind in the prohibited way.[31]

5. THE MODERN LAW OF PERFIDY AND RUSES IN OTHER MANUALS

The Air and Missile Manual[32] finds a rule expressed in similar terms to article 37(1) of API.[33] Importantly, perfidious action that results in damage but not in death, injury or capture does not constitute a breach of the law of armed conflict and, by extension, does not amount to a war crime.[34]

AMW then notes, most importantly in relation to the current discussion, that the mere fact "that a person is fighting in civilian clothing does not constitute perfidy"[35] although the same Manual notes that the person fighting in this way may not be entitled to combatant immunity and may thus be prosecuted and punished under domestic law.[36]

AMW also finds a rule as to ruses of war the effect of which largely reflects the rule as expressed in article 37(2) of API, but which employs somewhat different language.[37] It notes that the fact that the ruse results in death, injury or capture of personnel of the adverse party does not per se cause the attack to be prohibited as

[31] See Tallinn Manual, commentary accompanying Rule 60, paragraph 9.

[32] Program on Humanitarian Policy and Conflict Research, Harvard University, Manual on International Law Applicable to Air and Missile Warfare, published with a commentary in March 2010 and referred to collectively here as 'AMW'.

[33] AMW, Rule 111(a) and (b). Note also the US Commanders' Handbook on the Law of Naval Operations, NWP 1-14, paragraph 12.12 states a rule in similar language.

[34] AMW, commentary accompanying Rule 111(a), paragraph 7.

[35] AMW, commentary accompanying Rule 111(b), paragraph 4.

[36] The same paragraph of the AMW Commentary cites a useful example of perfidy, where the individual advances to an advantageous position "under the cover of being a civilian in order to fire on, and kill or injure, an unsuspecting enemy".

[37] The differences in language do not seem to produce significant difference in intended meaning.

perfidy provided that deception as to protected status is not involved.[38] It suggests as examples of lawful ruses the following activities in air warfare, namely "(a) mock operations[39]; (b) disinformation[40]; (c) false military codes and false electronic, optical or acoustic means to deceive the enemy (provided that they do not consist of distress signals, do not include protected codes, and do not convey the wrong impression of surrender)[41]; (d) use of decoys and dummy-construction of aircraft and hangars; and (e) use of camouflage".[42]

The same Manual also gives examples of air operations that would constitute perfidious conduct. The listed examples are "(a) the feigning of the status of a protected medical aircraft, in particular by the use of the distinctive emblem or other means of identification reserved for medical aircraft; (b) the feigning of the status of a civilian aircraft; (c) the feigning of the status of a neutral aircraft; (d) the feigning of another protected status; and (e) the feigning of surrender."[43]

Highly significantly from the perspective of the present text, whether or not such behavior is perfidious, AMW finds the following conduct is always prohibited, namely improper use by aircraft of distress codes, signals or frequencies and use of any aircraft which is not a military aircraft as a means of attack.[44] Improper use in this regard means any use outside normal purposes. So, for example, distress signals must be reserved for their humanitarian purposes,[45] and any military use of such signals that is outside the scope of humanitarian activity and which is, say, aimed at facilitating the undertaking of an attack, would be prohibited by the Rule. There is a fine distinction to be considered here. Thus, if a pilot of an aircraft sends a false distress signal that will clearly breach the Rule. If, however, the same pilot refrains from sending such a signal, but so flies his aircraft as to cause those on the

[38] AMW, commentary accompanying Rule 113, paragraph 3.

[39] AMW cites as examples air attacks on the Pas de Calais during the weeks leading up to D-Day in 1944 or the movement of, e.g., an aircraft carrier to create a false impression as to the likely nature of an attack. Similarly, simulated air attacks may be undertaken, as lawful ruses of war, as a device to persuade the enemy to activate its air defences and thus provide valuable targeting information. The common theme here is the presentation to the enemy of a false picture of what is occurring.

[40] AMW gives as an example an attempt to induce the enemy to surrender by creating the false impression that he is surrounded, or that an overwhelming attack is about to occur; AMW, commentary accompanying Rule 116(b), paragraph 2, where the distinction is noted between such lawful activities and the use of false information as to civilian, neutral or other protected status which would not be lawful; ibid, paragraph 3.

[41] The use of enemy IFF codes, or the use of the enemy's password to avoid being attacked when summoned by an enemy sentry or inducing a false return on the enemy radar screen indicating the approach of a larger force than is the case are all cited in AMW as lawful ruses; commentary accompanying Rule 116(c), paragraphs 2 and 3.

[42] AMW, Rule 116.

[43] AMW, Rule 114.

[44] AMW, Rule 115. Distress codes signals and frequencies do not for these purposes include IFF codes; AMW commentary accompanying Rule 115(a), paragraph 5.

[45] AMW, commentary accompanying Rule 115(a), paragraph 1.

ground to form the incorrect view that the aircraft has been damaged, that would not breach the rule.[46] Contrast the circumstance discussed in paragraph 4 of the same commentary, namely where the pilot of an aircraft simulates a situation of distress with the purpose of creating the false impression that personnel deploying from the aircraft by parachute are entitled to protection under article 42 of API.[47] In these circumstances, if the deploying personnel are in fact paratroopers "this could amount to prohibited perfidy if it leads to the killing, injuring (or capturing) of an adversary."[48]

The UK Manual gives the following examples of ruses: "transmitting bogus signal messages and sending bogus despatches and newspapers with a view to their being intercepted by the enemy; making use of the enemy's signals, passwords, radio code signs, and words of command; conducting a false military exercise on the radio while substantial troop movements are taking place on the ground; pretending to communicate with troops or reinforcements which do not exist…; and giving false ground signals to enable airborne personnel or supplies to be dropped in a hostile area, or to induce aircraft to land in a hostile area".[49]

6. IMPROPER USE OF CERTAIN INDICATORS

The other deception-related provisions of API that we will discuss in the present paper are to be found in articles 38 and 39 and relate to the misuse of the emblems specified in those Articles. Thus, article 38(1) prohibits making "improper use" of the red cross or red crescent[50] or of other emblems, signs or signals provided for in the Conventions or in the Protocol and further prohibits the deliberate misuse of other internationally recognized protective emblems, signs or signals.[51] Importantly, Article 9 of Annex I to API, as amended on 30 November 1993, addresses means of electronic identification of medical transports.

---

[46] Note in this regard that a damaged aircraft is not necessarily a disabled aircraft, neither is it necessarily a surrendering aircraft; consider the discussion at AMW, commentary accompanying Rule 115(a), paragraph 3.

[47] "No person parachuting from an aircraft in distress shall be made the object of attack during his descent."; API, art. 42(1).

[48] AMW, commentary accompanying Rule 115(a), paragraph 4 and note API, art. 42(3): "Airborne troops are not protected by this Article."

[49] U.K. Manual, para. 5.17.2.

[50] Additional Protocol III to the Geneva Convention, article 2(1), applies the same prohibition to the Red Crystal adopted by that Instrument also as a distinctive emblem.

[51] As the Tallinn Manual notes at paragraph 2 of the Commentary accompanying Rule 62, this would extend to the distinctive sign for cultural property, for civil defence, the flag of truce and the electronic protective markings set out in Annex I to API; Cultural Property Convention, articles 16 and 17, API, art. 66, Hague Regulations, art. 23(f) and API, Annex I, paragraph 9. See also AMW, Rule 112(a) and (b).

As the Tallinn Manual makes clear, these are absolute provisions that do not require death, injury or capture as an essential ingredient of a breach while the term 'improper use' is considered to comprise any use other than that for which the emblem, sign or signal was intended.[52] Accordingly, this and the following examples of improper use of emblems etc are prohibited irrespective of whether the acts concerned also amount to perfidy.[53] It will be noted from paragraphs 6 and 7 of the commentary accompanying Rule 62 in the Tallinn Manual that the Group of Experts was divided as to whether the rule is specifically restricted to misuse of the emblem, sign or signal as such or whether misuse of a domain name such as 'icrc.org' to like effect would also be prohibited. The author takes the provisional view that the former interpretation is *lex lata* while the latter view may reflect *lex ferenda*.

While the focus in those provisions is on 'improper use', in article 38(2) "[i]t is prohibited to make use of the distinctive emblem of the United Nations, except as authorised by that Organization".[54] While it is clear that the prohibition will extend to unauthorized use of the emblem by electronic means, the same division of opinion as described in the previous paragraph applies to whether breach of the rule requires use of the emblem as such.[55]

Any use of flags, insignia or military emblems of the enemy is prohibited "while engaging in attacks or in order to shield, favour, protect or impede military operations".[56] The Tallinn Manual adds the words "while visible to the enemy" to the rule, to reflect the majority view among the experts that "it is only when the attacker's use is apparent to the enemy that the act benefits the attacker or places its opponent at a disadvantage".[57] However, where the use of the enemy's emblem in cyber communications is concerned, the Tallinn Manual is explicit, opining "it is permissible to feign enemy authorship of a cyber communication", basing this view on State practice regarding lawful ruses.[58]

---

[52] Tallinn Manual, commentary accompanying Rule 62, paragraphs 3 and 4, citing in the latter instance the ICRC Study, commentary accompanying Rule 61.

[53] AMW, chapeau to Rule 112 and commentary accompanying Rule 111(a), paragraph 9.

[54] For the application of this rule in cyber operations, see Tallinn Manual, Rule 63, citing NWP 1-14, paragraph 12.4, the UK Manual paragraph 5.10.c and the AMW Manual, Rule 112(e).

[55] See commentary accompanying Rule 63, Tallinn Manual. It will be appreciated that if the United Nations becomes a party to an armed conflict, its military personnel who are combatants and the objects it uses to make an effective contribution to the hostilities will be lawful targets. Misuse of its emblem by an adverse Party to such a conflict would, in those circumstances, amount to improper use of an enemy emblem as opposed to misuse of the United Nations emblem; AMW, commentary accompanying Rule 112(e), paragraph 3.

[56] API, article 39(2), AMW Rule 112(c) and Tallinn Manual, Rule 64.

[57] Tallinn Manual, commentary accompanying Rule 64, paragraph 4.

[58] Citing the extract from the UK Manual noted earlier in the present paper.

Article 39(1) of API prohibits making use of "the flags or military emblems, insignia or uniforms of neutral or other States not Parties to the conflict" and the Tallinn Manual finds, subject to the traditional rules of naval warfare,[59] a customary rule expressed in identical terms.[60] Any such use is unlawful, so the word 'improper' is not included in Article 39(1) nor in the corresponding Rule in the Manuals. There was however, as the Tallinn Manual explains, division among the Experts as to whether the use of other indicators, such as the domain name of the neutral's Ministry of Defence, would constitute a breach of the rule in circumstances in which the emblem as such is not employed.[61]

7. ESPIONAGE

As AMW notes, "espionage consists of activities by spies", adding, perhaps rather more usefully, that "a spy is any person who, acting clandestinely or on false pretences, obtains or endeavours to obtain information of military value in territory controlled by the enemy, with the intention of communicating it to the opposing Party."[62] Rule 66(a)of the Tallinn Manual makes it clear that cyber espionage and other forms of intelligence gathering directed at an adverse party to the conflict do not breach the law of armed conflict.[63]

The Tallinn Manual describes as 'clandestine' acts that are undertaken secretly or secretively, whereas the term 'under false pretenses' refers to acts so conducted as to create the impression that the individual has the right to access the information concerned.[64] Importantly, a person who obtains information about an adversary while the information gatherer is located outside enemy controlled territory is

[59] See API, art. 39(3).

[60] Tallinn Manual, Rule 65; see also AMW, Rule 112(d).

[61] Tallinn Manual, commentary accompanying Rule 65, paragraph 4.

[62] AMW, Rule 118. Article 29 of the Hague Regulations of 1899 and 1907 provided: "An individual can only be considered a spy if, acting clandestinely or, on false pretenses, he obtains, or seeks to obtain information in the zone of operations of a belligerent, with the intention of communicating it to the hostile party. Thus, soldiers not wearing a disguise who have penetrated into the zone of operations of the hostile army, for the purpose of obtaining information, are not considered spies…." Note the Lieber Code stipulated that a "spy is a person who secretly, in disguise or under false pretense, seeks information with the intention of communicating it to the enemy"; Lieber Code, article 88 and see Brussels Declaration, article 19.

[63] Tallinn Manual, Rule 66(a), AMW, Rule 119. However, a combatant who acts as a spy loses the right to be a POW and may be treated as a spy if captured before he reaches the army on which he depends; Tallinn Manual, Rule 66(b).

[64] Tallinn Manual, commentary accompanying Rule 66, paragraph 2 citing API Commentary, paragraph 1779. See also AMW, Rule 120 where it is noted that an individual is not engaged in espionage if, while gathering the information, he is in the uniform of his armed forces. However, members of military aircrew who wear civilian clothes inside a properly marked military aircraft are not spies; AMW, commentary accompanying Rule 120, paragraph 2.

not engaged in espionage. In the cyber context, therefore, most acts of remotely undertaken information gathering will not constitute espionage, whereas close access cyber operations to obtain information from a targeted closed computer system using, for example, a memory stick will be espionage if the targeted computer is located within the enemy's zone of operations provided that the other elements of espionage are present.[65]

8. DO FORESEEABLE NOTIONS OF CYBER OPERATIONS CHALLENGE THE LAW?

We have now seen how the law regulates deception operations as they have been undertaken during traditional types of military operation. The question that now needs to be considered is whether the pervasive use of cyber deception to which we referred in the first section of this paper has implications for these traditional legal rules. To put it more succinctly, will these new kinds of cyber operation, and the associated extensive deception operations, challenge the law by requiring that existing legal rules be adjusted to permit such deceptions to be used more frequently, or will the existing rules prevail, for example because only deceptions that comply with traditional interpretations of the law will in fact be permitted and, thus, undertaken?

To make sense of this generalized question, we should very briefly consider a number of scenarios. They are purely illustrative, do not reflect all foreseeable kinds of cyber deception operation that may be relevant, but will at least give an indication as to the sorts of legal issue that may be expected to arise. The scenarios are:

A State A undertakes a remote access cyber attack making it appear that the attack has been undertaken by State B. State B is a co-belligerent of State A, so there is no breach of international law by State A as a result of the deception *per se*. However, if the subject of the cyber attack were to mount an automatic response attack against State B, this would likely breach international law as being an unlawful use of force or, even, an armed attack. This implies a need for caution to be exercised before responding, to seek to ensure that the true author of the initial attack is being engaged.

[65] If undertaken by a civilian, remote cyber information gathering and close access cyber espionage are likely to constitute direct participation in the hostilities, which, if undertaken by a civilian, would render him or her liable to attack while so engaged. It is also likely to breach the domestic law of the territory where the activity occurs and the persons concerned are therefore liable to be tried for the relevant offences; AMW, Rule 121.

B During an international armed conflict, State A undertakes a remote access cyber attack using a worm incorporated within an attachment to an Email and making it appear that the attack has been undertaken by State B. State B is a neutral and a reproduction of its national flag is employed to make the Email appear to have come from an authentic State B source. Making use of the neutral's flag clearly breaches API, article 39(1). If the flag were not to be used and the deception were based on use of the neutral government's domain name, such as '.gov.uk', the Tallinn Manual's Experts were divided as to whether such activity is unlawful.

Again, however, an automatic response against State B would, on the face of it, constitute an unlawful use of force or armed attack, and in both this and the previous example, it would seem advisable that diplomatic activity be undertaken to seek to confirm responsibility for the attack before a use of force, cyber or otherwise, in response is decided upon.

C A false Email sent to enemy personnel causes them to believe that they are being invited to attend a meeting to discuss the surrender of the unit sending the Email. The sending unit has no intention of surrendering, but the deceived personnel suffer a road accident on the way to the proposed meeting resulting in death or injury. The deception operation is not, arguably, the proximate cause of the accident and while the message was perfidious, the Rule is not broken because the death or injury are not proximately caused by the perfidy.

D A false Email sent to enemy personnel causes them to believe that they are being invited to attend a command group meeting. The Email appears, falsely, to have been sent by the Enemy superior commander to his subordinate commanders. As it is permissible to feign enemy authorship of cyber communications, the operation would appear to be lawful, even if enemy personnel are as a result killed or injured.

E Having hacked into the enemy computerized Common Operating Picture programme, false data is inserted making it appear that friendly forces are concentrated distant from their true location. This would be a lawful ruse.

If the false data were to make it appear that friendly forces are concentrated in or near a small town populated with civilians, and if the enemy as a result attacks that location causing incidental damage and casualties among the civilians, the perfidy rule would, arguably, not have been breached because no civilian status has been feigned in relation to the friendly forces themselves.

F Personnel from a State that is not party to API who are members of a military unit pretend to have civilian status. They dress in civilian clothes, alter the unit's website so as to make it appear civilian, include assertions in the website of its civilian status and omit all references to military ranks in any electronic communications from the unit.

On the approach of an attacking unit, the personnel from the State not party to the conflict are not attacked because of their apparently civilian status, but after that attack, succeed in capturing enemy personnel and in damaging their military equipment. No perfidious offence is committed as the customary perfidy provision does not extend to capture, and neither the customary nor the API rule extends to damage caused by the perfidy.

G A cyber operation deceives the targeted computerized perimeter security system to believe that enemy personnel are in fact friendly forces. The enemy personnel then enter the closed military facility protected by the security system and wreck the facility, capture its personnel and kill the commander. If the attackers enter in uniform, the operation would not be prohibited perfidy. If they enter in civilian clothes, it likely would be.

H A cyber operation deceives the targeted computer that protects the perimeter of a military, closed IT system. The deception causes the protecting computer to believe that an attachment to an Email has been received from a non-threatening civilian source and, thus, may be opened in accordance with IT protocols without undertaking certain preliminary checks. The attachment, when opened, causes the server to which the targeted computer is connected to shut down thus denying service to all users, with the result that the targeted unit's water purification system instantly malfunctions causing death and injury through disease/infections. According to the majority view among the Group of Experts, this deception of the targeted computer would be perfidy and, as it leads to death and injury, would be prohibited.

9. CONCLUSION

It is clear that deception operations will become of increasing importance as cyber warfare techniques become more widely employed in armed conflict. The traditional rules draw a vital distinction between lawful deception, and that which is prohibited because, causing death, injury or capture, it leads the adversary to believe that he is entitled to or is obliged to accord legal protected status. There is no reason to believe that this traditional distinction will be either eroded or abandoned in the cyber context. The focus in the definition of espionage on the geographical location of the spy may seem outdated in an age when remote access cyber operations may be employed to intrude into the most secret, protected and sensitive parts of the enemy's information architecture. Outdated or not, the geographical element in the espionage definition is customary, and thus binds all States, and seems unlikely to change.

The rules prohibiting the use of certain flags, emblems, insignia or uniforms also may appear to some to be somewhat anachronistic. The degree to which the capabilities of, and risks posed by, cyber operations will call the adequacy of these rules into question remains to be seen. For the time being at least, they have stood the test of time and are consistent, essentially, with the philosophy underpinning the perfidy rule.

Having put forward this case in support of the legal status quo, the author would offer one word of caution. It is this. If increasingly pervasive cyber capabilities are so used that deception operations become the rule rather than, relatively speaking, the exception, and if as a result little or no reliance can in future be placed on the information that would traditionally support targeting decision making, what are the consequences for the practical ability of combatants to comply with the distinction, discrimination, proportionality and precautions rules that lie at the core of targeting law? It seems to the author that some at least concrete basis for reliable decision making is central to the practical delivery of these protective principles. Widespread use of deception must not, it is suggested, become the cause of a slide into ‘anything goes’.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Legal Aspects of a Cyber Immune System

Janine S. Hiller
Department of Finance
Pamplin College of Business
Virginia Tech
E-mail: jhiller@vt.edu

Abstract: The malicious and criminal attacks against individuals, businesses, and nations on the Internet and in cyberspace must be mitigated in order to protect citizens and nations. One cyber security vision is the cyber immune system. Such a system would include automatic defense mechanisms based on incomplete attribution, continuous monitoring, pattern recognition, and the application of a set of rules designed to isolate or destroy the abnormal actor, or attacker. The cyber immune system would operate at a distributed level, at the speed necessary to thwart constant and ever changing threats. From a legal perspective, it matters if a state or private entity applies the system. For example, if a state actor is involved, then due process, and the protection of fundamental rights such as privacy and speech, are relevant to the action taken, while if a private entity applies the cyber defense then relevant legal issues include property, contract, and regulatory limits. While the automated nature of a cyber defense may present legal challenges to both state and non-state actors, it may mitigate the legal ramifications of human decision making if the system of rules is carefully crafted.

Keywords: *cybersecurity, privacy, property, speech, law*

1. INTRODUCTION

Concerted cyber attacks against the US banking system[1] are but one of the newest reported instances, among many, of the continuing and evolving threats against cyber entities. It is clear that "normal" cyber security is failing to mitigate threats and that new ideas for protecting citizens and nations should be considered. Technical security advances offer potential solutions for cyber defense, however they face legal uncertainties within a complex environment.

The original designers of the Internet focused on a free and open communications system, not foreseeing perhaps that the distributed design of the communications network would lead to its own insecurity.[2] But the values inherent in the design are those that imbue the medium with its power and ability to serve democratic principles. Novel applications of cyber security systems should incorporate society's values for privacy, freedom, and the rule of law into the distributed defense design. This task is made difficult because of the unique intersection of law and technology among different layers of state and non-state actors. Realizing that systems and actors will differ, this paper identifies, at a high level, the major legal issues that may arise in designing and implementing a cyber defense that is analogous to a human immune system composed of differing autonomous, distributed, learning systems that defend the person from attack. A holistic view of cyber defense is presented, emphasizing the potential contributions of a preventative, private law perspective. Because in many nations the cyber infrastructure is owned primarily by the private sector, actions that strengthen the cyber safety of those entities will ultimately strengthen national security. In addition, managing cyber security in the private sector will lead to fewer conflicts at the international level.

The type of technical system envisioned would include automatic defense mechanisms based on incomplete attribution, continuous monitoring, pattern recognition, and application of a set of rules designed to isolate or disable the abnormal actor, or attacker. Such a system would operate at a distributed level, at the speed necessary to thwart continuous and ever changing threats. The system would also be embodied systemically and limited to mitigative and preemptive actions, as opposed to individual, retributive action. From a legal perspective the

---

1 *See* Nicole Perlroth & Quentin Hardy, "Bank Hacking Was the Work of Iranians, Official Say," *New York Times* (January 8, 2012) *available* at http://www.nytimes.com/2013/01/09/technology/online-banking-attacks-were-work-of-iran-us-officials-say.html.

2 Chris C. Demchak, *Resilience and Cyberspace: Recognizing the Challenges of a Global Socio-Cyber Infrastructure (GSCI),* 14 J. Comp. Pol'y Analysis 254, 258-61 (2012) ("Cyberspace began as a pure document sharing mechanism for which security was about physical reliability, not human predatory behaviors.").

structure of the system is relevant and it matters if a nation-state or private entity applies the system. For example, if a state actor is involved, then due process, the protection of fundamental rights such as privacy and speech, are relevant to the action taken, while if a private entity applies the cyber defense then relevant legal issues include property, contract, and regulatory limits. While the automated nature of a cyber defense may present legal challenges to both state and non-state actors, it could possibly mitigate the legal ramifications of human decision making if the system of rules is carefully crafted.

2. IMMUNE TYPE DEFENSES

The goal of this section is to identify fundamental elements of an immune inspired cyber defense system that may invoke legal questions, thus facilitating discussion of the corresponding challenges of implementation in a democratic society. It is recognized that the technical level of discussion is general in nature and that the term cyber immune system, as described in this paper, could also incorporate common elements of certain artificial intelligence or intelligent systems.

Research in the 1990's described the metaphorical use of the human immune system to construct elements of a cyber security system.[3] These cyber defense elements seek to mimic the automatic actions of human cells and organs to respond to new, previously unknown threats, take defensive action, and internalize learning for future defense. Biancaniello et al. state that, "Artificial Immune Systems have enjoyed a number of application successes in Cyber Defense including web-server behavioral anomaly detection, network intrusion detection, the detection of malicious code execution, and operating system call monitoring."[4]

The US document, "Enabling Distributed Security in Cyberspace," describes current security as depending on reactive actions and human intervention.[5] Yet the Slammer worm infected 90 percent of its hosts in 10 minutes, scanning 55 million targets each second.[6] In order to defend against rapidly spreading, sophisticated, and persistent threats, the document identifies an Automated Course of Action (ACOA) as the first building block needed for a "Healthy Cyber Ecosystem."[7] The

---

[3] *See* Anil Somayaji et al., *Principles of a Computer Immune System,* 1997 New Security Paradigms Workshop 75 (1997).

[4] Paul Biancaniello et al., *AIR: A Framework For Adaptive Immune Response for Cyber Defense*, *available at* www.atl.lmco.com/papers/2021.pdf *at 3 (December 19, 2011), (an unclassified document prepared by authors from Delaware State University).*

[5] U.S. Dept. of Homeland Security, Enabling Distributed Security in Cyberspace 6 (2011).

[6] Id. at 6-7.

[7] Id. at 8-11.

human immune system is then used as a metaphor to describe the elements of such a system. The human system description includes multiple levels of defense, at both the cell and system level, including synchronization/communication, identification methods, and actions to destroy and/or immobilize an attack (for example, a virus). An automated cyber security system is conceptualized in a similarly decentralized and highly synchronized manner. Such a system could incorporate continuous monitoring, pattern recognition, and anomaly detection to identify non-entity threats, respond according to preset policies to block, shut down, or disable the threat, and then audit and share information among a system of users; all done automatically and at the speed of real time computer execution. The aggregation and maintenance of data is important within such a system so that adaptive/intelligent learning occurs. The ecosystem might include a centralized public entity that would facilitate sharing, learning, and techniques for immunization from future damage.

Within this broad description of a cyber immune system, certain data collection elements are required for effective implementation; IP and addressing information, deep packet inspection, data mining, and data retention. Like a human system that achieves immunities by "remembering" and defending against a virus, a fully operational cyber security/defense system will require longitudinal information about malicious actors and actions and continuous monitoring for both known and new threats. In addition, one must note that just like a human system, a cyber immune system will not operate perfectly; attribution may be based on probabilities, behavioral information, and past actions.

It is important to note that the *systematic* defense/security envisioned here is distinct from an individual "strikeback" offensive action.[8] Because these are individual retributive actions against particular perpetrators, they would not fall under an immune defense system that is adopted broadly in a community of users (the system) and operates to prevent damage and mitigate attacks. A private strikeback is legally suspect, although there have been arguments for supporting such action.[9] International law of warfare would apply to a nation taking such action, and would include such issues as attribution and self defense.[10] Adoption of an immune defense could potentially avoid the escalation of cyber conflicts by securing systems from attacks and vulnerabilities.

---

[8] For an exhaustive treatment of the law of cyber counterstrikes and a proposed way forward, see Jay P. Kesan & Carol M. Hayes, *Mitigative Counterstriking: Self-Defense and Deterrence in Cyberspace*, 25 Harv. J. L. & Tech. 415 (2012). Discussions indicate that industry offensive action is actually not a new phenomenon, although news reports are that it could be growing. *See* Dennis Fisher, Debate Over Active Defense and Hacking Back Crops up at RSA, Feb. 28, 2012, *available at* .http://threatpost.com/en_us/blogs/debate-over-active-defense-and-hacking-back-crops-rsa-022812 .

[9] *See Mitigative Counterstriking, supra* note 8, at 531-32.

[10] *See* Matthew E. Castel, *International and Canadian Law Rules Applicable to Cyber Attacks by State and Non-State Actors,* 10 Can. J.L. & Tech 89, 95-102 (2012). For a discussion of how the law of war would apply, see David E. Graham, *Cyber Threats and the Law of War,* 4 J. Nat'l Security L. & Pol'y 87 (2010).

Several examples can be used to illustrate components of an existing automated immune system, including continuous monitoring, data analysis, and automated action. The U.S. employs multiple information collection and monitoring methods within its National Cybersecurity Protection System, described as "an integrated system for intrusion detection, analysis, intrusion prevention, and information sharing"[11] in order to defend federal civilian systems. Different elements of the system collect network information, analyze the information to detect cyber threats, and distribute cyber security information across participating federal systems.[12] NCPS includes not only analysis and detection, but also intrusion prevention by agreement with Internet Service Providers that can take action against Internet traffic at the border of federal systems, i.e. as it enters or leaves those networks.[13] However, although some information sharing occurs voluntarily and will be expanded under the recent Executive Order,[14] the information collection system and defensive actions are limited to the federal civilian government and are not universally distributed.

In the private sector, Facebook describes its cyber system for security as "the Facebook Immune System because it learns, adapts, and protects in much the same way as a biological immune system."[15] Within their proprietary, closed platform, Facebook monitors users and their accounts in order to prevent criminal actions like stolen credit cards and passwords that can lead to economic losses. The automated system will not only halt the attack, it will take steps to destroy the "assets" of the attacker in order to dissuade future attacks. In 2011, Facebook utilized 2,000 servers, 200 models, and 20 billion daily checks to operate the system.[16] Being a social media company, Facebook faces unique risks; however, this example illustrates that a private entity will tailor its cyber security to meet the specific needs of its business, suppliers, and customers. It might be seen as a cyber immune system within that closed system, but does not reach the distributed and broader cyber immune model.

---

[11] U.S. Dept. of Homeland Security, Privacy impact Assessment for the National Cybersecurity Protection System (NCPS) 1 (July 30, 2012).

[12] Id. at 8-9 (includes EINSTEIN 1, 2, and 3, Security Information and Event Management (SIEM), Packet Capture (PCAP) as well as other technical elements).

[13] Id. at 18.

[14] Executive Order, "Improving Critical Infrastructure Cybersecurity" (Feb. 12, 2013), *available at* http://www.whitehouse.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity.

[15] "National Cybersecurity Awareness Month Recap and the Facebook Immune System," November 10, 2011, http://www.facebook.com/notes/facebook-security/national-cybersecurity-awareness-month-recap-and-the-facebook-immune-system/10150352042420766 .

[16] Id. *See also* Tao Stein et al. "Facebook Immune System," *available at* http://research.microsoft.com/en-us/projects/ldg/sns2011prog.aspx.

Japan has reportedly contracted with Fujitsu to develop a protective virus that will detect, trace, and disable malware or attackers across networks.[17] The unique aspect of the proposed virus is that it would act automatically to follow the attack back across multiple computers, collect information, and take action to neutralize the attack at each stage. Many details are unknown about the Japanese system, but its highly automated and distributed actions seem to meet some of the elements of an immune system.

It is also worth noting future potential developments of programs that can be likened to an immune system. In September, 2012, the U.S. Department of Homeland Security and the Department of Commerce issued a Request for Information entitled; "Developing a Capability Framework for a Healthy and Resilient Cyber Ecosystem Using Automated Collective Action."[18] The RFI sought information about the feasibility and challenges of pursuing a system that would include "automated information sharing and collective action, reference data, machine learning, behavior monitoring based on business rules, interoperable systems and organizational policies, and authenticated users and systems."[19] Reports linked existing programs in the Energy Department and the Federal Aviation Administration to this concept of a "learning, self-healing network."[20] Utilizing the concepts described in the NCPS, and intrusion protection platforms, a future system would, at least theoretically, provide for real-time automated responses to cyber intrusions across a wide infrastructure.

Interestingly, in February, 2013 a paper written by the New England Complex Systems Institute in 2008 for the Chief of Naval Operations Strategic Studies Group was released; it was titled, "Principles of Security: Human, Cyber and Biological."[21] The authors described the human immune system and its ability to evolve defenses. The report noted, by comparison, the inherent security weakness of the Internet architecture that transports communication packets in content neutral fashion. In conclusion the authors suggested two alternatives; distributed automatic security at

---

[17] Yomiuri Shimbun, "Govt working on defensive cyberweapon/Virus can trace, disable sources of cyber-attacks," *Daily Yomiuri Online* (January 3, 2012) *available at* http://www.yomiuri.co.jp/dy/national/T120102002799.htm.

[18] U.S. DEPT. HOMELAND SECURITY, DEVELOPING A CAPABILITY FRAMEWORK FOR A HEALTHY AND RESILIENT CYBER ECOSYSTEM USING AUTOMATED COLLECTIVE ACTION (Request for Information) (2012).

[19] Id. at 3.

[20] William Jackson, "Agency programs show outlines of future cyber ecosystem," *Government Computer News* (November 9, 2012) *available at* http://gcn.com/Articles/2012/11/09/Agency-programs-show-outlines-of-future-cyber-ecosystem.aspx. *See also* Peter M. Fonash, "Identifying Cyber Ecosystem Security Capabilities," Sept./Oct. 2012 Crosstalk 15 (2012) (cross referencing types of attacks with desired cyber ecosystem/defense design).

[21] Blake Stacey & Yaneer Bar-Yam, New England Complex Systems Institute, Principles of Security: Human, Cyber and Biological (2008).

the user level or a change in Internet protocols so that routers could inspect content for malware.[22]

In summary, currently there are partial automated cyber immune defense systems at some stage, public and private, but no complete system exists. Visions for a system include systems monitoring, longitudinal information collection, deep packet inspection, information sharing, system "learning," and proactive, automated action to takedown or quarantine bad actors based on behavioral and technical information. If a cyber immune system were to be employed at a national level, private sector actors as well as network administrators would be essential participants. In contrast to a military operation that depends on a hierarchy of command and control, a cyber immune system is distributed among all participants in order to exponentially increase the security of the network. Vast amounts of information about port scans, attack methods, signatures, behavioral actions, and the like is shared so that the immune system can learn about vulnerabilities and block attacks or cure weaknesses in defense, and redistribute the aggregated knowledge for individual action.

While many technical issues remain in the adoption of a metaphorical cyber immune ecosystem, they are matched by the legal and policy questions engendered as well.

3. LEGAL ISSUES

A system such as the cyber immune defense system described is never simply a technical solution to a thorny problem; it is "political to its very core,"[23] as the design and implementation will embody societal values and choices in a democratic society.[24] Data collection that aggregates great volumes of content related information longitudinally can identify patterns of harmful activity, yet can also threaten individual privacy and chill speech. Information sharing can provide the needed tools to prevent damage to systems and property, yet has the potential to thwart checks on government involvement in citizens' lives. The automated takedown or quarantine of websites, domain names, or software is necessary to respond in real-time to prevent illegal activity and maintain national security, yet its imperfect application can impede speech rights, violate property, and potentially undermine democratic discourse. The following discussion highlights these fundamental legal issues.

---

[22] Id. at 10-12.

[23] Helen Nissenbaum, *Where Computer Security Meets National Security*, 7 Ethics & Info. Tech. 61, 62 (2005).

[24] Id.

Legal protection of electronic property is built in part on criminal laws, including the Computer Fraud and Abuse Act (CFAA)[25] in the United States, and domestic laws that enforce the international Budapest (Cybercrime) Convention.[26] The CFAA makes unauthorized access of protected computers (including those connected to the Internet, by interpretation) a crime; intentional unauthorized access to federal computers does not require damage, while intentional or reckless access to other computers can require that damage occur, such as the loss of intellectual property or the degradation of the system.[27] The international Cybercrime Convention and the European Union Framework Decision on attacks against information systems[28] provide similar prohibitions against illegal access to information systems, illegal system interference, and illegal data interference.[29]

A cyber immune defense imagines a distributed approach that goes beyond the traditional deterrence effect of criminal actions, therefore requiring a broader analysis of actions by not only the government, but also the private sector. Application of the system should take into account the ways that the design of the technology implicates the important areas of speech, privacy, and property. The following sections discuss these areas and the basic laws that apply based on whether the action is led by government or the private sector.

A. SPEECH

Government Action. Freedom of speech is enshrined in fundamental laws across the globe, and the First Amendment in the US prevents the government from limiting free speech; even computer code has been interpreted to be a form of speech.[30] The recent Middle East changes provide a reminder of how important free speech is to political discourse; a discourse that occurred significantly due to Internet communications. Although protection of speech may vary in application between leading democracies,[31] it is undeniable that the right of speech is essential to the preservation of fundamental freedoms.

---

[25] Counterfeit Access Device and Computer Fraud and Abuse Act of 1984, 18 U.S.C. § 1030 (1984).

[26] Convention on Cybercrime, *opened for signature* Nov. 23, 2001, E.T.S. No. 185, *available at* http://conventions.coe.int/Treaty/en/Treaties/Html/185.htm.

[27] *See* Chris Kim et al., *Computer Crimes,* 49 Am. Crim. L.Rev. 443, 460-62 (2012).

[28] Council Framework Decision 2005/222/JHA of 24 February 2005 on attacks against information systems, *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32005F0222:EN:NOT

[29] *See* Liis Vihul et al., Legal Implications of Countering Botnets 9 (2012).

[30] *See, e.g.,* Universal City Studios, Inc. v. Corley, 273 F.3d 429 (2d Cir. 2001).

[31] For example, the US and German conceptions of freedom of the press and speech differ. *See,* Christopher Witteman, *Information Freedom, a Constitutional Value for the 21st Century,* 36 Hastings Int'l & Comp. L. Rev. 145 (2013) (speech protected from a broader principle in Germany).

In order to identify malicious actions through behavioral information, signatures, and the like, a cyber immune system would automatically collect information and data from users' traffic longitudinally, thereby posing a real potential harm to the essential values of privacy and speech. The widespread collection of information about individual communications is extraordinarily sensitive, especially when an immune system would go further than collecting address and IP information, and would undertake deep packet inspection in order to detect and take action to neutralize malicious activity.[32] This type of packet inspection, reportedly used by China to block the websites it censors,[33] poses a great threat to individual liberties. Government application of these technologies to civilian networks is particularly problematic from the US standpoint; the current administration firmly opposed legislation, ultimately defeated, that would have allowed government agencies to monitor domestic private communications in order to actively defend them from attack.[34]

The rights to private life and freedom of expression and opinion are also protected in the Universal Declaration of Human Rights and in the treaty, the International Covenant on Civil and Political Rights.[35] In addition, in 2011 the UN Special Rapporteur for freedom of expression released a report that discussed the importance of Internet communications,[36] and ensuing coverage labeled the report as a declaration that Internet access is a human right.[37] Statutes in Estonia, Finland, France, and Costa Rica for example, provide a right to Internet access for citizens.[38] Any automated system will need to incorporate strong protections for protecting access in order to ensure rights to free speech.

---

[32] *See* Ted Stevenson, "Network Security Essentials: Deep Packet Inspection," Feb. 28, 2012, *available at* http://www.enterprisenetworkingplanet.com/netsecur/network-security-essentials-deep-packet-inspection.html (deep packet inspection is necessary to stop sophisticated attacks).

[33] Alex Wang, "What is Deep Packet Inspection?' Feb. 1, 2012 *available at* http://www.pcworld.com/article/249137/what_is_deep_packet_inspection_.html.

[34] Ellen Nakashima, *When is a cyberattack a matter of defense?* Wash. Post, Feb. 27, 2012 *available at* http://www.washingtonpost.com/blogs/checkpoint-washington/post/active-defense-at-center-of-debate-on-cyberattacks/2012/02/27/gIQACFoKeR_blog.html. The issue of monitoring foreign communications is a separate issue, and not discussed in this article.

[35] Kent Roach, *Must We Trade Rights for Security? The Choice Between Smart, Harsh, or Proportionate Security Strategies in Canada and Britain,* 27 Cardozo L. Rev. 2151, 2152-53 (2006) (speech rights may also be restricted when balanced with other interests under the doctrine of proportionality).

[36] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, delivered to General Assembly*, U.N. Doc. A/HRC/17/27 (May 16, 2011).

[37] Nicholas Jackson, *United Nations Declares Internet Access a Basic Human Right*, The Atlantic, June 3, 2011, *available at http://www.theatlantic.com/technology/archive/2011/06/united-nations-declares-internet-access-a-basic-human-right/239911. See also* Young Joon Lim & Sarah E. Sexton, *Internet as a Human Right: A Practical Legal Framework to Address the Uniqure Nature of the Medium and to Promote Development,* 7 Wash. J.L. Tech. & Arts 295, 297 (2012).

[38] Victoria Ekstedt, Tom Parkhouse & Dave Clemente, *Commitments, Mechanisms & Governance, in* National Cyber Security Framework Manual 163-66 (Alexander Klimburg, ed., 2012).

Private Action. Actions by private parties that affect speech may not be prohibited in the same manner as those by government entities. Businesses control the use of their systems, and to meet the goal of maintaining network quality ISPs often have the right to manage and protect network traffic. United States law, for example, allows providers to monitor and even disclose communications in order to maintain service levels.[39] Agreements, formalized in contracts between service providers and their customers, delineate these management rights. Furthermore, general terms of use between private entities and the broader community of users can negotiate use limitations and access rights. The recent voluntary Copyright Alert System agreement between ISPs and copyright owners, whereby ISPs will monitor and notify users of potential copyright violations, is an example of a kind of mediation activity by ISPs.[40]

B. PRIVACY

Government. In the electronic world, speech and privacy are intertwined, as surveillance of communications can breach privacy of information and chill speech. The legality and extent of surveillance by governments varies greatly. A survey of law enforcement access to data in ten countries showed that in the midst of an investigation that in all ten countries access to electronic data was allowed; eight did not require approval of a formal request.[41] In comparison, the Fourth Amendment of the US Constitution prohibits unreasonable searches and seizures and requires probable cause for a warrant to obtain access to places when there is a reasonable expectation of privacy.[42] Thus, the law restricts government access to the content of electronic communications with judicial approval, but is not interpreted to restrict access to address information such as header or IP information. The Electronic Communications Privacy Act (ECPA), Stored Communications Act (SCA), and Wiretap Acts as well as other state and federal laws, protect the rights of citizens to privacy and autonomy.[43]

The ECPA, amended by the SCA, protects the privacy of electronic communications

---

[39] *See* Scott J. Glick, *Virtual Checkpoints and Cyber-Terry Stops: Digital Scans To Protect the Nation's Critical Infrastructure and Key Resources*, 6 J. Nat'l Sec. L. & Pol'y 1, 8 (2012).

[40] *See*, Peter Groh, *Through a Router Darkly: How New American Copyright Enforcement Initiatives May Hinder Economic Development, Net Neutrality and Creativity,* 13 U. Pitt. J. Tech. L. & Poly 1 (2012).

[41] *See* Steven C. Bennett et al., *Storm Clouds Gathering for Cross-Border Discovery and Data Privacy: Cloud Computing Meets the U.S.A. PATRIOT Act,* 13 Sedona Conf. J. 235, 247 (2012).

[42] *See Virtual Checkpoints, supra* note 39, at 9-12

[43] For a detailed discussion of how a myriad US laws meet the requirements of Section 15 of the Cybercrime Convention to safeguard human rights, for example, see Discussion Paper, Data Protection and Cybercrime Division, Directorate General of Human Rights and Rule of Law, *Article 5 Conditions and Safeguards under the Budapest Convention on Cybercrime,* Nov. 8, 2011, *available at* www.coe.int.

and applies to both the government and service providers. Police must seek a warrant to obtain communications in some cases, or a subpoena under other circumstances. Both criminal and civil penalties for violations are possible. However, exceptions allow entities to share information related to the investigation of computer trespass, and ISPs are allowed to share information in emergency situations.[44]

The Cybercrime Convention requires that competent authorities have access to specific data held by a person or system in whatever method it is stored, including traffic data. An ISP may be required to assist collecting and accessing the data. The convention anticipates that the request will be pursuant to an active investigation of wrongdoing, however.[45]

Private Action. As described above, the ECPA is the primary US law protecting privacy of electronic communications, and it prevents access by private parties, with some exceptions. One of the exceptions is based on consent of the party. For example, Google has reportedly shared information with government agencies in order to trace the source of a series of cyber attacks; arguably terms of use agreed to by customers allow Google to share personal information for the purpose of 'protecting the rights or property of Google or our users."[46]

In the EU, the Data Protection Directive, and other telecommunications acts,[47] apply to the private sector and ISP actions, and protect personally identifiable data. Through harmonized national laws, data collection requires user consent, is limited to the intended purposes, and individuals have the right to information about the data that is held about them. Differences in national laws occur, such as whether IP addresses are protected personal information.[48] An ISP involved in collecting personally identifiable information for a cyber immune system will invoke the provisions of the Directive unless consent is obtained.

C. PROPERTY

One of the major purposes of a cyber immune system is to protect the property of citizens and government from attack, therefore supporting the goal of national security. Property, though, can exist in multiple forms. Intellectual property, such

[44] *See* Gregory T. Nojeim, *Cybersecurity and Freedom on the Internet,* 4 J. Nat'l Security L & Pol'y 119, 125-28 (2010).

[45] *Cybercrime Convention, supra* note 26, at arts. 16-21.

[46] Stephanie A. Devos, *The Google-NSA Alliance: Developing Cybersecurity Policy at Internet Speed,"* 21 Fordham Intell. Prop. Media & Ent. L.J. 173, 209-212 (2010).

[47] *See* Vihul, *supra* note 29, at 49-53 (also comparing the national laws applied to ISPs in Estonia and Germany).

[48] Id. at 19 (national laws may differ in application however).

as trade secrets, business plans and the like, supports the economic stability of both business and the country, while the property of privately held critical infrastructures can consist of electronic controls that affect physical performance, such as the electric grid. Computer systems themselves are a form of property in which the right to exclude others is incorporated. The computers, controls and most of the ISP's[49] and networks that make up the Internet are primarily privately owned. In the United States, "virtually all broadband networks"[50] fall into the private ownership category, therefore implicating laws of private property. Ironically, the same laws that criminalize cyber attacks may also limit proactive cyber defense.

Government. Government action related to the rights of speech and privacy can also affect property in the electronic environment. The requirement of due process and fundamental fairness in areas of property and liberty could apply to an automated action taken in a cyber immune system; if the government takes down a website or restricts Internet access, principles of notice and an opportunity to be heard become relevant.[51] If malicious cyber actors use "innocent" computers to launch an attack and an automatic defense is triggered, innocent parties may be negatively affected by government action. In addition, if the implementation of an automatic cyber immune defense occurs across networks it could violate property rights in privately owned computers if it involves unauthorized access to private parties' proprietary system, or if it is beyond the authorization of a network provider, even though it intends to disarm a criminal actor.[52]

Private Action. Common law concepts of trespass to property can be applied to computer intrusions in addition to the cause of action for unauthorized access. An automatic system that accessed a website in violation of its terms of use has been held in the US to give rise to a claim of trespass;[53] without owner consent, such as an automatic virus update, a cyber immune system implemented by a private entity such as an ISP could run the risk of violating property rights. The argument has been made, however, that self-defense could allow mitigation across network property

[49] ISP and network operator are used interchangeably to designate an entry point to the network. While the paper does not discuss the potential involvement of Tier One telecommunications companies, the backbone operators, those companies may have some of the same opportunities for monitoring. (There are however, more difficult questions for monitoring at this level.) *See* James Andrew Lewis, Speech at the Sasakawa Peace Foundation: Rethinking Cybersecurity-A Comprehensive Approach (Sept. 12, 2011), *available at* http://csis.org/publication/rethinking-cybersecurity-comprehensive-approach.

[50] Charles B. Goldfarb & Lennard G. Kruger, Cong. Research Service, 7-7500, Infrastructure Programs: What's Different About Broadband? 2 (2009).

[51] *See* Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process,* 40 Ga. L. Rev. 1 (2005). *Also see* Sean M. Condron, *Getting It Right: Protecting American Critical Infrastructure,* 20 Harv. J.L. & Tech. 403, 416-18 (2007) (noting due process importance, but also suggesting a balance).

[52] *See* James P. Farwell, *Industry's Vital Role in National Cyber Security,* 2012 Strategic Stud. Q. 10, 30 (2012).

[53] eBay v. Bidders Edge, 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

lines.[54] In the EU, the 2009 Telecom Directive[55] requires public communication providers to 1) provide secure services, 2) report breaches, and 3) share a summary of material breaches with the European Network and Information Security Agency (ENISA).

4. DISCUSSION OF CYBER IMMUNE DEFENSE WITHIN THE GLOBAL SOCIO-CYBER CONTEXT[56]

Envisioning and implementing an automated cyber immune defense should intentionally preserve the fundamental rights that the security ultimately seeks to protect; property, privacy of communication, and speech. Legal limitations to protect these rights differ based on who will undertake the defensive steps, whether it be maintenance of a database to identify malicious actors or installation of software to purge victims' infected computers, for example.

Distributed security will require the participation of both private and state actors, both for effectiveness and for policy reasons.[57] ISPs may be particularly situated to play a role in the security ecosystem. Logs at the infrastructure level showed recently that 162 of 168 Fortune 500 companies were compromised by hackers at some point of time,[58] and an ISP has "unparalleled visibility into global networks"[59] being "well positioned to aid" in "proactive" actions.[60] An ISP is located within network infrastructure between victim and attacker, perhaps a kind of neutral zone, handling traffic that is not within the "perimeter" of either side. Automated actions taken to disable or immobilize an attack or bad actor could be designed as part of network management, analogous to how actions to stop spam have been taken in the past. Defense and security at this system point might defuse, at least in part, the

[54] Kesan, *supra* note 8, at 520-21.

[55] EU Directive 2009/140/EC

[56] The term Global Socio-Cyber is found in Demchak, *supra* note 2 (Resilience and Cyberspace: Recognizing the Challenges of a Global Socio-Cyber Infrastructure).

[57] *See* Paul Rosenzweig & James X. Dempsey, *Einstein 3.0, in* Patriots Debate 115-34 (Harvey Rishikof, Stewart Baker & Bernard Horowitz, eds., 2012).

[58] Joseph Menn, *Hacked companies fight back with controversial steps*, *Reuters*, June 18, 2012 (Neustar found evidence of a breach at some point of time at companies).

[59] William J. Lynn, III, *Remarks on Cyber at the RSA Conference*, Feb. 15, 2011, *available at* http://www.defense.gov/speeches/speech.aspx?speechid=1535 (they may also "have the best operational capacity to respond").

[60] OECD, "Proactive Policy Measures by Internet Service Providers against Botnets," OECD Political Economy Paper No. 199, at 8, *available at* http://dx.doi.org/10.1787/5k98tq42t18w-en. For an argument that government should be the entity in control see Jay P. Kesan & Carol M. Hayes, *Thinking Through Active Defense in Cyberspace*, in Proceedings of a Workshop on Deterring CyberAttacks: Informing Strategies and Developing Options for U.S. Policy 334 (2010).

debate about how far beyond its own systems a victim can go to defend itself against cyber intrusions. In addition, at this juncture ISP actions rather than government action could mediate the potential threat of government overreach.[61] The same is arguably true of the predictive and learning aspect of a cyber immune system that requires the collection and longitudinal analysis of enormous amounts of potentially personally identifiable information.[62]

The automated nature of a cyber immune system could potentially incorporate actions that would effectuate legal standards and strengthen the protection of civil liberties. An immune defense would automatically identify and disable malicious code and cyber threats based on a reasonable and sufficient level of evidence, but the standard could potentially be less sensitive to attribution questions because it is not applied by a government actor. If an ISP outside of government control undertakes robust action it would probably not be considered an act of a nation state.[63] Establishing a means for redress for mistakes and a waiver of liability for ISPs if actions are taken in good faith and according to reasonable security standards are important considerations.

An automated cyber immune system that is implemented at the ISP level might contribute significantly to national security and property protection while maintaining access and facilitating speech for the community. National security can be strengthened by private actions that increase the security of computers and systems of computers from attack, and ISPs seem to be in a good position to aid in that protection.

If a nation implemented an automated process, then perhaps established levels of technical predictability could form the basis, at least in part, for standardized due process and judicial approval. In addition, the question of intent towards a particular nation, as in an act of war, might be negated if action was taken towards all system threats automatically rather than being an individual decision against a particular nation. This design and implementation might forestall heightened global conflicts.

The discussion leaves detailed comparative analysis of important legal areas such as jurisdiction and electronic communications surveillance [64] for future discussion, but it may be noted that these issues will be resolved differently in unique legal cultures that address important social goals. For example, the recent OECD study of ISP actions to defeat botnets outlines different approaches of eight countries

---

61 *See* Michael Chertoff, *Foreward,* 4 Nat'l Security L. & Pol'y 1, 5 (2010).

62 *See* Patriots Debate, *supra* note 57, at 123-134.

63 *See* Scott J. Shackelford & Richard B. Andres, *State Responsibility for Cyber Attacks: Competing Standards for a Growing Problem,* 42 Geo J. Int'l L. 971, 985-88 (2011).

64 See for example, Legal Implications of Countering Botnets, *supra* note 29 (comparing in detail the statutory provisions in Estonia and Germany, for example).

and notes that future international cooperation will require development of communication between different participants, governmental or ISP.[65] It is highly likely that an immune system design would be different from nation to nation and that communication between systems and nations would be essential.

5. CONCLUSION

The adoption and implementation of a cyber immune system is not an easy technical task; in comparison, the thorny legal and ethical issues across global boundaries are equally daunting. While the automated nature of a cyber defense may present legal challenges to both state and non-state actors, perhaps it can also mitigate the legal ramifications if the system of rules is carefully crafted. The design of the technical system and its implementation should not only secure cyberspace, it should also incorporate legal and ethical principles that will preserve the essential values of a democratic system that are enabled by features of Internet communications.

[65] Id.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Towards a Cyber Common Operating Picture

Gregory Conti
Cyber Research Center
United States Military Academy
West Point, New York, USA

John Nelson
Cyber Research Center
United States Military Academy
West Point, New York USA

David Raymond
Cyber Research Center
United States Military Academy
West Point, New York, USA

Abstract: Commanders enjoy a refined common operating picture of the kinetic battlespace. While still imperfect, today's military command posts represent centuries of refinement and maturation enhanced by cutting-edge technology. Cyberspace's emergence as an operational domain, however, presents unresolved challenges to this status quo. Techniques for maintaining situational awareness and command and control of cyber operations, particularly joint cyber/kinetic operations, are ill-defined, and no current solutions provide military decision-makers with a comprehensive cyber common operating picture, or CCOP. This paper provides a framework for designing such systems. We focus on the problem of cyber-only operations as well as joint cyber-kinetic operations. Our analysis indicates that the CCOP problem is tractable, but non-trivial, requiring substantial effort realized through evolutionary and revolutionary research approaches.

Keywords: *cyber operations, cyber COP, cyber Common Operating Picture, CCOP, cyber situational awareness*

1. INTRODUCTION

Cyberspace's emergence as an operational domain challenges military organizations' current ability to provide commanders with enough critical information to lead operations involving cyberspace. This challenge rises from the inherent differences between kinetic warfare and combat realities in cyberspace. The days of a battlefield commander sitting in an operations center receiving staff briefings, which took hours to prepare, to make a handful of decisions that will take hours or days to execute, are anachronistic in the cyber warfare era. Unlike nuclear missiles, which take about 30 minutes for global transit, leaving time for hurried human decision-making, network packets take milliseconds. Thus, distance and reaction time approach zero in the cyber domain. Therefore, a cyber Common Operating Picture (CCOP) system that provides situational awareness despite cyberspace's largely opaque nature, enhances a leader's ability to make quicker critical decisions, and leverages automated responses that can operate at machine speeds is essential. Absent a CCOP, leaders are effectively blind to an entire operational domain where adversaries coordinate, operate, and hide. Significant advantage has historically gone to militaries that more effectively apply new technologies. Cyberspace is no different.

A CCOP's design is complex and must allow monitoring of the physical and virtual battlespace and provide actionable information. To prevent operator overload, such systems provide tailored and timely information at each military echelon. However, operators are not just passive observers of the battlespace, but are active participants, and the system must facilitate automated and manual Command and Control (C2) of kinetic and cyber forces. This paper provides a framework for the design of CCOP systems. Thus, we provide necessary underlying contextual information unique to the military domain as well as critical analysis of potential approaches. We do not claim an ultimate solution to this significant problem; we do, nevertheless, contribute a novel analysis of the problem space and a framework to inform future work.

We define *cyber* as the combination of Computer Network Attack (CNA), Computer Network Exploitation (CNE), Computer Network Defense (CND), and Global Information Grid Operations. Note that we explicitly omit the cognitive domain, i.e. information operations, but acknowledge that future CCOP systems will likely pursue this extension to parallel emerging military doctrine. We define *common operating picture* and *situational awareness* using U.S. military doctrine. A *COP* is "a single identical display of relevant information shared by more than one command that facilitates collaborative planning and assists all echelons to achieve situational awareness." *Situational awareness* is the "the requisite current and predictive knowledge of the environment upon which operations depend —

including physical, virtual, and human domains — as well as all factors, activities, and events of friendly and adversary forces across the spectrum of conflict." Finally, *Battlespace* is an extension of the notion of the ground battlefield, to include air, land, sea, space, and importantly, cyberspace [1].

This paper is organized as follows. Section 2 places our research into the field of related work. Section 3 discusses the challenge of linking cyberspace and kinetic warfighting operations. Section 4 examines techniques for complementing visualization with machine processing. Section 5 analyzes key facets of a CCOP system's design. Section 6 provides our conclusions and suggests directions for future work.

2. RELATED WORK

Important related work surrounds the creation of a CCOP, including work in network monitoring, intrusion detection, incident response, security visualization, and military command center design. This section highlights the work most germane to this paper.

Command centers began transitioning from physical map and acetate overlay to computerized displays in the 1990s. Military doctrine and technology have since significantly improved. For example, the U.S. military updated its doctrine to include significant coverage of visualization and COP concepts, but only in the physical, not cyber, battlespace [2]. In terms of technology, the U.S. Army's blue force tracker system Force XXI Battle Command Brigade and Below (FBCB2) is representative of current systems that use GPS data to place military units on map-based displays. FBCB2 will upgrade into the Joint Battle Command - Platform (JBC-P), which provides mobile C2 and improved network communication capability. Tactical and operational command posts use Command Post of the Future (CPOF) to provide the battlespace COP from battalion- to division-level. CPOF provides a suite of tools for collaborative, real-time, multi-echelon C2. At the strategic and operational levels of war fighting, systems such as the Global Command and Control System (GCCS) provide a common operational picture including friendly and enemy status information. Other systems, such as the Advanced Field Artillery Tactical Data System (AFATDS) provide automated support for planning and controlling kinetic weapons, and other systems such as the Battle Command Sustainment Support System (BCS3) support logistics functions. Many deployable, hardened systems can survive austere environments, but require the space and consistent power of a command post or military vehicle; some systems, however, are battery-operated, handheld devices for battlefield usage, such as the Forward Entry Device (FED), linking artillery observers with fire support. Current systems represent the state-

of-the-art in kinetic warfighting for situational awareness and for commanding weapon systems and subordinate units, but importantly, do not extend to the cyber domain.

Computer network monitoring does indeed occur in government and industry network operating centers, primarily designed to monitor network operation, and to a degree, to detect and defend against cyber-attacks [3]. They possess limited physical domain awareness, are primarily defensive, and lack offensive capabilities.

The speed with which decisions and actions must occur in cyberspace operations will increasingly surpass human capacity and already requires automated approaches. Consider the Defense Advanced Research Projects Agency's newly announced Plan X program. While limited details are available, the program seeks "revolutionary technologies for understanding, planning, and managing cyberwarfare in real-time, large-scale, and dynamic network environments." Plan X emphasizes "visualizing and interacting with large-scale cyber battlespaces" and envisions "hardened 'battle units' that can perform cyberwarfare functions such as battle damage monitoring, communications relay, weapon deployment and adaptive defense." [4] Still in its genesis, research generated by this program will be germane to CCOP development.

Existing visual analytics tools may be integrated into a future CCOP system. Representative examples include IBM's Analyst's Notebook, which translates disparate information into actionable intelligence; Palantir, which fuses data from diverse data sources into a unified model to accelerate analysis and harden defenses; HP's ArcSight, which provides visibility into enterprise-level IT infrastructure; and Splunk, which allows multiple data source analysis, including logs, configuration files, and alerts; as well as the products of the start-up PixlCloud, which employ cloud resources to visualize and understand big data [5,6,7,8,9].

Academics are also developing visualization techniques suitable for potential CCOP integration. A full description is beyond this paper's scope, but we recommend studying the proceedings of the Symposium on Visualization for Cyber Security, the IEEE Visual Analytics Science and Technology Conference, the ACM Conference on Computer Supported Collaborative Work, and IEEE Information Visualization, for historical and emerging ideas. In addition, Conti's *Security Data Visualization* and Marty's *Applied Security Visualization* provide useful overviews of design techniques and insight into candidate visualization technologies [10,11]. Many of the visualization and interaction techniques useful for a CCOP exist today, but must be carefully integrated into a seamless system designed around large scale, potentially highly-automated, cyber warfighting needs.

Visualization is only part of a CCOP system, which also requires automated decision-making and analysis techniques. Butler suggests using decision analysis for cyber

operations, which could be integrated into hybrid human-machine or machine-only cyber operations decision-making [12]. Butler's solution, or similar higher-level analytics, would likely become critical components in a CCOP system. In addition, as the future portends friendly algorithms fighting against enemy algorithms in the cyber battlespace, we suggest exploring Wall Street's high-frequency trading for important insights [13,14]. Finally, Boyd's classic work on decision-making and OODA loops might illuminate the dynamics of cyber warfare operations, particularly regarding human and machine cognition [15]. The CCOP must enable the user and the machine to cycle through the OODA loop faster than adversaries.

Our work's novelty springs from the gap between the robust military technology—excellent at tracking and issuing commands in the physical realm, but lacking cyberspace integration—and telecommunication industry systems, which monitor networks, but are unable to plan cyber operations, particularly if large scale and offensive in nature. A CCOP solution demands convergence and integration, but not all the required pieces exist today. Filling these gaps is the role of the CCOP systems we propose.

3. LINKING CYBERSPACE AND KINETIC OPERATIONS

The physical world and cyberspace differ dramatically. Geographic regions define the physical world, where military operations are divided into sectors of responsibility. Cyberspace is a manmade network whose components reside in physical space, but which is a complex and constantly evolving dynamic system modifiable by computer code. Minutes, hours, days define physical world's time. Cyberspace components can operate in milliseconds or less. For example, network packets travel near light speed, and computer code is executed by commodity processors at billions of operations per-second. The military marks physical world distance by meters and kilometers. Cyberspace distance effectively approaches zero; the time-space differential is nearly negligible. Humans are slow, easily tire, and error-prone, but possess ingenuity. Computers can manipulate symbols for years and rarely make errors, but only on algorithmic problems. For additional discussion on these topics consult Miller's work [16].

In the land domain alone, military operations are incredibly complex, requiring a thorough understanding of enemy and friendly disposition, the current mission, and an executable vision. Maneuver, artillery, reconnaissance, and air defense activities must be deliberately synchronized with intelligence, engineer, communication, military police, and other supporting units. Modern U.S. military doctrine includes early steps toward integration of "soft" force, including information, psychological,

and civil military operations, to influence the adversary and civilian populace. As the operation unfolds, forces seek to answer leaders' information requirements, take risk reduction and force protection measures, follow rules of engagement, and minimize negative environmental impacts. As casualties occur, supplies deplete, and systems break, force sustainment activities help maintain maximum operational potential. Simultaneously, signaleers seek to maintain reliable and robust communications [17]. Even the best plans, however, rarely survive initial enemy contact; all leaders—both friendly and adversary—must adapt. The result is Clausewitz's "fog of war," where combatants must make decisions with limited information while solving ill-defined problems, with limited time, and lives at stake [18]. Air, sea, space, and cyberspace operations are similarly complex and uncertain. To illustrate this complex environment, we offer the model in Figure 1, which demonstrates how cyberspace crosscuts the physical domains of air, land, sea, and space. While not an operational domain (in U.S. Military doctrine), we propose a second crosscutting plane for the electromagnetic spectrum, which acts as a substrate for some aspects of cyberspace. The CCOP's overarching objective is to link these domains in time and space into a single operating picture.

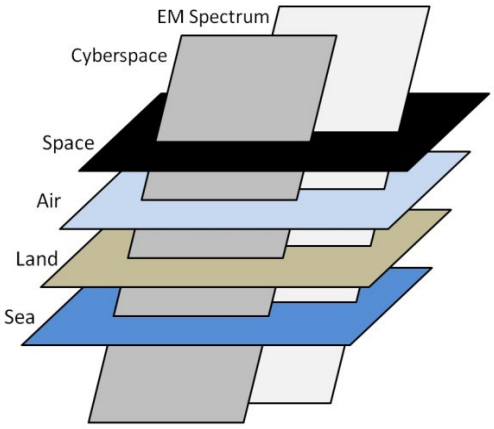

Figure 1. Cyberspace is unique among operational domains because it is manmade and crosscuts each physical domain, akin to a parallel dimension.

4. COMPLEMENTING COP VISUALIZATION WITH MACHINE PROCESSING

Visualization helps clear the proverbial fog of war. Carefully designed visualizations create windows onto information supportive of decision-makers by tapping into

humans' high-bandwidth visual-recognition capacity. Visualization systems are far more than the graphical pie-and-bar charts found in office application suites. They are inherently interactive, contain carefully-crafted displays, and help users efficiently accomplish complex tasks. However, they are not the complete solution. Visualization systems tightly integrate humans into the loop, but while such systems enhance human decision-making, they still are significantly constrained by mankind's weaknesses. Over time, we anticipate the reduced utility of visualization systems alone because human intelligence and perceptual capabilities are constant, computer displays grow at a linear rate, but data requiring analysis has exploded exponentially. A scalable solution is to assign complementary CCOP tasks to human operators and machines, treating each as an integrated system. The right balance is critical. Human processing is in short supply and by nature limited in performance, so humans must perform their specialties (primarily pattern detection, analysis, and creative interpretation) and machines must operate as designed (speedily, accurately, and tirelessly operating on symbols). The best solutions will come from humans' developing insights using visualization and then employing tools to structure this insight in ways that allow computers to do the bulk of future work. The reverse is also possible: machines can alert humans to information that requires human interpretation, see Figure 2. Think, for example, of malware analysts creating antivirus signatures. The signatures can then be automatically distributed across the entire enterprise antivirus system.

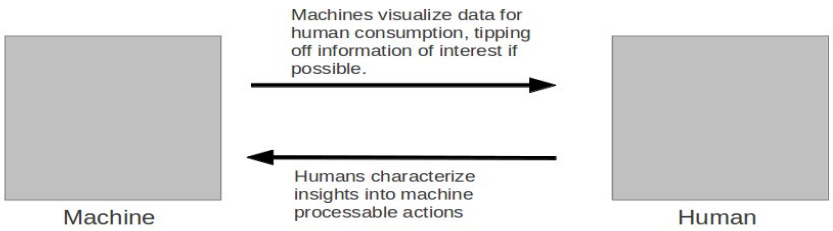

Figure 2. In a CCOP, humans and machines are complementary, tapping into mankind's high bandwidth visual processing system and applying the machine's tireless ability to follow algorithmic instructions.

5. GUIDING THE DESIGN

An effective CCOP system's design requires a deep understanding of system users and their operational environments. An understanding of user tasks, available data, and the available technology's capabilities is also crucial.

A. CCOP USERS

Military organizations typically operate in three echelons: tactical-level (corps and below), operational-level (theater), and strategic-level (national), each with varying missions, capabilities, and areas of responsibility. Tactical units maintain smaller sectors of responsibility and are often directly engaged with enemy kinetic forces. Tactical forces are usually younger, composed primarily of enlisted personnel, warrant officers as technical experts, and officers serving as generalist leaders. The tactical battlefield is often austere, stressful, dirty, with scarce resources, including limited power and network bandwidth. Tactical units are nomadic, reducing the ability to improve their environments. In contrast, operational-level units maintain much larger sectors, often nation-state or larger. Operational headquarters are typically well-developed, fixed locations, and the human dimension includes more senior personnel as well as military, civilian, and foreign representatives from myriad organizations. Strategic headquarters, often located in urban settings, rarely deploy and enjoy easier access to high quality and reliable power, significant bandwidth, and other crucial resources.

A near-future CCOP system may have users trained primarily as kinetic soldiers, with little understanding of cyberspace. The ability to code will initially be uncommon. However, military technologists with cyber warfare expertise will be increasingly common; they will operate CCOP systems and act as intermediaries who translate technical matters for non-technical audiences. Coding skills will thus increase, but some users will possess only general IT and sysadmin-like skill sets. CCOP system products, such as reports, will be consumed by primarily kinetic decision-makers up to the general officer-level, who will likely have minimal technical experience. Military operations rely heavily on skilled planners, primarily trained for kinetic operations, but who will begin receiving training on integration of cyber effects. These planners will increasingly interact with some CCOP systems.

A CCOP's initial success will be a system that addresses operations only in the cyber domain. However, a primary challenge will be how they seamlessly fuse the physical domain with cyberspace for planning and execution of combined arms operations (artillery, infantry, armor, etc.), joint domain operations (ground, air, sea, and space) with both non-expert (kinetic) and expert (cyber specialist) operators and information consumers. For success, the CCOP system must seamlessly interoperate with existing kinetic military command and control systems. This transparent interoperability is crucial for the system's successful employment in a dynamic operating environment.

But what are the cyber responsibilities, operations, and capabilities mandated at each military echelon, particularly at the tactical level? (for early analysis see

Grigsby, who advocates combined cyberspace and electronic warfare efforts in support of tactical operations) [19].

B. TASK ANALYSIS

A detailed listing of a CCOP's required tasks is beyond this paper's scope. We instead provide an overview of major task areas. At a high level, an *ideal* cyber COP system provides:

- Accurate real-time location (both physical and, where applicable, virtual) and status of cyber and kinetic forces, including friendly, neutral, and adversary.
- The ability to provide machine- and human-based C2 of assigned friendly units throughout ongoing cyber operations.
- Seamlessly integrated displays and processing of information for the air, land, sea, space, and cyber domains.
- Appropriate situational awareness of the environment's tactical, operational, and strategic levels.
- Predictive analysis to anticipate enemy actions and reactions.
- Decision support to help leaders analyze options and make decisions across cyber/physical domain operations.

These objectives are complex and unrealistic in the near term. Many friendly forces, such as special operations forces on covert missions restrict their activities to a closely constrained group. "Need to know" controls on classified information will deny some CCOP users access to important data and create situational awareness gaps. Interoperability issues will frustrate communication between sister services, worse still within multinational coalitions. Adversary forces will actively mask their activities and their intent. Even neutral entities and non-governmental organizations will not necessarily aid, and may frustrate, tracking their activities. In cyber warfare the entire global Internet is a potential battleground; billions of pieces of electronics are potential combatants. Decision-making will occur in multiple forms based on willingness to accept risk, legal constraints, and operational necessity, including humans in the loop, humans on the loop, and purely machine decision-making [20]. Because of the complexity, initial success means accomplishing *some* of the desired tasks, but built upon an extensible and robust framework to facilitate future expansion. Table I provides a high-level overview of potential tasks suitable for a CCOP system [21].

Table I. Partial List of High-Level Tasks for an Idealized Cyber COP System

Maintenance	Generate detailed maintenance data suitable for human technicians and automated diagnosis and repair.
Operational Execution	Coordinate highly-complex cyber and kinetic operations; seamlessly allow integration of offense, defense, and exploitation activities.
Electronic Warfare	Integrate electronic warfare capabilities into operations; control friendly and shape enemy electromagnetic spectrum usage.
Forensics	Import insights from forensics systems, capture relevant forensic data from cyber events, and export it to external forensics applications.
Interoperability	Support secure integration and data exchange with a wide variety of systems, including kinetic systems as well as sister-service, multinational, and interagency systems using open and standardized formats.
Targeting	Enable rapid direction of cyber fires despite agile virtual adversaries. Assist with target set development, deconfliction of targets, and the matching of capabilities to desired targets.
Network Analysis	Provide continuous mapping and rapid understanding of the cyber battlespace, including enemy, friendly, and neutral entities, as well as critical nodes. Support study of network bandwidth constraints as being suitable for desired capabilities and to assist in forecasted analyses based on node and link availability. Suggest network paths based on operational needs. Keep pace with cyber maneuver as friendly and enemy operations unfold.
Mission Analysis	Provide support for cyber military decision-making process, including mission analysis, course of action (COA) development, COA analysis (wargaming), COA approval, and orders production.
Mission Rehearsal	Allow operators to rehearse missions, including phasing, sequencing, and timing, and analysis of projected effects.
Battlespace Visualization	Visualize cyber terrain, including large-scale dynamic networks, ideally in real-time, and facilitate delineation of unit sectors of responsibility in the physical and virtual realms.
War Plans	Development of strategic level war plans is beyond the scope of this paper, but automated integration of war plans into a CCOP system will likely be beneficial and a CCOP system may be useful in developing plans perhaps via wargaming or models.
Identify Friend or Foe	Modern kinetic weapon systems use technology to identify whether entities are friendly or enemy; we envision this capability may be possible with cyber platforms.
Battle Damage Assessment	Provide battle damage assessment to analyze forecasted vs. actual effects, including the ability to monitor physical and informational destruction and modification, as well as collateral damage [22]. Provide mechanisms to feedback learning from operations into future planning and prediction sub-systems.
Rules of Engagement	Assist with compliance of authorized rules of engagement, including alerting when approaching legal and ethical boundaries during the planning and execution of cyber operations.

Order of Battle	Monitor the status of friendly, adversary, and neutral order of battle, including irregulars, insurgent groups, criminal organizations, potential insider threats, as well as nation-state organizations along with associated real-world human identities and virtual personas.
Sensor Management	Manage both physical and cyberspace sensors, including issuing of instructions and extracting data.
Training	Possess training and operational modes that allow operators to employ the same system in exercises, simulations, during individual and collective training, as well as operational engagements, supporting the common military practice of "training as one fights."
Capabilities	Provide database of available capabilities and cyber weapon systems, including cost and estimates of risk in usage. System should facilitate integration of new capabilities, awareness of those in use by others, and an ability to remove outdated capabilities from operational consideration. System should suggest candidate capabilities as part of planning process. Integrate notional capabilities for planning and testing purposes.
Weapon System Deployment	Monitor status of cyber weapons platforms and issue commands either manually or via code to automate execution of some stratagems. This goal includes a requirement to synchronize large numbers of cyber weapon systems with millisecond-level precision.
Resiliency and Survivability	Operate effectively despite attack and under degraded network conditions. Provide scalable, reliable, and guaranteed services under all except the most extreme conditions, utilize local caching of data to operate despite network outages, and possess robust backup and failover capabilities, including redundant, load-balanced systems. If the system does fail, it should fail gracefully and securely.
Deception Resistance	Resist human and machine attempts to deceive or otherwise influence decision-making [23]. The system must resist detection despite aggressive threat reconnaissance.
Deception Planning	Provide support for deceptive cyber operations and activities. See the work of O'Connor for examples [24].
Confidentiality, Availability, and Integrity	Operate securely, protect data confidentiality and integrity, and make data broadly available when needed.
Information Operations	Integrate appropriate data from existing information operations systems and planning.
Defensive Operations	Provide comprehensive awareness of friendly networks' health and welfare, including security policy compliance. Appropriately and timely alert human operators of potential and ongoing attacks. Provide shared warning capabilities with allies. Detect, prevent, and respond to attacks and assist with planning and executing counterattacks and adapting defenses. Provide indications of defense failure and recovery activities. When possible identify and isolate attackers (hardware, software, and human). Assist with performing attribution of attacks, despite use of proxies and anonymization.

Intelligence	Assist cyber, SIGINT and all-source analysis. Monitor indicators and warnings relevant to unit's operations. Assist enemy order of battle development, including information on emerging actors, threat signatures, and important cyber events [25]. Fuse information from sensors and intelligence-related cyber missions.
Decision Support	Present options to the commander or operator. Facilitate crosstalk among other friendly decision-makers in the battlespace. Provide decision-support functionality including information from historical and current missions and predictive analysis, including degree of uncertainty, potential risk, desired effects, collateral effects, and legal constraints, for candidate courses of action. Assist in performing intelligence gain-loss calculus. Allow user to display details on the internal logic used by the system.

C. TECHNOLOGY ANALYSIS

Available technology significantly constrains a CCOP's design, particularly at lower echelons. Cloud-based resources can partially decrease the disadvantages of limited resources near the tactical edge. However, cloud resources, while offering the tactical user reach-back capability, are inherently dependent on network connectivity. When networks fail, which is a common battlefield occurrence, a poorly-designed system is effectively useless. Besides, variations in bandwidth and network reliability at each echelon, processing power, display sizes, electrical power sources, and other characteristics vary dramatically (see Table II).

Table II. Technology to Support a Cyber COP System varies dramatically based on military echelon.

	Processing	Network	Interface	Power	Typical Display Size
Strategic HQ	High – Extremely High	High	Keyboard Mouse	Reliable, with generator as backup	up to wall size displays.
Operational / Theater HQ	Average	Average	Keyboard, mouse	Generator, possible host nation commercial	up to 60"
Tactical HQ	Modest	Modest bandwidth and possibly intermittent connectivity	Keyboard, mouse	Generator, possibly unreliable commercial	up to 42"
Tactical Vehicle	Limited	Limited bandwidth and intermittent connectivity	Touch, keyboard	Battery, generator	up to 15"
Tactical Individual	Limited	Limited bandwidth and intermittent connectivity	Touch, small keyboard	Battery	3" - 15"

As the table indicates, screen size, processing power, and network capabilities vary dramatically. A CCOP system must account for these aspects. A "one-size-fits-all" solution is unlikely; instead solutions tailored for each echelon, which account for available technical platforms and network resources, will likely be the most promising approach. Despite these differences, similar interfaces, software modules, and interoperable data sources might maximize ease of use and minimize coding and training requirements. To ameliorate dependence on network connectivity, caching and localized processing can provide resilience against network or other failures.

Some military units embrace innovation and will likely develop prototype solutions. These systems will illuminate promising approaches for future adoption, but will initially frustrate standardization and interoperability. One potential solution is to create an extensible system that actively supports end-user development, such as custom visualizations using the Ozone widget framework, but provided under an overarching standardization framework [26].

Human and technological limitations will constrain the system's visualization aspects. Visual representation of large-scale data remains an open problem since limited pixels populate even the largest display. However, the ability to zoom and filter combined with higher-level analytics, such as attack trees or decision analysis algorithms, can maximize the limited resource of human time and attention. Systems based on formal methods may increase commanders' confidence. Advances in automated analysis and fusion of text, sound, images, video and other sensor data will increasingly enhance capabilities. Gaming and simulation engines may serve as viable frameworks for integration into a CCOP system and are also intimately familiar to computer gamers in the military.

D. *INFORMATION FLOWS*

A CCOP system relies on its information flows, which can be in a raw form, aggregated, summarized, filtered, anonymized, or combined with other data flows. Transformations might occur upstream, perhaps due to bandwidth constraints, or could occur directly on the system to provide desired insights or prevent user-information overload. However, latency, completeness, and accuracy are constant challenges. Clock drift will cause subtle variations in time-stamped data despite simultaneously occurring events. Data classification will prevent some users from accessing needed information as will data-sharing restrictions among inter- and intra- national and agency partners, including between privately-owned, civilian, military, and government entities.

Internet data collection is particularly pernicious. The Internet is the operational

battlespace, yet simultaneously many CCOP information flows will occur over this same network. Out-of-band communications, such as separate networks for observation and reporting, are expensive, but likely required for critical information flows feeding a CCOP. Importantly, these parallel networks will be high-priority targets and require effective safeguards. As a constantly changing, dynamic system, comprised of billions of computing devices, global, real-time, and comprehensive knowledge of the Internet is an impossibility. The sheer number of states surpasses today's information processing capability and will remain so because increased processing capability spurs the Internet's complexity. However, partial mapping of the Internet's state is possible but time consuming and risky. Packet-based mapping increases detection likelihood and risks unintended impacts on the observed systems, such as crashing a system or triggering automated defenses. Many Internet-connected systems are walled gardens, including social networks and virtual worlds, protected by robust authentication and other means. Others take more extreme measures, creating peer-to-peer distributed networks, which ride over opaque, encrypted channels across the Internet substrate. In these cases, traffic analysis based on message externals may be the only way to garner system information.

The Internet was not designed with attribution in mind. Trust of data should be constantly suspect. Deception is easy and common. Threat, neutral, and friendly forces will mask identities or use traps like honeynets to spoof legitimate systems' characteristics.

Kinetic battlefield and cyberspace sensors are key components of the collection, processing, and dissemination chain. Some information derives from intelligence sources; others arrive from open source intelligence, private industry, and increasingly sensors placed on individual soldiers and weapon systems. Information-sharing agreements are necessary, as are automated transformations to convert data format. Similarly, automated-language translation will be necessary. Adversary data will always be incomplete or contradictory due to counterintelligence activities. Friendly force data will provide a better but also incomplete picture.

The enduring bandwidth problem can be reduced by fusion, intelligent data filtering, and generation of high-level semantic information flows (e.g. alerts) that disseminate critical information. Bandwidth, link length, and uptime degrades significantly at the network's tactical edge. Expensive and unreliable connectivity will exist under the best of circumstances, and CCOP systems must be partially functional despite loss of or degraded connectivity during a cyber conflict.

CCOP systems require significant interoperability. But military services have historically resisted military-wide interoperability in lieu of service-tailored

systems, as have defense contractors, who feel data interoperability threatens vendor lock-in. Designing systems for interoperability will be more efficient than trying to bolt-on post-deployment interoperability. See Sweeney's analysis of Blue Force Tracking (BFT) systems for lessons learned from kinetic systems [27].

E. INTERACTION

Visualization's power derives through interaction. A key tenet from the information visualization community is Schneiderman's "mantra" : "[O]verview first, zoom and filter, provide details on demand," a common and powerful paradigm oft-employed by the best information visualization systems. Static displays alone undercut a CCOP system's power. Many existing operations centers forego interaction with their large-screen displays, which are too often underused for cable news, UAV feeds, a map or two, or maybe a few Excel-derived bar charts. Today, real work generates from the analyst's desktop. Part of the solution thus requires creating systems that spur individual and team interest and use, rather than visitor "eye candy." We acknowledge, however, that one person's fancy graphics may have value when tailored smartly for senior decision-makers.

The ultimate solution presents data in functional ways, at the strategic, operational, and tactical-level, with user-determined success. The CCOP should help users accomplish tasks quickly and efficiently. The system must map data to a visual display smartly and efficiently. Many resort to Excel-class graphics, but much more intuitive and interactive options are available. The visualization research community regularly generates employable precision visualization and interaction techniques, which represent a powerful, largely-untapped resource. Additionally, empowering users to generate their own visualizations using technologies such as the Ozone widget toolkit mentioned earlier and then create Apple App Store-like environments for community-based sharing may prove useful. We also recommend evaluating the efficacy of the CCOP systems using real-world users in laboratory, training, and operational environments to determine the system's overall impact on task completion, error rate, and speed, as well as developing an understanding of system limitations.

6. CONCLUSIONS AND FUTURE WORK

Constructing an effective cyber common operating picture system remains an elusive but surmountable goal. Deficiencies are inevitable for the foreseeable future. A way forward involves step-by-step research at the intersection of cyberspace with other domains: physical, electromagnetic, information, and cognitive. We should then seek seamless integration of these disparate domains, not just cyberspace.

Complete knowledge of even a single domain is unlikely, so future work must focus on developing the sensors, processing systems, and communication networks that provide enough, and the right type of information, at the right time to provide actionable information to support informed decisions by CCOP human and machine users. Throughout this R&D process, user studies based on existing systems must ensure the validity of each candidate solution. Although problematic due to security or competitive concerns, this research data and task analyses derived from studying real-world users should be shared to drive future innovation. Humans, however, are not the complete solution. Whenever possible, we must offload appropriate work onto machines, allowing humans to focus on work humans can best provide.

Soon we will see candidate CCOP solutions from academia, industry, and from within the military. Now, though, a panacea is highly unlikely—most solutions will merely be evolutionary improvements. Purchasers should be wary of far-reaching claims. However, visualization thoughtfully-designed in a way that complements human and machine strengths while ameliorating their weaknesses, bears great promise. We can learn from the mature kinetic warfighting processes and systems refined over the centuries, as well as from major telecommunication providers, and assimilate their best ideas. Gaps remain, but as we outlined, a viable design process to combine these insights and fill these gaps with new solutions exists. Ultimately, the solution will be iterative, requiring constant evolution based on user-feedback and system evaluation in operational environments far removed from the laboratory. The true success of a CCOP system hinges upon trust, acceptance, and adoption by the operators and decision-makers whom it supports.

REFERENCES:

[1] JP 1-02 DOD Dictionary of Military and Associated Terms, U.S. Department of Defense, Oct. 17, 2008.

[2] FM-3 Operations, U.S. Army, Feb. 2008.

[3] “Theater Network Operations and Security Center.” U.S. CIO/G-6, Architecture Community. Available: http://architecture.army.mil/technical-view/tnosc.html

[4] N. Shachtman. “Darpa Looks to Make Cyberwar Routine with Secret ‘Plan X.’” Wired Danger Room Blog, Aug. 21, 2012.

[5] “IBM i2 Analyst’s Notebook.” International Business Machines Corporation. Available: http://www.i2group.com/us/products/analysis-product-line/ibm-i2-analysts-notebook

[6] A. Vance and B. Stone. “Palantir, the War on Terror’s Secret Weapon.” Business Week, Nov. 22, 2011.

[7] “Cyber.” Palantir Corporation. http://www.palantir.com/solutions/cyber/, last accessed Sep. 17, 2012.

[8] "HP ArcSight Security Intelligence." Hewlett Packard. Available: http://www.hpenterprisesecurity.com/products/hp-arcsight-security-intelligence/

[9] "Product Overview." Splunk corporation. Available: http://www.splunk.com/product

[10] G. Conti. Security Data Visualization. San Francisco: No Starch Press, 2007.

[11] R. Marty. Applied Security Visualization. New York: Addison Wesley, 2008.

[12] R. Butler, D. Deckro, and J. Weir. "Using Decision Analysis to Increase Commanders' Confidence for Employment of Computer Network Operations." IO Sphere, Fall 2005.

[13] C. Steiner. Automate This: How Algorithms Came to Rule Our World. New York: Portfolio Hardcover, 2012.

[14] N. Popper. "Searching for a Speed Limit in High-Frequency Trading." The New York Times, Sep. 8 2012.

[15] D. Ford. "A Vision So Noble: John Boyd, the OODA Loop, and America's War on Terror." Create Space, 2010.

[16] M. Miller, J. Brickey, and G. Conti. "Why Your Intuition About Cyber Warfare is Probably Wrong." Small Wars Journal, Nov. 29, 2012.

[17] This paragraph draws heavily upon the U.S. Army's five paragraph operations order format.

[18] C. Clausewitz, On War. USA: Empire Books, 2011.

[19] W. Grigsby, G. Howard, T. McNeill, and G. Buehler. "CEMA: A Key to Success in Unified Land Operations." Army, Jun. 2012, pp. 43-46.

[20] P. Hew and E. Lewis. "Situation Awareness for Supervisory Control: Two Fratricide Cases Revisited." International Command Control Research and Technology Symposium, 2010.

[21] A complete cataloging of tasks is far beyond the scope of this paper, but we suggest studying the "The National Cybersecurity Workforce Framework" developed by the U.S. National Institute of Standards and Technology (NIST). Available: http://csrc.nist.gov/nice/framework/

[22] R. Fanelli and G. Conti. "A Methodology for Cyber Operations Targeting and Control of Collateral Damage in the Context of Lawful Armed Conflict." International Conference on Cyber Conflict (CyCon), Tallinn Estonia, Jun. 2012.

[23] G. Conti, M. Ahamad and J. Stasko. "Attacking Information Visualization System Usability: Overloading and Deceiving the Human." Symposium on Usable Privacy and Security, Jul. 2005.

[24] T. O'Connor. "About Face: Defending Your Organization Against Penetration Testing Teams." SANS Information Reading Room, Dec. 2010.

[25] "Department of Defense Strategy for Operating in Cyberspace." U.S. Department of Defense, Jul. 2011.

[26] "Ozone/Synapse Download Portal." Potomac Fusion. Available: http://widget.potomacfusion.com/main/home

[27] M. Sweeney, "Blue Force Tracking: Building a Joint Capability," U.S. Army War College, Mar. 15, 2008.

Chapter 4.

Cyber Command – Towards Automatic Operations

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Complexity and Emergence in Ultra-Tactical Cyberspace Operations

Jeffrey L. Caton
President
Kepler Strategies LLC
Carlisle, Pennsylvania, U.S.A.
Jeff.Caton@keplerstrategies.com

Abstract: This paper explores how the concepts of complexity and emergence can affect cyberspace operations that occur beyond human perception and intervention, such as automated cyber attack responses. It first introduces the concept of the ultra-tactical as an additional realm of operations in the traditional strategic-operational-tactical framework. The context of this realm is compared to human cognitive processes as well as machine processes used to aid human decision making. Potential biases intrinsic in both processes are identified and evaluated. Factors that contribute to the complexity of cyberspace environment in ultra-tactical time scales are reviewed and the potential impact of emergent events on automated decision making protocols are examined. Futuring methodologies are used to develop feasible operational scenarios which are in turn used to evaluate the benefits and risks inherent in implementing automated responses that operate without human cognitive interaction. Specific focus of the analysis includes determining if automated responses will be robust enough to accommodate the dynamic nature of cyberspace and if they can differentiate adversarial threats from natural emergent behavior.

Keywords: *complexity, emergence, automated response, futuring scenarios*

The views expressed are those of the author and do not necessarily reflect official policy or position of any Department or Agency within the U.S. Government. All information and sources for this paper were drawn from unclassified material.

1. INTRODUCTION

In an October 2012 speech, U.S. Secretary of Defense Leon Panetta [1] warned of a potential "cyber Pearl Harbor; an attack that would cause physical destruction and the loss of life." To guard against such a catastrophe, he called for "common, real-time understanding of the threats in cyberspace" concluding that "after all, we need to see an attack coming in order to defend against that attack." This statement implies, perhaps unintentionally, that such cyberspace operations will follow the timelines of commanders in the traditional physical domains. However, attacks in cyberspace can occur in timescales measured in nanoseconds. This paper explores how the concepts of complexity and emergence can affect such cyberspace operations that occur beyond human perception and intervention, such as automated cyberspace defense and attack responses.

2. THE ULTRA-TACTICAL ENVIRONMENT

General Keith Alexander, Commander, U.S. Cyber Command [2] in his 2012 Congressional testimony highlighted the need for the U.S. military to have a "pro-active, agile cyber force that can 'maneuver' in cyberspace at the speed of the Internet." In his 2013 testimony [3], he mentioned that the inter-agency and international exercise CYBER FLAG "introduced new capabilities to enable dynamic and interactive force-on-force maneuvers at net-speed."

But how does one characterize and codify operations at such speeds? A useful model is one that expands the operational realm of cyberspace—the "network speeds"—as part of a more traditional framework. In this case, let us define the ultra-tactical environment as an expansion of the tactical portion of the traditional tactical-operational-strategic operations model.

Consider a one-second timeframe and some illustrative physical events that occur within it (Figure 1). The time required for this page to be processed from your retina to your frontal lobe is about 25 milliseconds. Light will traverse the Earth's equator in 130 milliseconds, during which time an M-4 carbine projectile will travel about 110 meters. Your average eye blink takes about 350 milliseconds. For cognitive processes, a Chess Grand master will discern danger from an opponent's move in about 650 milliseconds--this value represents an approximate threshold for the ultra-tactical environment [4].

Clearly in the ultra-tactical realm are processes and events that occur well below one second. This includes CPU processing speeds (GHz/nanoseconds), memory access, and hard drive seek times. On the opposite end of scale are macro processes

and events that are well above one second. These include activities that require deliberate cognitive processes for decision making, such as intelligence assessment, course of action development, and at the further reaches, policy development. Thus, the actual implement of cyberspace operations occur mostly in the realm below that which humans can comprehend. Certainly, this is an assertion that motivates many security professionals to develop defensive—and perhaps offensive—tools that function automatically in cyberspace. What implications are there for such automated processes occurring in this ultra-tactical realm?

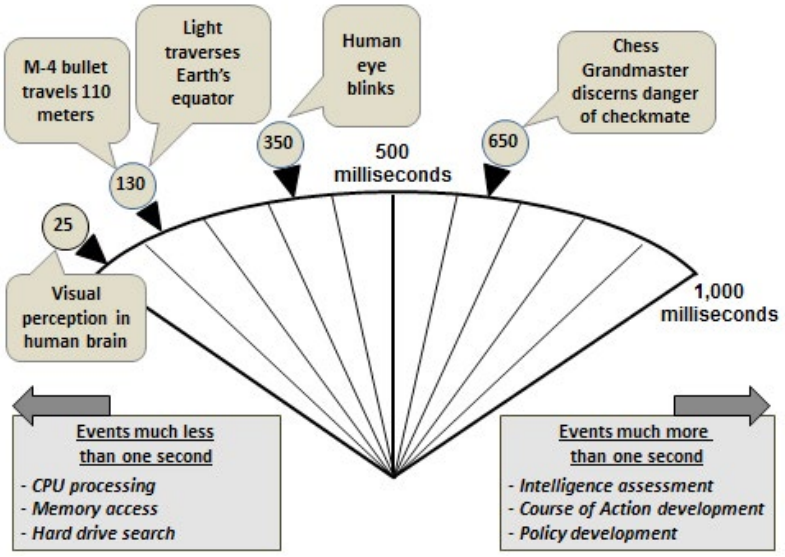

Figure 1. Typical events occuring within one second

3. CONTEXTS OF COMPLEXITY AND EMERGENCE

Geers [5] describes the dynamic nature of cyberspace as an environment where "insurmountable obstacles and golden opportunities can appear and disappear as if by magic." The flow of data occurs across nodes that may exist and disappear within fractions of milliseconds based on internal model attributes that prescribe a desired endstate (such as a software update). He proposes a codification of this phenomenon as one of the ten distinctive aspects of the cyber battlefield framework – specifically, "frequent software updates and network reconfiguration change Internet geography unpredictably and without warning." What are the factors that

contribute to the complexity of cyberspace environment in ultra-tactical time scales and what are the potential impacts of emergent events on automated operations to include decision making protocols?

Czerwinski [6] describes seven basic attributes of complex adaptive systems (properties: aggregation, nonlinearity, flows, and diversity; and mechanisms: tagging, internal models, and building blocks) and he argues that their interactions are fundamental to national security processes and warfare. *Aggregation* relates to the emergence of complex large-scale features from the interactions of less complex agents. *Tagging* facilitates formation of the aggregation by providing agents with traits that can be used for filtering. *Flows* relate to the development of networks among agents that are dynamic in scope as well as in adaption to appearing and disappearing nodes. *Diversity* relates to complex systems creating or fostering communities of agents "marked by perpetual novelty." *Internal models* give systems "the power to anticipate" using two model types: *tacit* which aim for implicit prediction of desired future state, and *overt* for explicit exploration of alternatives.

In sum, one can argue that cyberspace writ large is becoming more like a force of nature than a controlled and predictable network, especially in the ultra-tactical realm. As with the traditional physical domains, what humans are able to perceive and comprehend are manifestations of synergistic trends, properties, and characteristics of an infinitely dynamic environment. What are some of the implications of structure, scale, commonality, and diversity in cyberspace ultra-tactical environment?

A. BLACK SWANS AND DRAGON-KINGS

Emergent events based on models of the micro system dynamics that occur in the ultra-tactical realm may produce macro behaviors through self-organization and synchronization. Sornette [7] studied the dynamics of systems with large numbers of mutually interacting parts, specifically looking for mechanisms of self-organization that may produce surprising emergent behavior at the macroscopic level. In general terms, events that are statistical outliers with novel behavior are often referred to as "Black Swans" which tend to form in regions of self-organized criticality based on the degree of heterogeneity and interaction strength among the parts involved (see Figure 2). They are statistically expected, but not discretely predictable. The concept of Dragon-Kings refers to the existence of transient organization into extreme events that are statistically and mechanically different from their smaller siblings. They may be catastrophic events resulting from the strong coupling of highly homogenous parts in a complex system, and they do not need large perturbations to occur.

Examples of these phenomena are found in natural studies, such as organism networks and ecology in biology; plate-tectonics and erosion in geology; as well as applications in social sciences and economy. Unfortunately, Sornette concludes that "extreme events occur much more often than would be predicted or expected from the observation of small, medium, or even large events." How can this apply to cyberspace operations?

To be prudent, we should address certain ultra-tactical security measures that may drive macro behavior in cyberspace toward the Dragon-King realm. Specific examples include measures that push for increased system homogeneity and predictable interaction, such as: standardized desktops and intrusion detection systems; centralized networks; limited input/output portals; and automated responses. Geer and others [8] argued over a decade ago in their controversial report on Microsoft that use of a "single, dominant operating system in the hands of all end users is inherently dangerous" and that this danger is "exacerbated by tight integration between applications and operating systems." Their methods and findings are consistent with the Dragon-King characteristics of homogeneous systems that are tightly coupled.

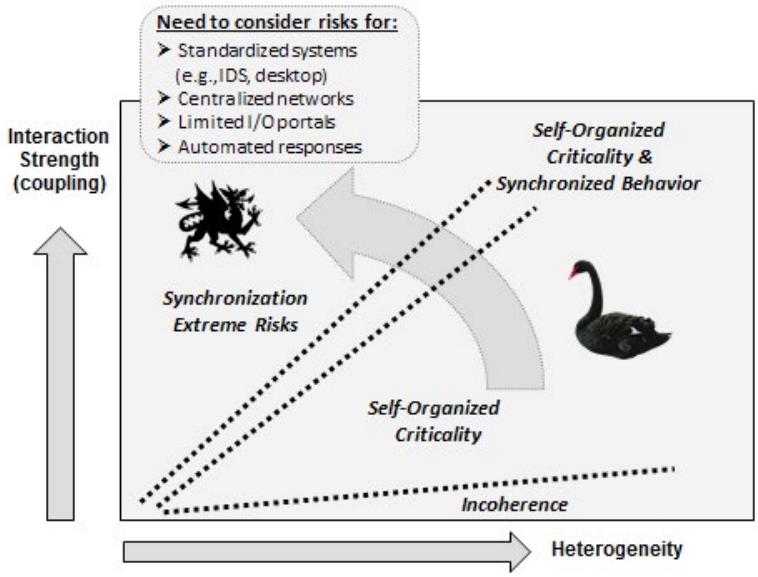

Figure 2. Conceptual emergent behavior

When considering the benefits of activities such as interoperability and cloud computing, we also need to balance the risks. This requires examination of risks posed not only by the threat vectors that these measures may open (or close) to a cognitive adversary, but also those environmental and design threats posed by the self-organization and synchronization they may introduce into the system. Of particular concern are unanticipated and undesired results emerging from ultra-tactical processes to support human situational awareness and decision making.

B. COUPLING IN COMMAND SYSTEMS

Geers [9] characterizes attack and defense in cyberspace as a "game of cat-and-mouse" since over time the opposing forces will develop complex algorithms to counter their foe; these will inherently include some guesswork and miscalculation. But the larger objective of these processes may be to support the command and control of military forces, offering an unwelcome vector of opportunity for emergence from the ultra-tactical realm to drive anticipated behavior in the tradition tactical environment.

Moffat [10] identifies six key properties of complexity important to networked command systems used in warfare. *Nonlinear interaction* can lead to "surprising and non-intuitive behavior;" *decentralized control* can facilitate emergent behavior generated through local coevolution; *self-organization* can occur without external guidance; *nonequilibrium order* means there is never a steady-state; *adaptation* involves clusters or avalanches of local interaction that are constantly being created or dissolved; *collectivist dynamics* reflect the ability of elements to influence each other and cause ripples effects throughout the system. These properities are consistent in principle with the system dynamics and behavior that produce Black Swans and Dragon-Kings.

Another approach [11] is to examine modern military command and control through the lens of the Perrow safety engineering model using two main parameters—the interaction of parts (linear or complex) and the coupling characteristics (tight or loose). Of the four basic combinations of these parameters, systems that are tightly coupled with complex interactions (i.e., those in the realm of Dragon-Kings) are the highest risk. This is due in part to the conflicting operating requirements—that is, control of complex interactions is best decentralized; control of tightly coupled processes are best centralized. So, designing a centralized command and control system for automated cyberspace operations (defensive or offensive) is a high risk venture from both the perspectives of complexity modeling and safety engineering.

C. ULTRA-TACTICAL OPERATIONS GONE AWRY

To better understand these concerns regarding such behavior in cyberspace, consider the 2010 flash crash of U.S. futures and securities markets [12]. On May 6, 2010, major equity indices in both U.S. futures and securities markets suddenly plummeted 5 to 6 percent in a matter of minutes. During this time, over 20,000 trades across more than 300 securities were executed at prices more than 60 percent away from their values just moments before. Many of these trades were executed at prices of a penny or less, or as high as $100,000 – ranges that would not have been approved by rationale humans. Most of these trades were cancelled via formal intervention after the market closed.

One could assert that such trading operations have evolved far beyond the original intent of a stock market to connect investors with capital to prospective revenue-generating venues. Instead, it has largely moved toward making large volumes of purchases and sales to leverage microscopic changes in the perceived value – often with little regard for the long-term prospects of the stock (or market writ large). Osorio and others [13] observed that as early as 2001, the distribution of high-frequency stock market events included autocorrelations in volatilities and volumes caused in part by a herding attitude amongst traders. These effects were magnified as trading became faster and more automated. By May 2010, market dynamics were dominated by automated responses implemented with the willing abdication of the cognitive. Automated algorithms--individually well designed--interacted in such a way as to produce a Dragon-King that dropped market value dramatically. Although the U.S. Government report outlines many contributing factors to this event, no one has been able to determine exactly how it occurred or how to prevent future occurrences. A reasonable conjecture is that the internal models of the algorithms were tacit ones concerned only with immediate profit opportunities with little overt elements to examine alternatives or consider the overall system stability.

Further examination [14] of the ultra-tactical transactions surrounding the “flash crash” uncovered over 18,000 ‘ultra-fast” Black Swan events—either mini-spikes or mini-crashes—that had millisecond-scale durations. In this light, perhaps the proper cybersecurity perspective to adopt is one less worried about a “cyber Pearl Harbor” and more concerned about a “cyber tsunami” or “cyber Super Storm Sandy.”

4. CONTEXTS OF HUMAN COGNITION

Recall that the concept of the ultra-tactical is that of an additional realm of operations in the traditional strategic-operational-tactical framework and that

its discrete processes occur well below the level of human cognitive processes. However, the ultra-tactical processes and their aggregate results may be used to aid human decision making in traditional operational spectrum where human cognition dominates.

A. ENHANCED DECISION MAKING

Figure 3 depicts a full operational spectrum from strategic down to ultra-tactical time scales. At the strategic level, deterrence is practiced based on existing policy; at the operational level, deliberate responses to cyberspace activity reflect doctrine and planning; and at the tactical level, more immediate deliberate responses are based on tactics, techniques, and procedures. In the ultra-tactical realm, automated responses are based on a priori design. Anticipating that these designs will be standardized and coupled, it may also create a breeding ground for Dragon-Kings as well as a quandary for implementing either centralized or decentralized control of the processes.

But, within this spectrum, where and when should human cognition be engaged to enhance the overall process? Risky emergent behavior is possible at any level, but the time scale to address any emergence increases in the ideal case; that is, strategic issues may have a greater luxury of time for examination and policy may be broad to allow flexibility in application. When implementing automated responses we have willingly abdicated the option of cognitive processes based on what we think may occur. In doing so, we must ensure these responses can differentiate adversarial threats from natural emergent behavior and that they are robust enough to accommodate the dynamic nature of cyberspace.

Tyugu [15] examines the use of command and control agents in cyber warfare, noting the trend toward increasing use of automatically operating entities, with one critical factor being the speed of automatic decision making. Regarding the command and control of intelligent agents (i.e., ones that have some independence) he notes that their behavior is harder to predict due to possible misinterpretations of the situation, the command, and priorities. These agents may operate autonomously oriented toward a goal using a beliefs-desires-intentions framework, perhaps following a tacit internal model focused on a desired state vice examining alternatives. This situation may be exacerbated in multi-agent systems, with a specific threat being the "formation of unwanted coalitions by agents," an outcome consistent with the adaptation and collective dynamics of Moffatt. Klein and others [16] have explored initial frameworks to react to detected attacks (such as denial of service) using automatic responses, hoping to improve speed and reliability.

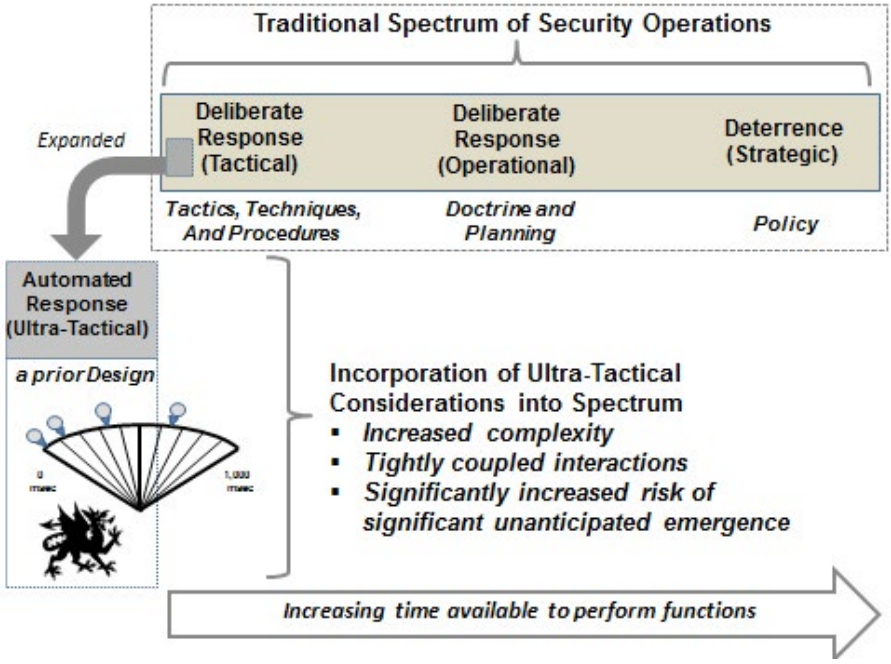

Figure 3. Operational spectrum with expanded ultra-tactical events

Accomplishing this requires information from a diversity of resources passed along many paths—definitely a useful opportunity to apply ultra-tactical processes, but the design should also guard against emergence in the response options that are generated.

B. POTENTIAL BIASES

A crucial part of building confidence in the design of ultra-tactical processes is to fully consider and mitigate the consequences of unchecked cognitive biases in their design. To add further challenge, cognitive bias encountered during design or operation introduces the dilemma of actual versus perceived reality. For example, aural and optical illusions exploit shortfalls in cognitive processes, sometimes to the degree that you cannot force your perception to recognize the reality once the illusion is revealed. For example, the McGurk effect demonstrates how human perceive different sounds from identical sounds under different visual references of human mouth movements [17]. Such mechanisms are more than mere parlor tricks; fully understanding these phenomena is crucial to achieving objective and insightful situational awareness during both the design and operation of cyberspace systems.

Since all the activities in the spectrum are developed a priori to some degree, they are all sensitive to changes in the dynamic cyberspace environment. MacNulty [18] has examined how values, cultures, and beliefs relate to mental models and perceptions across the spectrum of conflict. Because of different value systems, individuals and groups may not perceive the world in the same way and therefore may not respond to communications, hardships, and crises as predicted. Tyugu [19] notes that many human factors influence the development of command and control models. These factors (such as intent, rules & constraints, roles & responsibilities, and situational assessment) could introduce significant biases into the development of automated agents operating in the ultra-tactical realm. Geers [20] included as his ninth aspect of the future cyber battlefield that "the intangible nature of cyberspace can make the calculation of victory, defeat, and battle damage a highly subjective undertaking." Indeed, it will remain a challenge to define the beliefs-desires-intentions parameters for intelligent agents that will operate effectively and appropriately in both the desired future environments as well as potential alternative futures. Realizing that there is no way to predict the complex future, how can one evaluate ultra-tactical processes in future situations?

5. FUTURING METHODOLGIES

Futuring methodologies can develop feasible operational scenarios for use in evaluating the benefits and risks inherent in implementing automated responses that operate without human cognitive interaction. Clearly, there exist many probable futures to consider for the given spectrum of cyberspace activities. These futures will have various degrees of dynamic activity, but at the ultra-tactical scale, all will deal with a cyberspace environment that is constantly changing. Thus, merely applying a tacit model of linear or even exponential extrapolation to define a discrete future has limited applicability. Instead, it is useful to develop an envisioned future scenario without the constraints of having to plot a logical path to its existence (a potentially fruitless situation given the nature of complexity and emergence).

A useful tool for assessing future events is to develop sets of future scenarios that encompass areas defined by divergent conceptual axes. Ogilvy and Schwartz [21] offer a simple and effective model for developing sets of scenarios that use deductive logic to build outcome plots—based on two dimensions of uncertainty--that can capture the scope of many possible outcomes. They recommend having diversity in teams that develop scenarios to help reduce individual biases. Of course, implementation requires the commitment of resources and preferably external facilitators.

Figure 4 depicts an example to illustrate the process of constructing a futuring

scenario diagram. The first dimension of uncertainty (the diagram x-axis) addresses the use of automated defenses in cyberspace--at one extreme is use limited to only the military, the opposing end is global use. The second dimension of uncertainty (the diagram y-axis) is the degree to which military cyberspace operations use the Internet—at one extreme is stand-only operations separate from the Internet, the opposing end is operations fully integrated into the Internet.

The axes of the plot form quadrants offering potential situations for detailed scenario development as does the center of the plot in most cases. It is useful to name the quadrants using simple titles that quickly convey the essence of the situation. In our example, the center scenario is called "Status Quo" and could be developed as an extrapolation of the current situation of the presence cyberspace automated defenses in both military and global applications and the partial use of the Internet by military systems. The upper left quadrant is called "Spill Over" since it indicates a situation where only the military has automated defenses with the possibility that the effects of the automation could spill over into the Internet. The upper right quadrant, "Mixed Signals," signifies how the global use of automated defenses and full integration into the Internet makes it difficult to differentiate the effects caused by military operations from those occurring from other sources.

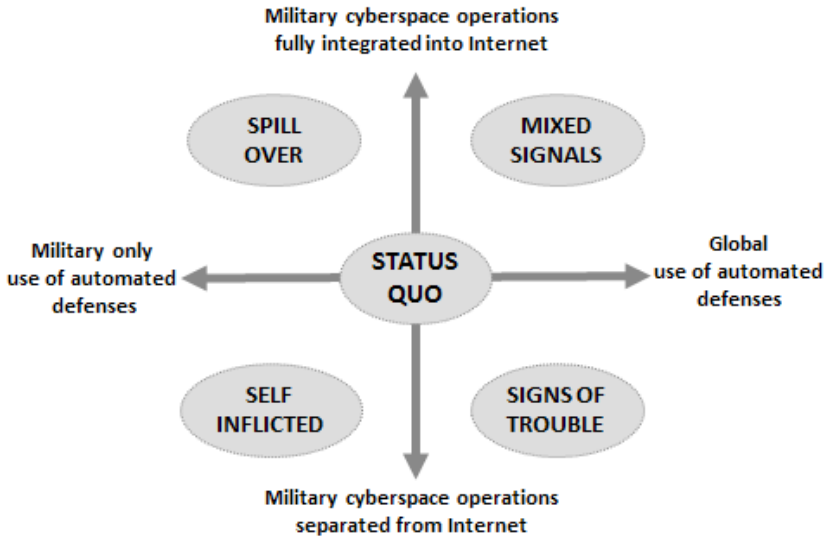

Figure 4. Example futuring scenarios

Since it represents an environment of a diverse community of moderately coupled systems, it would be a likely source for Black Swan events. The lower right quadrant is

called “Signs of Trouble” since the military could observe problems with automated defenses in the global Internet environment, but not be directly affected because its systems are separate from Internet. Finally, the lower left quadrant, “Self Inflicted,” represents the case where only the military uses automated defenses and that these are limited to stand alone systems, thus any problems must be internally generated. Because this quadrant is an environment of largely homogeneous systems that are tightly coupled, it is likely to spawn Dragon-King events.

Once the initial framework of the scenarios is complete, details can be added to better describe the possible future. Each scenario can then be explored to identify possible issues, challenges, and opportunities as well as how they may be addressed, mitigated, or exploited. The more detailed scenarios can then be compared to identify common themes as measures or actions that work effectively in multiple scenarios; these are good candidates for resilient design consideration. This process can be repeated using different dimensions of uncertainty to generate new scenarios. Clearly this is an iterative process that can be accomplished in a collaborative workshop venue. Remember that development, examination, and comparison of the scenarios help provide extensive and robust insight into what *may* happen, not a discrete and limited prediction of what *will* happen. Emergent events (e.g., Black Swans and Dragon-Kings) may also be examined using “Wild Card” scenario methods [22]. The presence of such emergent events can impact the situational awareness of decision makers in the scenarios. For example, Tyugu [23] extended his concerns regarding multiple agent operations into a “Scary Scenario” where very intelligent cyberspace agents may follow intentions and priorities of their own—potentially drawing response from other defensive agents. Such a future emergent event could be viewed as a Dragon-King resulting from complex interactions originally designed for goals quite different from those that emerge, all forming and evolving at potentially ultra-tactical speeds.

A broader value of developing scenarios of alternative futures is their use to assess the vision, mission, and goals for the organization’s desired future [24]. The comparison of these futures may provide insight to weaknesses in the current strategies that can be adjusted to provide a more robust and resilient future strategy. Healy [25] developed five scenarios to examine the future of conflict and cooperation in cyberspace. This included an assessment of the stability and likelihood of these futures occurring. These scenarios could serve as possible starting points for brainstorming dimensions of uncertainty to construct future ultra-tactical vignettes.

6. SUMMARY

Military cyberspace operations—offensive and defensive—envisioned for the near future may make extensive use of automated response processes that occur well below the threshold of human cognition. This realm can be modelled as an ultra-tactical portion that expands from the traditional tactical-operational-strategic spectrum. Complex interactions in this realm will lead to unanticipated emergent behaviour with potentially significant negative effects on planned operations. Current agents designed to operate automatically may be limited to tacit internal models that focus on a desired future outcome and may not consider the alternative futures to reduce risk. Their design may also reflect unchecked biases embodied in the beliefs-desires-intentions objectives of their desired outcome. Futuring scenarios can facilitate the examination of a wide range of possible alternative outcomes that can be incorporated into the development of more robust and resilient processes in the ultra-tactical realm.

REFERENCES

[1] L. Panetta. Remarks on Cybersecurity to the Business Executives for National Security, New York, 11 October 2012.

[2] K. Alexander. Statement before the House Committee on Armed Services, Washington, D.C., 20 March 2012.

[3] K. Alexander. Statement before the Senate Committee on Armed Services, Washington, D.C., 12 March 2013.

[4] J. Caton. "Beyond Domains, Beyond Commons: Context and Theory of Conflict in Cyberspace," presented at the 4th International Conference on Cyber Conflict, Tallinn, Estonia, 2012.

[5] K. Geers. Strategic Cyber Security. Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence, 2012, pp. 103, 109.

[6] T. Czerwinski. *Coping with the Bounds: Speculation on Nonlinearity in Military Affairs*. Washington, D.C.: National Defense University, 1998, pp.7-27.

[7] D. Sornetter. "Dragon-Kings, Black Swans and the Prediction of Crises." Int. J. of Terraspace Sci. and Eng., pp. 1-18, 2009.

[8] D. Geer et.al. "Cyber*In*security: The Cost of Monopoly. How the Dominance of Microsoft's Products Poses a Risk to Security." Computer and Communications Industry Association, Washington, D.C., 24 September 2003.

[9] K. Geers. Strategic Cyber Security. Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence, 2012, pp. 41.

[10] J. Moffat. Complexity Theory and Network Centric Warfare. Washington, D.C.: DoD Command and Control Research Program, 2003, pp. 42-43.

[11] T. Czerwinski. "Command and Control at the Crossroads," *Parameters*, vol. XXVI, pp. 121-132, Autumn 1996.

[12] "Finding Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues." Washington, DC: U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission, Sep. 30, 2010.

[13] R. Osorio, L. Borland, and C. Tsallis. "Distributions of High-Frequency Stock Market Observables" in *Nonextensive Entropy: Interdisciplinary Applications*. Edited by M. Gell-Mann and C. Tsallis. New York: Oxford University Press, 2004, pp. 321-334.

[14] N. Johnson et al. "Financial black swans driven by ultrafast machine ecology." technical working paper, Cornell University Library, Ithaca, NY, Feb. 2012.

[15] E. Tyugu. "Command and Control of Cyber Weapons," in *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 335-341.

[16] G. Klein et al. "Enhancing Graph-based Automated DoS Attack Response." *Proc. 2009 Conf. on Cyber Warfare, NATO Cooperative Cyber Defence Centre of Excellence*, Tallinn, Estonia, 2009.

[17] L. Brancazio and J. Miller. "Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect." *Perception & Psychophysics*. 67(5), pp. 759-769, 2005.

[18] C. MacNulty. "Values, Resiliency & Strategy in Cyberspace." presented at Army War College Cyber Futures Workshop, Carlisle Pennsylvania, 13 December 2011.

[19] E. Tyugu. "Command and Control of Cyber Weapons," *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 335.

[20] K. Geers. Strategic Cyber Security. Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence, 2012, pp. 109.

[21] J. Ogilvy and P. Schwartz. "Plotting Your Scenarios." Emeryville, California: global Business Network, 2004.

[22] J. Dewar. "The Importance of 'Wild Card' Scenarios." Santa Monica, California: RAND, 2009.

[23] E. Tyugu. "Command and Control of Cyber Weapons," *Proc. 4th International Conference on Cyber Conflict*, 2012, pp. 340-341.

[24] "The Future Belongs to Those Who…A Guide for Thinking About the Future." Alexandria, Virginia: Institute for Alternative Futures, 2009.

[25] J. Healy. *The Five Futures of Cyber Conflict and Cooperation*. Washington, D.C.: The Atlantic Council, 2011.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Patterns of a Cooperative Malware Analysis Workflow

Daniel Plohmann
Cyber Defense Research Group
Fraunhofer FKIE
Bonn, Germany
daniel.plohmann@fkie.fraunhofer.de

Sebastian Eschweiler
Cyber Defense Research Group
Fraunhofer FKIE
Bonn, Germany
sebastian.eschweiler@fkie.fraunhofer.de

Elmar Gerhards-Padilla
Cyber Defense Research Group
Fraunhofer FKIE
Bonn, Germany
elmar.gerhards-padilla@fkie.fraunhofer.de

Abstract: In recent years, an ever-increasing number of IT security incidents have been observed, often involving malicious software. In order to cope with the threat posed, it is essential to have a structured analysis workflow for assessment and mitigation.

In this paper, we give a thorough explanation of the malware analysis workflow specified and employed by our team of analysts. It was deducted from observed work patterns and best practices with a strong focus on enabling collaboration, i.e. analyses conducted by multiple analysts in parallel in order to achieve a speed-up. The proposed workflow starts at the point where one or more malware samples have already been extracted. It consists of four phases as a whole, each with its own goals, constraints, and abort conditions.

The first phase aims at gaining an overview of the current situation and specifying goals of the analysis and their respective priorities. The second phase features a preliminary analysis used to sharpen the picture of the threat, using methods of Open Source Intelligence (OSINT) and automated tools in order to obtain a quick assessment enabling first mitigation. In addition, one objective is to facilitate and prepare a more granular dissection of the malware sample, e.g. by unpacking and deobfuscation. The third phase comprises an in-depth analysis relying heavily on reverse engineering of selected parts of the malware. The selection may be influenced by earlier findings or focus on prominent aspects like nesting, functionality, or communication protocols. The final phase builds upon the results of the preceding phases, leading to tailored mitigation concepts for the specimen analysed.

For each of the proposed phases, we give an overview of potential key tools, e.g. helping to gain information or improve collaboration. On a higher level, we highlight challenges to cooperative analysis and our approach to handle them. In this regard, the workflow contains adoptions of principles known from agile software development methodologies. For example, Scrum is used for management of tasks and coordination, aiding the creation of a reproducible and reliable chain of results.

Keywords: *malware analysis, workflow, cooperation*

1. INTRODUCTION

Malware, short for malicious software, is a prevalent tool for digital crime and targeted attacks. Being well-organised, the underground ecosystem around malware constantly unleashes new threats almost entirely aiming at generation of financial gain for miscreants. Additionally, an increasing number of cases including the use of malware for politically motivated espionage and sabotage campaigns [1, 2, 3] have been observed in recent years as well.

The analysis of malware, especially when trying to achieve deeper insights on concrete inner workings, is a time-consuming task. It usually involves notable manual effort which by itself requires significant expertise to be carried out.

In this paper we explain the workflow used by our team of malware analysts, developed on behalf of the German Federal Office for Infomation Security (BSI). The primary goal of this workflow is to speed up in-depth analysis of malware by parallelization of multiple analysts' efforts working in an environment tailored for collaboration. While we are aware of other opportunities for team collaboration, the proposed workflow represents our collection of work patterns and best practices.

Our contributions are the following:

- We identify challenges for the process of cooperative malware analysis
- We propose a workflow designed to overcome these challenges
- We provide an outline of best practices for malware analysis based on our experience

The remainder of paper is structured as follows. Section 2 examines the challenges to cooperative malware analysis. In section 3, an analysis workflow addressing these challenges is described. Section 4 covers related work and section 5 concludes this paper.

2. CHALLENGES TO COOPERATIVE MALWARE ANALYSIS

Being able to effectively conduct malware analysis requires a considerable skill set on its own [4]. This paper focuses on additional challenges posed by close cooperation of multiple analysts working on the same case and how to benefit from joint resources. Useful insights can be taken from the research field of computer-supported cooperative work (CSCW). In this paper, we limit ourselves to two key aspects of collaboration identified by CSCW, awareness and articulation work.

Awareness can be defined as an understanding of the activities of others, providing context for the own activities [5]. Transferred to malware analysis, this can be seen as a need of synchronization of both case specific knowledge and state information among analysts. In consequence, this highlights the importance of a shared space for documentation and proactive signalling of relevant findings. Furthermore, being aware of the individual analysts' proficiencies helps with quick identification of the right contact points in case of specific questions or when delegation of a task is desired.

Articulation work as defined by Strauss [6] is the coordination of lines of work and the interactions necessary for finding work-related agreements [7]. In terms of cooperative malware analysis, this means the breakdown of analysis objectives into manageable tasks that can be processed in parallel. The integration of work results into a consistent product, e.g. documentation in the form of an analysis report or accompanying proof-of-concept code is covered by this as well.

In order to overcome these challenges, we have developed a workflow to handle cooperative malware analysis by adopting components of the Scrum methodology [8] known from agile software development. The three pillars of Scrum are transparency, inspection, and adaption. Internally, transparency forces analysts to create and experience awareness while providing a clear view on the current status of investigation to the outside. Frequent inspection of documentation artefacts propagates knowledge in the team and enables emergence as analysts may encounter artefacts to which they can contribute. Adaption allows controlling progress towards the defined analysis goals. The roles as defined for a Scrum team are treated not as consequently as intended in the concept. The person in charge of keeping contact with a client becomes Product Owner of the analysis. The role of the Scrum Master is taken by different team members on a per case basis. For details on the duties of these roles, please refer to [8].

Scrum as a process management framework has several features that perfectly fit for malware analysis such as being lightweight and flexible. The course of investigation usually has only little foresight, which causes a strong need of short-term decisions based on new findings. Having an iterative progress with incremental results allows keeping focus on superordinate analysis goals and frequently partitioning workload into distinct tasks, thus avoiding plural effort on the same objectives. If a task has been finished by an analyst, it is peer-reviewed by another member of the team. By this, knowledge is shared in the team and a higher quality standard is assured. A typical Scrum task board as used to visualise tasks and their progress is shown in Figure 1, in this example for the analysis of a malware's C&C protocol.

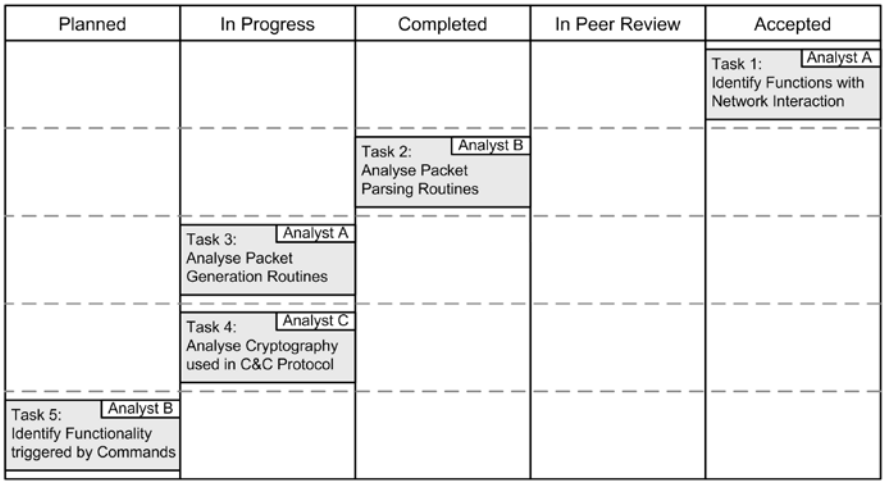

Figure 1. A typical view on a Scrum board as provided by common task management systems.

Tasks are scheduled to be worked on in so called sprints, a time-box framed by an obligatory sprint planning and sprint review meeting. These meetings are a very effective instrument for overcoming the challenges of awareness and articulation work as defined earlier. While planning meetings are used to define which of the remaining tasks are selected for the upcoming sprint, how complex they are and how they should be approached, reviews are used to inspect accomplished analysis work and to gather feedback on it. The duration of sprints should be chosen in compliance with given analysis time frames. After completion, a sprint retrospective is held to enable inspection of the workflow as a methodology itself, in order to successively gather input for possible adjustments and improvements.

To further enable collaboration, one or more tools covering three classes of functional aspects should be available within the workflow. In the following they will be described as distinct services, while all aspects could as well be served by just one tool. First, a documentation system such as a wiki is needed. It should be accessible both by analysts (read/write permission) and all other parties involved (read permissions). The documentation system is used for all kinds of note taking as well as continuous delivery of the analysis report for a case. Second, a tool supporting task management and being compatible with the Scrum-like workflow allows tracking progress. Third, a case or file repository is used as storage for intermediately generated data. It also serves as version control system for own code created in order to support the analysis.

Social aspects of teamwork that may also influence the analysis performance are out of scope of this paper.

3. MALWARE ANALYSIS WORKFLOW

In this section, our malware analysis workflow is described in detail. It consists of four phases. Each has increasing analysis depth and a scope on delivering details towards defined goals. The phases are not to be seen as completely disjoint but with a partial overlap. The workflow as a whole has been designed with the intention to support a thorough and thus long-term investigation of the characteristics of selected malware families but may also prove useful as component in incident management framework. An overview of the workflow is shown in Figure 2.

For the workflow described in this paper, we assume that the starting point of analysis is an already extracted suspicious file. It is assumed that the initial compromise point has already been identified, thus the process of how this file was obtained, e.g. by means of digital forensics, is out of scope for this paper. Note that the workflow also can be easily adapted to cover forensic activities.

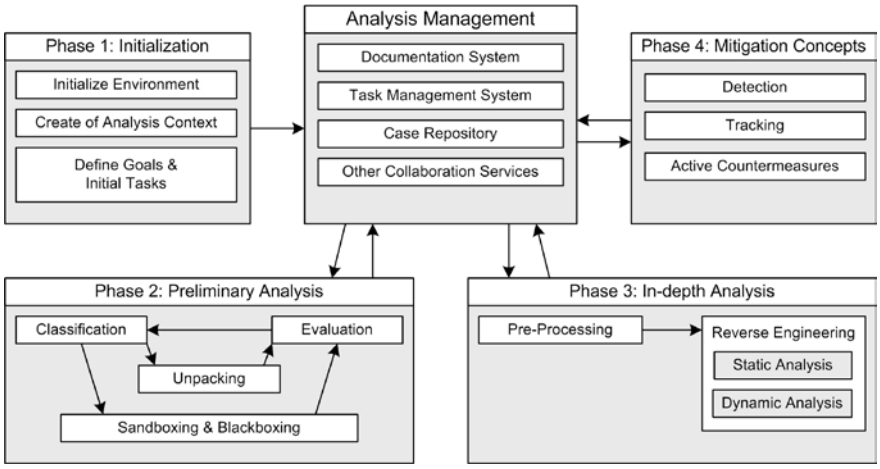

Figure 2. Overview on the proposed cooperative malware analysis workflow.

During all phases and their respective steps, written documentation is produced already alongside the analysis, covering the concrete intention and scope of the analysis activities, a brief description of the procedure taken and its outcome. Additionally, automatable log files such as network traces are always recorded as it is not known whether these are repeatable at a later point in time. Furthermore, the notes taken increase the understanding among analysts involved. The documentation allows assessment of the current analysis progress as well. We use a report template as basis for documentation in order to maintain consistency in our reports. The template may be extended to suffice the needs of special cases.

Where possible, illustrations are used to clarify findings as they can be considered a valuable supplement [9]. The documentation system also serves as a growing knowledge database that can be queried against.

When handling incidents, it is also of interest to collect evidence on whether the malware subject to analysis is part of a targeted attack or the infection occurred rather by chance and is potentially related to digital crime, e.g. through a mass campaign using spam to infect as many targets as possible. A targeted attack is often indicated by the nature of the malware used and may be uncovered by identifying additional compromises with similar malware within the environment of the originally affected system. This investigation is often performed by forensics.

In the following, dynamic analysis will refer to techniques requiring execution of the code subject to analysis, whereas static analysis operates without execution.

A. *ANALYSIS ENVIRONMENT*

Before outlining the workflow, a typical cooperative work environment tailored to our workflow is presented. The structure of this environment shall maximise flexibility. By design it is not bound to static infrastructure in order to suffice the requirements of possible incident response tasks that may be connected to investigations outside of our own facilities. An overview is given in figure 3.

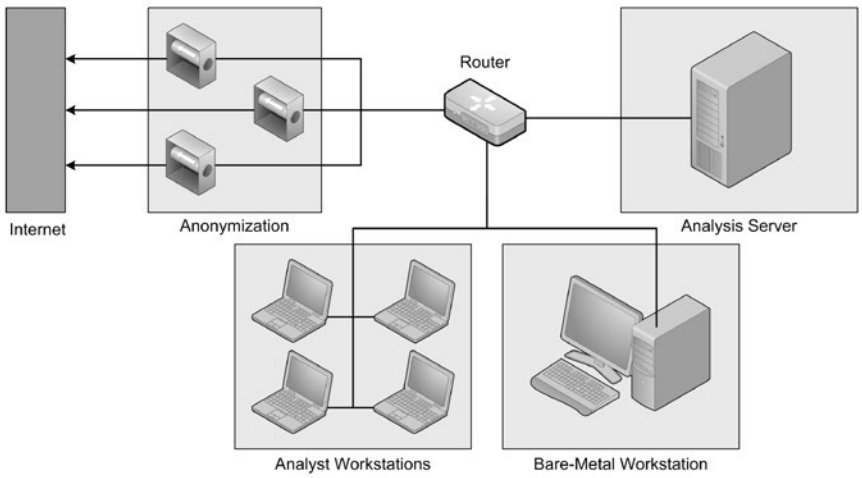

Figure 3. Structure of the work environment the workflow was developed for.

Every analyst uses his own workstation, which usually is a laptop in order to ensure mobility. The main requirement to the hardware is being able to run multiple virtual machines in parallel. This allows simulation of small networks within one workstation. The virtual machine images used for analysis are derived from identical templates per laptop in order to ensure reproducibility across the different hardware instances. Operating systems used for analysis are oriented on the most common variants found in the wild [10], equipped with typical consumer software (e.g. document processors, web browsers) in different patch levels. Furthermore, each workstation has a baseline of similar commonly used analysis tools as needed during the different phases. A template of this basic setup is maintained centrally and can be rapidly rolled out to new devices. Should the malware be virtualization aware to a degree that seriously aggravates the progress, analysts can temporarily shift to bare-metal machines.

The analysts' workstations are complemented by an analysis server enabling collaboration. As described in section 2, cooperative malware analysis faces challenges that can be partly overcome by providing services to the analysts. The analysis server hosts the documentation and task management system, as well as the case repository. In our implementation of the workflow, we use the following three tools to reflect these services:

- Confluence [11], a web-based wiki as documentation system
- JIRA [12], a web-based project tracking tool as task management system
- Stash [13], a web-based Git [14] management software as case repository

If the workflow is more targeted towards incident handling and less for in-depth analysis, the freeware RTIR [15] is a considerable option. Furthermore, special services such as collabREate (see section 3.D) can be hosted on this machine.

All machines are linked by a router, configured as a firewall. Internet Access is provided through a consumer uplink, optionally achieving anonymization through Tor [16] into a VPN provider in order to be able to control the outside origin IP address. This is useful when investigating threats with regional limitations. However, this uplink or permission to use the Internet for research at all depends on the nature of the investigated case.

B. PHASE 1: INITIALIZATION

The goal of the first phase is to set up the framework for the start of a new analysis case. Optimal starting conditions are created for the team by using baselined analysis templates. It is important to continuously maintain these templates during

non-analysis periods in order to be able to start the actual analysis without any delay.

1) Initialization of the Work Environment

During the first step, the work environment is prepared. A new repository is created for hosting files that will accumulate during the work on this case. Depending on the tasks performed, for example this can include one more malicious executables or droppers, memory dumps taken during execution, network traces recorded while examining the malware, short snippets of code aiding the analysis, and related documents obtained during research. In parallel, a new space is created within the documentation system in order to allow collaboration on the documentation to be created for the case. This setup takes at most a few minutes.

2) Creation of Analysis Context

Next, in an initial meeting the scene for the analysis case can be set. The goal is to create context and a roadmap for the team of analysts. The information known on the incident thus far is discussed and inserted in the documentation system. This includes details on when and how the incident was recognised, all data available to the analysts such as potentially relevant files and network traffic recordings as well as the original role of the affected system in order to derive a threat scenario, e.g. a laptop used for travelling, a workstation, or a web server facing the Internet. Points of contact and communication channels are defined in case of questions that may arise. Constraints for the analysis with regard to certain search terms, network access for the malware sample etc. are defined in case of a need for confidentiality.

3) Definition of Analysis Goals and Initial Tasks

The final step for the first phase is a definition of overall analysis goals. Example goals are the extraction of indicators of compromise (IOC) to allow detection on related systems, the identification of remote communication entities or a description of the malware's functionality. The more specific these goals are, the more constrained and focused an analysis can be performed. Each goal should be coupled to an abort criterion, which can be either a timeframe or result and is agreed upon between analyst team and client. At this stage, potential obstacles that may arise during analysis should be outlined. Goals can also be adjusted based on findings later on. From these goals, initial tasks are derived and scheduled in the first sprint planning. The first sprint usually targets to complete the preliminary analysis as defined in phase 2.

C. PHASE 2: PRELIMINARY ANALYSIS

After framing the case in the first phase, a preliminary analysis can be conducted. The goal of this phase is to obtain a first impression of the characteristics and effects caused by the malware to assess its threat potential. This phase also serves as a fundament for all further analysis phases by providing indicators, which can influence steps taken in the following phases. All interim results produced in this and later phases are directly shared with relevant parties such as the security operations centre through the documentation system to allow taking mitigative actions. This is a common good practice as described in [17] and [18]. The second phase of the workflow iterates around three central activities: classification, behavioural analysis, and unpacking to enable deeper analysis. Iterations are closed with an evaluation of the current analysis state.

1) Classification

The step of classification during this phase targets the malware sample as a whole. The focus lies on static analysis in a cursory manner, using tools to automatically extract and assess features. This involves various techniques that can be compared for establishment of a hypothesis on the identity of the malware. In case of a well known specimen, there might be publicly available report coverage on this family that can be incorporated into the case, aiding the speed of analysis. Calculated hashes of suspicious file serve as unique fingerprints. Identification of the file format and examination of its fields (e.g. PE header of an executable) as well as statistical measures such as file entropy constitute the file's outer appearance. Some tools provide detection mechanisms for well-known protection schemes or can help with identifying the programming language or compiler used. Strings and other constants present in the data ascertainable by pattern matching sharpen the picture, especially "low hanging fruits", such as domain names, IP addresses, suspicious registry keys and file names, or similar. It has to be noted that these features are usually not immediately visible due to packing. In this case, a memory dump using a framework like Volatility [19] can serve as a rough approximation to unpacking as detailed later in this section. Furthermore, scans with locally available commodity antivirus software or a matching against the database of online services such as VirusTotal [20] can give hints on the family. Additionally, outstanding data fragments such as very expressive strings can be examined with methods of Open Source Intelligence [21]. However, it has to be kept in mind that such data can always be forged in order to mislead an analyst.

2) Behavioural Analysis

The goal of behavioural analysis is to gain an insight on how the malware affects the system during and after infection. Both sandboxing and blackboxing can be

either performed in parallel to get a perspective from different tools or blackboxing can be used to explore the pointers obtained from sandboxing in more depth. Fully automated sandboxing can give valuable hints on the malware's interaction with the victim system. Results are usually stored machine-readable and accompanied by comprehensive presentation of analysis results. However, sandboxes may lack granularity desired by the analyst. Regardless of the approach, a selection of optionally fake network services should be provided to the analysis system to potentially increase the malware's execution depth by allowing basic network availability checks to succeed. This can be achieved e.g. with InetSim [22]. Nevertheless it has to be kept in mind that the execution behaviour of malware in a sandboxed or blackboxed environment can drastically differ and have more expressiveness if the malware has access to its C&C channel.

3) Unpacking

The third step in the second phase is unpacking, if necessary. Oftentimes, a proprietary protection scheme is used on the malware in order to evade detection by antivirus software and aggravate analysis [23]. However, due to the commercialization of the malware economy, "crypting" of binaries is offered as a service. In many cases the packing layers only wrap the original malware which is then loaded and executed directly from memory, similar to the technique described in [24]. As a result, by intercepting execution at the right moment, the original binary can be recovered from memory prior to its execution. In some cases, the recovery process involves manually reconstructing parts of the original binary e.g. fixing its API imports. Another step of unpacking is removing additional layers of obfuscation if applied to the binary. The recovery of a "clean" unpacked binary is crucial for a success in the following phases and thus deserves special attention. Optimally, the recovered binary is executable after freeing it from its protection schemes. For the further explanation of the workflow, we assume that a binary ready for deep analysis has been recovered in this step and consider details of actual unpacking process as out of scope of this paper.

4) Evaluation

As has been mentioned in the beginning, the three steps are iterated in multiple evaluation stages, as an unpacked malware sample can provide additional insight and can be more accurately classified and behavioural results can be incorporated into classification.

D. PHASE 3: IN-DEPTH ANALYSIS

The third phase shifts the analyst's view from the outside to the inside of the malware, towards the actual code level. The goal of this phase is gaining detailed

understanding of the inner workings of the malware. A solid understanding of these aspects serves as a basis for threat assessment and successful mitigation. In this phase, both static analysis of disassembled code as well as dynamic analysis by debugging selected code fragments are used. The proposed workflow heavily relies on reverse engineering via static analysis. The preferred tool for this type of analysis is Hex Rays' IDA Pro [25] and its ecosystem of extensions and plug-ins such as the Hex Rays Decompiler [26] that can convert disassembly to pseudo code text resembling the C programming language.

1) Pre-Processing

As reverse engineering on machine level usually is very time-consuming, it is worthwhile spending time on a pre-processing step for reducing the expected analysis effort. This can be done e.g. by trying to identify known algorithms imported from library code, recognition of formerly documented functions by matching their characteristics against repositories of already known functions. Another technique is prioritising the analysis on potentially interesting portions of the disassembly, e.g. by extracting the occasions of system API usage, hinting at the higher-level semantics represented by the respective fractions of the code. This is possible because it can be assumed that a compiled binary roughly reflects the structure of its source code [27].

The identification of library code is supported by IDA Pro's FLIRT [28]. Additionally to the shipped modules, further signature databases are available [29]. Another approach for identification is automating queries with constants found in disassembled code against search engines [30], for example Google code.

Similar to the use of publicly available library code, code reuse in malware can be regularly observed as well. This goes without saying for specimen of the same malware family advancing over time. Therefore, the recognition of formerly documented functions is a promising attempt to reduce reverse engineering effort. Zynamics' BinCrowd [31] was such an approach, using the BinDiff technology for a centralised repository, but the product has been discontinued. Out of their own need, CrowdStrike have made the CrowdRE tool and repository publicly available [32]. This service allows the exchange of annotations created with the Hex-Rays Decompiler plug-in, based on both exact and fuzzy matching of functions. As the service is only available in conjunction with the proprietary database provided by CrowdStrike, limitations to application of this service can arise when working on classified cases. Instead of trying to match individual functions against a growing repository, the tool collabREate [33] focuses on keeping annotations consistent among the IDA Pro databases used by instances of multiple analysts. The data exchange happens in real time and thus contributes to synchronization of the analyst's view. A direct transfer of annotations from one IDA database to another

based on function matching can be achieved with BinDiff, which is useful for migrating annotations along consecutive versions of related malware.

Besides the incorporation of existing knowledge into the case, heuristics can be applied in order to increase orientation within the code. IDAscope [34] is an IDA Pro plug-in that aims at helping to identify interesting parts of the binary. By analysing the frequency and usage of selected API calls for all functions, it can lead to semantically interesting locations. For instance, the presence of API calls for network and file access within the same function can be taken as an indicator for download functionality. The plug-in furthermore implements detection of cryptographic algorithms both based on signatures and a heuristic approach as described in [35].

2) Reverse Engineering

After these steps of pre-processing, the further analysis strategy depends strongly on the individual case.

Dynamic analysis is used to complement static analysis wherever it appears to be of avail. For example, observing the generation of system-dependent dynamic values or processing of communication through debugging the respective code fragments can drastically speed up figuring out details of the algorithms used. Debugging can also be used to prove reasoning about functional aspects that may have been made during static analysis. Well-proven debuggers for this task are e.g. OllyDbg [36] and WinDbg [37] on Windows and gdb [38] on Linux operating systems.

In general, there exist three main categories the analysis can be directed towards. They immediately benefit typical goals of the final phase. The three categories are: nesting strategy, functional capabilities including potential spreading mechanisms, and the communication protocols used.

An exact understanding of the malware's nesting strategy allows the creation of tailored detection or even protection methods based upon its indicators of compromise. This is covered in more detail with the description of detection in phase 4.

The in-depth analysis of functional capabilities can serve as a base for threat assessment. Identifying information stealing capabilities gives a clue on potentially exfiltrated data. It is noteworthy that an analysis through reverse engineering can reveal functionality hidden in the binary that has not yet been observed active in the wild or during prior analysis phases, e.g. during sandboxing. The reason for this is that most functionality in malware is triggered by specific commands that have to be given by the actor using the malware. Further functionality of interest can be update or downloading behaviour, potentially widening the degree of compromise.

Finally, understanding the communication protocol and likely used cryptographic routines enables the analyst to decipher traffic generated by the malware or imitate the malware in order to extract information from the C&C entity. Furthermore, this part of analysis may reveal potential backup communication channels that have to be taken into concern when planning countermeasures.

In our workflow, we use both collabREate and CrowdRE for synchronization and data exchange within IDA Pro. While collabREate's real time updates serve as immediate notifications indicating the functions currently being examined by the individual analysts, CrowdRE is used to submit the documentation of decompiled functions after their analysis is finished.

In addition to this synchronization on a technical level, periodic but time-boxed meetings in person or via voice based conferencing software are used to exchange information on the latest progress. This procedure is an adoption of the stand-up meetings known from Scrum [8]. The intervals between those meetings depend on the time criticality of the case, with typically one up to four meetings per day. The meetings are timeboxed with about 15 minutes shared among all analysts in which they report their findings oriented on the following three questions:

- What did I accomplish during the last time box?
- What am I going to do in the next time box?
- What problems may I face in my analysis?

E. PHASE 4: PROVIDING EVIDENCE AND LONG-TERM MITIGATION CONCEPTS

To not have the malware analysis case end in itself, the final phase aims at providing tailored mitigation concepts for the malware specimen. Evidence in form of a written documentation on the analysis case is already available at this point as it has been created throughout the former phases alongside analysis. The mitigation concepts considered for this paper represent only a selection of possibilities and cover the following aspects: detection, tracking, and active countermeasures. Management of tasks connected to the mitigation strategies can again be achieved with the adapted Scrum. It has to be noted that parts of the proposed methods are potentially in conflict with given law in some countries and should be considered by legitimated authorities such as law enforcement only.

1) Detection

As has been mentioned in the description of phase 3, a thorough understanding of the malware's nesting strategy can lead to comprehensive detection methods,

extending a pure signature based approach. In case of deterministic IOCs, reliable detection is possible and even the creation of a temporary tool serving as a virtual vaccine can be considered to protect other machines [39]. Network based detection nowadays is increasingly challenging as it has to be assumed that the traffic generated by malware is completely encrypted. One common product of in-depth analysis of the communication protocol is the identification of C&C entities such as domain names or IP addresses used as point of contact. While being less generic, these serve as primary indicators for network based detection. In some cases protocol characteristics such as fixed ports or characteristic data fragments e.g. caused by use of a static encryption key can be identified that are sufficient for detection. A custom traffic decrypter tailored to the malware is also of high value in case of available full packet captures recorded during the incident.

2) Tracking

A prerequisite to tracking a malware's C&C channel is the understanding of its communication protocol as obtained during phase 3. One possible goal of tracking is being able to monitor the channel for commands, updates, and changes to the C&C infrastructure. Potential use cases are the extraction of templates for spam mails to build better filters [40], get knowledge about announced targets of DDoS attacks [41], or track the evolution of a malware specimen by mining new malware samples and analysing the differences to preceding versions. In case of a distributed architecture (P2P botnet), implementation of a crawler [42] or sensor [43] is an option in order to globally identify infected machines and inform affected parties.

3) Active Countermeasures

As far as active countermeasures are concerned, various options exist. Based on the identified C&C points of contact, an abuse notification to the responsible registrars and hosting providers can be issued. In case of neglect of these notifications, an attempt can be made to orchestrate a takedown supported by a court of law [44]. In case of a P2P-based C&C channel, there may be the option to abuse protocol characteristics in order to achieve a sinkholing effect or partition the network until it is rendered unusable.

The mitigation concepts presented likely require the creation of proof-of-concept or production code in order to be carried out. The utilization of Scrum as process management paradigm in the proposed workflow as well as the structure of the analysis environment support this naturally and allow seamless mixing or shifting of tasks with a focus from analysis to software development. For software development, it should be adhered to known good practices [45]. Especially providing tests for code can serve both as illustrative examples of usage and ensure stability of rapidly prototyped projects.

4. RELATED WORK

In this paper, we addressed challenges to cooperative malware analysis and proposed a structured workflow to overcome them. We are not aware of directly comparable work, addressing both aspects of human collaboration and different progressive phases of malware analysis.

In [46], Wedum describes a systematic approach to malware analysis divided in three phases. An overview of common methodologies used in malware analysis is given by Sikorski in [47] and an explanation of selected tools and techniques with their respective use cases is provided by [48]. In [49], Willems and Freiling give a survey on reverse engineering and countermeasures. Song et al. have developed a BitBlaze, a system for binary analysis [50] usable in the context of malware.

A selection of frameworks for synchronization of analysis results [31, 32, 33] has been already covered in section 3.D.

A comparative survey on automated dynamic analysis systems has been performed by Egele et al. [51]. Blaszczyk discusses automation versus in-depth malware analysis in [52].

5. CONCLUSION

In this paper we gave a thorough overview of a proposed cooperative malware analysis workflow. By adapting selected elements of the Scrum methodology, challenges originating from collaboration such as need for synchronization and work partitioning have been targeted.

A major issue with an efficient workflow in general is the lack of inter-operability of analysis and documentation tools. Oftentimes, the output of analysis tools has to be extensively edited in order to be usable in a report.

Another area with high potential for improvement is further semi-automation of static analysis as described in phase 3. Currently, best practices for recovering details of functionality from binary code are connected to tedious human efforts. Further exploitation of methods for recognising and classifying the structure of the program and its control flow may speed up the analysis by providing better orientation and understanding the relationship of functions and interactions.

There already exist examples of malware specimen, where an effective mitigation can only be derived from a deep understanding of their communication protocol strategy due to the nature of their distributed command and control channels [53, 54]. Thus, we firmly believe it is worthwhile and necessary to research into the

optimization of a workflow for malware analysis, especially targeting families as a whole.

REFERENCES

All websites successfully retrieved on January 09th, 2013.

[1] W32.Stuxnet Dossier. Falliere, N., O Murchu, L., Chien, E. Symantec Whitepaper, 2011.

[2] Protecting Your Critical Assets - Lessons Learned from “Operation Aurora”. McAfee, 2010.

[3] Shamoon the Wiper - Copycats at Work. Kaspersky GReAT, 2012.

[4] The Malware Analysis Body of Knowledge (MABOK). Valli, C., Brand, M. Published by: Edith Cowan University, School of Computer and Information Science, 2008.

[5] Awareness and coordination in shared workspaces. Dourish, P., Belotti, V. In: Proceedings of the 1992 ACM conference on Computer-supported cooperative work (CSCW), 1992.

[6] Work and the Division of Labor. Strauss, A. In: The Sociological Quarterly 26 (1), 1985.

[7] Continual Permutations of Action. Strauss, A. Published by: Aldine de Gruyter, 1993.

[8] The Official Scrum Rulebook. Schwaber, K., Sutherland, J. E-book, published 2011 on: http://www.scrum.org/Scrum-Guides

[9] Syntactic Theory of Visual Communication. Lester, P. Essay, published 2006 on: http://commfaculty.fullerton.edu/lester/writings/viscomtheory.html

[10] OS Platform Statistics and Trends. http://w3schools.com/browsers/browsers_os.asp

[11] Confluence, Team Collaboration Software. http://atlassian.com/software/confluence

[12] JIRA, Issue and Project Tracking Software. http://www.atlassian.com/software/jira

[13] Stash, Enterprise Git Repository Management. http://www.atlassian.com/software/stash

[14] Git – distributed version control system http://git-scm.com

[15] RT for Incident Response. http://bestpractical.com/rtir

[16] Tor Project. https://www.torproject.org

[17] Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. Hutchins, E., Cloppert, M., Amin R. In: The 6th International Conference on Information-Warfare & Security, 2010.

[18] Exploring Security Countermeasures along the Attack Sequence, Sakuraba, T., Domyo, S., Chou, B., Sakurai, K.. In: Proceedings of the 2nd International Conference on Information Security and Assurance ISA, 2008.

[19] The Volatility Framework. https://www.volatilesystems.com/default/volatility

[20] VirusTotal. http://virustotal.com

[21] NATO Open Source Intelligence Handbook, 2001.

[22] InetSim: Internet Services Simulation Suite. http://www.inetsim.org/

[23] Binary-Code Obfuscations in Prevalent Packer Tools. Roundy, Miller, B. Published by: University of Wisconsin, 2012.

[24] Dynamic Forking of Win32 EXE. http://www.security.org.sg/code/loadexe.html

[25] Hex-Rays IDA Pro. www.hex-rays.com/products/ida/

[26] Hex-Rays Decompiler Plugin. http://www.hex-rays.com/products/decompiler/

[27] Recovering the Toolchain Provenance of Binary Code, Rosenblum, N., Miller, B., Zhu, X. In: Proceedings of the International Symposium on Software Testing and Analysis (ISSTA), 2011.

[28] IDA FLIRT Technology: In-Depth. http://www.hex-rays.com/products/ida/tech/flirt/in_depth.shtml

[29] IDA FLIRT Signatures. http://woodmann.com/collaborative/tools/index.php/Category:IDA_FLIRT_Signatures

[30] RE-Google. http://regoogle.carnivore.it/

[31] BinCrowd. http://www.zynamics.com/bincrowd.html

[32] CrowdRE. https://crowdre.crowdstrike.com

[33] CollabREate. http://www.idabook.com/collabreate

[34] IDAscope. http://idascope.pnx.tf

[35] Dispatcher: Enabling Active Botnet Infiltration using Automatic Protocol Reverse-Engineering. Caballero, J., Poosankam, P., Kreibich, C., Song, D. In: Proceedings of the 16th ACM conference on Computer and communications security (CCS), 2009.

[36] OllyDbg. http://http://www.ollydbg.de/

[37] Debugging Tools for Windows (WinDbg). http://msdn.microsoft.com/en-US/windows/hardware/gg463009/

[38] The GNU Project Debugger. http://www.gnu.org/software/gdb/

[39] Using Infection Markers as a Vaccine against Malware Attacks. Wichmann, A. Gerhards-Padilla, E. In: Proceedings of the 2nd workshop on Security of Systems and Software resiLiency (3SL) in conjunction with IEEE International Conference on Cyber, Physical and Social Computing (CPSCom), 2012.

[40] Spamcraft: An Inside Look At Spam Campaign Orchestration. Kreibich, C., Kanich, C., Levchenko, K., Enright, B., Voelker, G., Paxson, V., Savage, S. In: Prcoceedings of the 2nd Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET), 2009.

[41] Active Botnet Probing to Identify Obscure Command and Control Channels. Gu, G., Yegneswaran, V., Porras, P., Stoll, J., Lee, W. In: Proceedings of the Annual Computer Security Applications Conference (ASAC), 2009.

[42] Measurements and Mitigation of Peer-to-Peer-based Botnets: A Case Study on Storm Worm. Holz, T., Steiner, M., Dahl, F., Biersack, E., Freiling, F. In: Prcoceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET), 2008.

[43] Towards Complete Node Enumeration in a Peer-To-Peer Botnet. Kang, B., Chan-tin, E., Lee, C.P., Tyra, J., Kang, H., Nunnery, C., Wadler, Z., Sinclair, G., Dagon, D., Kim, Y. In: Proceedings fo the ACM Symposium on Information, Computer & Communication Security (ASIACCS), 2009.

[44] Bredolab Takedown, Another Win for Collaboration. Williams, J., 2010. http://blogs.technet.com/b/mmpc/archive/2010/10/26/bredolab-takedown-another-win-for-collaboration.aspx

[45] Clean Code, A Handbook of Agile Software Craftsmanship. Martin, R. Published by: Prentice Hall International, 2008

[46] Malware Analysis; A Systematic Approach. Wedum, P. Published by: Norwegian University of Science and Technology, Department of Telematics, 2008.

[47] Practical Malware Analysis. Sikorski, M. Published by: No Starch Press, 2012.

[48] Malware Analyst's Cookbook. Ligh, M., Adair, S., Hartschein, R. Published by: John Wiley & Sons, 2010.

[49] Reverse Code Engineering - State of the Art and Countermeasures. Willems, C., Freiling, F. In: it - Information Technology 54 (2), 2012.

[50] BitBlaze: A New Approach to Computer Security via Binary Analysis. Song, D., Brumley, D., Yin, H., Caballero, J., Jager, I., Kang, M., Liang, M., Newsome, J., Poosankam, P., Saxena, P. In: Proceedings of the 4th International Conference on Information Systems Security (ICISS), 2008.

[51] A Survey on Automated Dynamic Malware Analysis Techniques and Tools, Egele, M., Scholte, T., Kirda, E., Kruegel, C. In: ACM Computing Surveys 44 (2), 2012.

[52] Automation vs. In-depth Malware Analysis. Blaszczyk, A. Essay, published 2012 on: http://www.hexacorn.com/blog/2011/11/21/automation-vs-in-depth-malware-analysis

[53] What we know (and learned) from the Waledac takedown. Williams, J., 2010. http://blogs.technet.com/b/mmpc/archive/2010/03/15/what-we-know-and-learned-from-the-waledac-takedown.aspx

[54] The Lifecycle of Peer-to-Peer (Gameover) ZeuS. Stone-Gross, G., 2012. http://www.secureworks.com/cyber-threat-intelligence/threats/The_Lifecycle_of_Peer_to_Peer_Gameover_ZeuS/

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Architecture for Evaluating and Correlating NIDS in Real - World Networks

Robert Koch
Faculty of Computer Science
Universität der Bundeswehr München
85577 Neubiberg, Germany
robert.koch@unibw.de

Mario Golling
Faculty of Computer Science
Universität der Bundeswehr München
85577 Neubiberg, Germany
mario.golling@unibw.de

Abstract: Research in the field of IT security - in this case especially the Evaluation and Correlation of Intrusion Detection Systems (IDS) - implies special demands for the construction and operation of IT systems. In order to (i) evaluate multiple IDS under absolutely identical conditions and to (ii) check their reactions especially against novel attack patterns / attacker behaviour, all attack related actions (i.e. all traffic) have to be forwarded to all IDS in parallel at real-time. In addition, an attractive target needs to be offered to potential attackers, awaking the outward semblance of real-productive systems / networks including the corresponding behaviour.

In particular, the correlation of IDS seems a promising approach to compensate the individual deficiencies of IDS. For example, while knowledge based systems are only able to detect previously known attacks, anomaly based systems suffer from higher False Alarm Rates (FARs). Even more, periodic performance evaluation studies, e.g., by NSS-Labs, have illustrated that numerous IDS are not configured properly and have a much worse system performance and detection capability than announced by the vendors. However, changing parameters of systems in productive networks (for the correlation of IDS as well as for their evaluation) can result in an enhanced endangerment of the security or even a breakdown of the network in case of horrible misconfigurations.

To overcome these shortcomings, we present an architecture that supports research in the field of IT security and simultaneously ensures that all actions associated with an attack get recorded and a spill over of the attack from the research to the productive environment is prevented. Each test system is supplied with an unaltered live record of the network traffic. This allows an assessment of the detection as well as a comparison of different NIDS concepts/products. In addition, different correlation strategies of alerts of multiple systems can be evaluated. Furthermore, superior configurations can be identified and assessed without endangerment of the productive network.

Keywords: *intrusion detection, optimization, real-world evaluation, comparative evaluation, intrusion detection correlation, test environment*

1. INTRODUCTION

Detecting and Defending attacks against networks and systems is an intense research area for over 30 years. Although a high number of security mechanisms have been developed, for instance numerous proprietary as well as Open Source (Network) Intrusion Detection Systems (NIDS), the situation is not easier. In contrast, the number of attacks, security incidents and malicious software is increasing constantly during the last years. So does the quality of the attacks. Nowadays, attacks are much more targeted and technically sometimes very complex. Even more, attack toolkits which perform sophisticated automated attacks are available on the underground market and can be purchased with Service Level Agreements, guaranteeing that the product will not be detected by todays widely used IDS for a specific amount of time, resulting in multimillion Dollar losses caused by cyber-crime every year.

In order to evaluate the protection mechanisms of existing and newly developed IDS, exposing them to real word attacks seems beneficial. Thus, in order to perform analyses on the reactions of IDS on novel attack patterns / attack behaviours, a secured and controlled research environment - where real attacks are allowed knowingly - is almost indispensable.

Up to now, also modern IDS are often not able to detect sophisticated attacks [1]. This is not only because of new and yet unknown attack techniques, but also because of misconfigurations, erroneous detection engines, etc. For example, studies by NSS-Labs have shown that most systems are configured badly. In addition, it has been demonstrated, that the detection performance in real-word networks by current IDS can be much worse than specified by the producer [2]. E.g., one system was only able to analyse 3 percent of the expected amount of traffic. Besides that, depending on the classification of the IDS, also several shortcomings can be found. For example, while knowledge-based systems are only able to detect already known attacks, anomaly-based systems suffer from higher False Alarm Rates (FARs).

Summarized, the most important real-world problems regarding the use of state-of-the-art IDS are:

- High False Alarm Rates
- Undetectable attacks
- Complex configuration
- Intense administration
- Data Encryption

Therefore, an environment is required which enables a testing and optimization of parameters/configurations as well as an in-depth evaluation of the performance and detection capabilities of IDS without an endangerment of the productive network. Such an environment can also be used for research and development of, e.g., correlation techniques for IDS. Especially with the exchange of attack information between different installations respectively of different autonomous systems spread around the globe and the corresponding analysis/correlation, it is possible to defend against new attack waves. For example, the Internet Storm Center of the SANS Institute collects data from over 500.000 Sensors around the globe. ATLAS from Arbor Networks or the exchange of statistical data by Cisco IPS are other examples for collaboration and the generation of early warnings.

In order to (i) assist the consumers by selecting the corresponding NIDS that best fits into their environment and to (ii) support the combination and correlation of different NIDS (like a anomaly-based with a knowledge-based NIDS), a common environment is needed which provides the network operator with the ability to install more than one NIDS without an influence on the productive network on the one hand and among the different NIDS on the other hand.

Taking into account the idea of evaluating and correlating NIDS, a comprehensive architecture that allows secure and manageable research within a subnet of an - otherwise productively used - network is presented in this publication. Each isolated NIDS is provided with a copy of the sniffed data traffic. In contrast to evaluations of system performances and capabilities with the help of synthetic data-sets like Lincoln Lab DARPA intrusion detection evaluation 98/99 [3], our architecture supports real-world data and real-time detection capabilities for a realistic system assessment. An injection of malicious traffic onto the productive network by the NIDS is prevented and also no modification of traffic during transit is possible.

The operator of the network will have the opportunity to use the proposed architecture on three different modes:

- **Test environment:** The customer implements a productive IDS next to one or more test IDS environments of the same system. These test IDS environments will be used to modify and optimize IDS policies. The customers can verify and validate their modifications without risking harmful influence on the network.
- **Validation of different systems:** For those customers that don't know which product fits their purposes best, the proposed architecture can be used to test different IDS. Thus the customer will be able to validate the IDS implementations against each other based on real traffic. The customers will use this mode in order to choose the best system they will later implement into their network.

- **Correlation of the results of different NIDS:** As IDS will often produce false-positive and false-negative alerts, it would be beneficial to have multiple IDS implemented in order to validate results and to achieve a common operational picture. For this operational picture the evaluation and correlation unit is placed outside the test environments.

The paper is structured as follows: In Section 2, a practical scenario of research on Intrusion Detection will be given. In Section 3, we will discuss related evaluation techniques and approaches as well as requirements for the architecture. Based on that, the architecture of the system will be presented in detail in Section 4. Finally in Section 5, we will present results of a proof-of-concept implementation as well as a case study to show the benefits of the architecture, before the conclusions are drawn.

2. SCENARIO

In this section, the need for a holistic architecture for evaluating and correlating NIDS is illustrated using a practical, real-world scenario (see Figure 1). The special feature of the scenario is the integrative approach of different components; from Intrusion Detection (through multiple sensors) over live analysis (automated correlation of data) to post-mortem analysis (IT forensics). In the following, the individual components are presented:

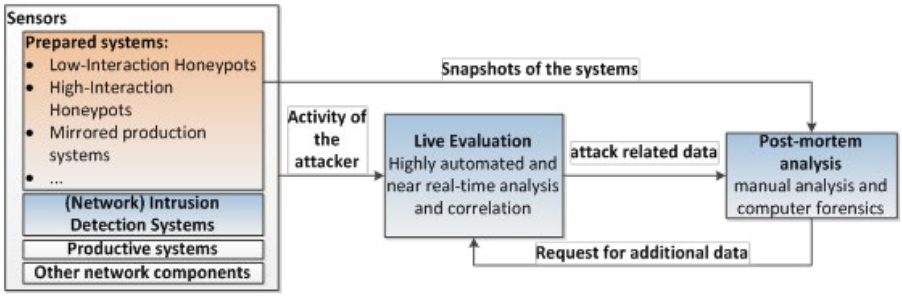

Figure 1. Overview of the components

A. SENSORS

Attackers are attracted with specially prepared systems (clients running Windows XP, Windows 2003 servers as well as low-interaction and high-interaction honeypots) with deliberately (simulated) vulnerabilities in the research environment. The

course and the behaviour of the attacker are recorded multiple times – both by host and network components (host / network intrusion detection systems, honeypots, switches with monitoring port, etc.).

B. LIVE EVALUATION

The sensors are forwarding the recorded data to a central database for analysis and correlation of the activities of the attackers. The alerts generated form the basis for further investigations. The evaluation of the high amount of alerts is first carried out by an automatic correlation. For this, already several approaches are existing, which - however - all have individual shortcomings and needed to be improved [4, 5,6]. Current approaches only consider alerts of IDS, but no additional sources such as Honeypots and log data included.

C. POST-MORTEM ANALYSIS

For the reconstruction of an attack, additional data or snapshots can be requested using specific criteria, such as time stamps, source or destination. This extends the database for forensic examinations. Finally, automated countermeasures can be taken through Intrusion Prevention Systems (IPS).

3. RELATED WORK

A. INTRUSION DETECTION SYSTEMS

IDS can be classified using numerous characteristics, where the most important one is the detection technique. Two different mechanisms are used, namely:

- **Knowledge-based detection**: Detection of attacks by the use of knowledge about malicious events, e.g., by matching a set of known misuse patterns (signatures) against a stream of packets or events.
  - **pro:** low false alarm rates, precise diagnostics
  - **con:** insensitivity to attack variations, difficulty of signature maintenance/ updates, incompleteness of known patterns, huge databases, difficult to reach near-real-time evaluation of network links with high data rates
- **Anomaly-based detection:** Detection of attacks by measuring deviation from statistical models of normality.

- **pro:** detection of unknown attacks, adjustment to traffic/process drift
- **con:** high false alarm rates, anomalies ≠ attacks

Both methods of detection have their pros and cons. Often, knowledge based systems are preferred since they typically provide lower FARs and precise diagnostics: Lower FARs because of the knowledge-based detection technique which produces less false alarms than behaviour-based systems do, which work on models, measurements and thresholds.

The major shortcoming of behaviour-based systems is their high FAR. Especially benign, but yet unknown user behaviour often results in numerous false positives. By that, the number of false alarms can achieve thousands of messages per day in a large network – resulting in an inability to distinguish between true and false alerts.

B. REQUIREMENTS FOR THE ARCHITECTURE

Already a number of general requirements have been derived (e.g. [7, 8]):

- **Security:**
  - *Hidden to the attacker;* in order to analyse the behaviour of an intruder in detail (for instance the exploits used or the steps performed), the presence of intrusion detection tools has to remain completely hidden to the attacker
  - *Multilevel system of interlocking security mechanisms;* despite the careful selection of systems and a consistently focus-oriented security configuration of services, individual security mechanisms can fail (e.g. due to software bugs in the operating system). Through the implementation of additional protective measures at different levels, a multilevel security system still provides protection even if a policy or a device fails.
  - *No influence on productive systems or data streams*
  - *Emergency routines,* e.g. out of band communication to shut down all other communication links
  - *Simulation of a productive behaviour*
  - *Protection of the productive systems*
  - *Recording of all activities*

- **Legal Requirements:**
  - *Prevention of further proliferation of malware and at the same time:*
    - *Minimal and controlled communication* from the research environment to the Internet (especially needed for the analysis of malware behaviour and botnets)
    - *Manipulation of dangerous outgoing data packets*
  - *Consideration of the legal rules and any liability*
- **Scalability:**
  - *Handling X devices, Y users and Z applications*
  - *Evaluation results are independent from data rate*
- **Management challenges:**
  - *Cope with lots of alerts*
  - *Reduce FAR* with the use of intelligent alert correlation [9]
  - *Provide sufficient alert messages* for adequate incident diagnostics
- **Comparison:**
  - Optimization and experimental policies can be tested *without influence on the productive system*
  - *Test behaviour of different IDS based on the same data*

C. EVALUATION CRITERIA

Measuring performance or efficiency of IDS is widely discussed, e.g. Debar et al. give a definition in [10]:

- **Accuracy:** Proper detection of attacks and absence of false alarms. Thus, inaccuracies are anomalies or intrusive traffic flagged as legitimated information.
- **Completeness:** Property to detect all attacks. Without a global knowledge about attacks or abuses of privileges it is much harder to evaluate this measure.
- **Performance:** Rate at which audit events are processed. Real-time detection requires a good system performance.
- **Fault Tolerance:** IDS itself should be resistant to attacks; otherwise new peril points would be opened.

- **Timeliness:** Propagate analysis as quickly as possible in order to react and thus prevent the attacker from harmful malicious actions.

Customers who would like to use the most efficient IDS in their company network have the problem of how to evaluate IDS against the evaluation criteria presented. Therefore, customers have to make a selection of some IDS, implement them, and compare the results afterwards to evaluate the systems concerning their requirements. Several commercial and open-source IDS are available, which make use of knowledge-based or anomaly-based detection methods.

4. ARCHITECTURE

Our proposed architecture divides the network under consideration into several isolated and specialized subnets, namely a Research Network, a Productive Network and an Evaluation Network. At the network border and the border gateway, the datastream is taken of the public network, duplicated one or multiple times and distributed to the different evaluation systems and networks. Test Access Point (TAP) Devices [11], SPAN/Mirrorports and Firewall Kernelmoduls are used for that. Security mechanisms prevent an extravasation from the evaluation networks, e.g., the duplication process is secured by data diodes. By that, only one data direction is possible.

Within the evaluation network, multiple security systems like IDS can be installed, e.g., same systems with different configuration for the optimization of parameters or different systems with complementary detection techniques, e.g. anomaly- and knowledge-based systems. Based on the copied data stream, the evaluation and correlation systems can also initiate active reactions like blocking firewall-ports, generating reports, etc., on dedicated servers therefore not influencing the data on the productive network at all. Also, the results of the different systems can be compared to find the best system respectively configuration for a specific network. Beyond that, the results of different systems can be correlated and, e.g., a majority decision can be taken or more sophisticated correlation techniques can be used or investigated. Figure 2 gives an overview of our architecture. The different components will be described in detail as follows.

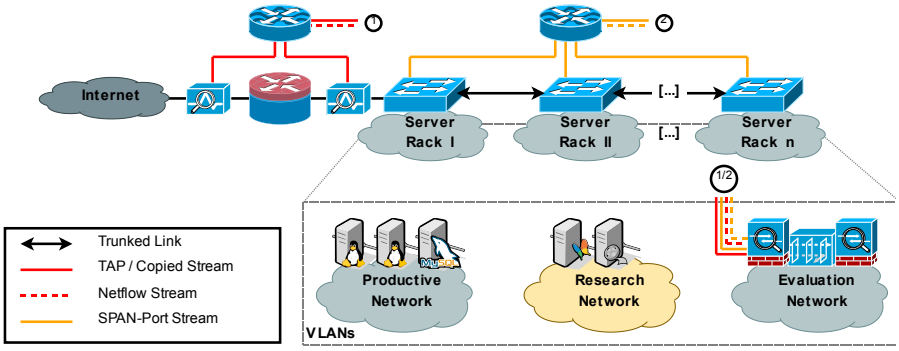

Figure 2. Components of the architecture and integration into company network infrastructure

A. NETWORK BORDER

The border gateway connects all internal networks with the external network, typically an ISP or Internet Backbone. For being able to give basic security against attacks from the external network, a firewall is implemented into the border device. Anyway, because all network traffic is needed for evaluation (not only the filtered one), TAP devices are installed behind as well as in front of the firewall. The TAP devices are able to copy all incoming and outgoing traffic including all Layer 1 and Layer 2 errors. The copied data stream of the TAPs is sent to Cisco Routers which are generating Netflow statistics on the one side; session monitoring is used to multiply the data stream for the evaluation systems on the other side. After the filtering by the firewall, the remaining network traffic is sent through multiple switches which are connected by Trunks/Tagged Ports [12] to the end systems. For complete evaluation and analysis of the internal network, each switch in turn is connected to a second TAP device by its respective monitoring port. The internal network consists of the following parts:

B. RESEARCH NETWORK

This network is used for the examination and evaluation of systems and services relevant to security. To attract attackers and allure malware, honeypots as well as specially prepared real-world servers and services are run in this environment. All systems belonging to the research network can communicate to each other and connections to the Internet are allowed initially. Because of its special nature being open for attacks, only rudimentary or even none filtering is done for this network in the beginning. By an adaption of the rule-set of the firewall, special configurations can be done, e.g., for focusing on individual security aspects.

C. PRODUCTIVE NETWORK

All IT-systems which are required for the operational day-to-day use are set into this network, e.g., office computers, network printers or storage. While the network constraints of this subnet guarantees an usability of the productive network like in the original state without a separation of different networks, multiple security features are activated to protect the corresponding systems:

Switch-based options like fine-grained Access Control Lists (ACLs), Port Security and Community/Isolated port-based VLANs [13,14], which are a basic element for the realization of our architecture. By the use of these elements, combined with a restrictive configuration of the firewall, the productive network can be safely operated in our architecture. Even more, (anonymized) data from the productive network can be mirrored to the research network, therefore generating a more attractive and especially more realistic target environment. Finally, all systems placed in the productive network are monitored by the security systems of the evaluation network, too, enabling an additional protection layer.

D. EVALUATION NETWORK

All network traffic of the complete environment is duplicated for the evaluation network. More precisely, there are even multiple copies of the streams, done by the TAPs in front/after the firewall and the different SPAN-ports of the internal switches. This enables an in-depth evaluation of all network traffic and traffic characteristics, opening up the possibility to build and evaluate complex intrusion and insider detection systems as well as the development of new correlation strategies. The systems of the evaluation network are independent from the systems of other networks, the only receive all data of the other networks but typically without retral information flow because of the TAPs and diodes used. Within the evaluation network, several IDS can be connected or used to exchange evaluation results in order to investigate strategies for reducing the FARs. Based on our architecture, all data as well as corresponding statistics are available for multiple evaluation systems. Even more, the different data streams enable the specialisation of detection systems, e.g., for insider detection, early warning or correlation-based attack detection. Of course, by providing at least the complete external and internal network traffic and the respective flow data, an extensive amount of data has to be processed in the evaluation network. The effective analysis and reasonable storage of parts of the data streams has to be done by the security systems of the evaluation network. Anyway, a central storage of the data is done based on flow data because of the amount of traffic and data protection regulations [15]. Metadata is stored, too; for example, alerts of the different IDS. Based on that, attack sequences can be reconstructed later on.

E. MANAGEMENT NETWORK

For the surveillance of the systems and switches, a separated network is created in our architecture. The systems under surveillance are connected by an independent network interface, which is only used for the management aspects. Also, the systems can be configured and managed by this separated network, e.g., by the use of Nagios and OpenNMS. Attacks based on the management network are prevented based on a strict configuration of the underlying Community VLAN.

F. INTERCONNECTOR

The Interconnector is a coupling device to provide specific communication channels across the different networks. Because of that, the coupling device is particularly critical for security and must be secured especially. To prevent attacks conducted over the coupling device, only minimal functionality is implemented into the device and a comprehensive firewall is integrated. Because of its strict orientation to IT security, the Interconnector is be realized based on OpenBSD [16]. The installation is limited to absolutely necessary packets. For the communication across the networks, OpenSSH is the only service used based on Public Key Authentication.

5. PROOF OF CONCEPT AND CASE STUDY

For the fulfilment of our future research in the area of Intrusion Detection and network security and as a Proof of Concept, a network environment based on the proposed architecture has been realized in our labs. At the moment, the subsequent elements are used for the setup:

- **TAP-Device:** for multiplying the data streams
- **Manageable Switches:** with configured SPAN-Ports for the analysis of the internal network
- **Cisco Routers:** with Netflow capability for the generation of flow data
- **Console Server:** for the management of switches, routers and IPS

Four racks with numerous different servers and systems are integrated in our network at the moment. The connection coming from the Internet is secured by a firewall. OpenBSD is used for this because of its strict orientation to IT security, with a minimum installation of packets and services.

A TAP-device is installed in front of the firewall to be able to capture all in- and outgoing data. For each connection conveyed through the TAP, multiple copies

of the data stream can be tapped with separated channels for the incoming and outgoing network packets. Next, the data stream of the TAP is sent to the evaluation network and to Cisco Routers with Netflow Capability for the generation of respective statistical data which is also sent to the evaluation network. Also, additional copies of the data stream can be generated via monitoring. The regular data stream after the firewall is routed by manageable switches which are trunked together and configured for the different VLANs in use. The configuration of the firewall respects the structure of the VLANs, e.g., the traffic to the research network typically is not filtered while traffic to systems of the productive network is altered and controlled. To enable an in-depth view onto the internal network, for example for the detection of Insider Activities, each manageable switch is configured to sent all traffic to a SPAN-Port.

All SPAN-Ports are collected by the evaluation network and used for Insider and Extrusion Detection. Because the amount of data transported in the different segments of the internal network is not as high as the traffic volume running through the external interfaces to the Internet, the SPAN-Ports are typically able to copy all data without a loss of packages. Note, that this is not possible when all ports of a switch are heavily used: Because the SPAN-Ports are “regular” ports with the same capacity like all other switch-ports, packets will be dropped randomly in case the volume of the aggregated packets is higher than the data rate of the SPAN-Port. Therefore, this technique can be used to reliably supervise individual network segments, e.g., the systems of one rack, but can’t be used at the borders of the network where TAP devices are needed. It also should be mentioned, that trying to increase the traffic on one switch to force a high amount of dropped packets (e.g., to conceal an attack) is not possible, because this results in a strong deviation to the normal network behaviour, and thus easily detectable by behaviour-based IDS. The data stream of the SPAN-ports is sent to Cisco Routers for the further distribution and the generation of Netflow data.

At the moment, the analysis and evaluation of the Intrusion Detection is done with the following systems (amongst others) in particular:

- **Cisco IPS 4345** (Signature Release S690, January, 16th 2013)
- **Snort Version 2.9.3.1** (snapshot-ruleset downloaded by PulledPork / SourceFire VRT rules, December, 20th 2012 and Emerging Threats rules)
- **Bro Version 2.1**, including snort2bro-translated signatures
- **Suricata 1.3**, Emerging Threats and SourceFire VRT rules
- **Flowmatrix Version 0.30**

Because most systems are running on Virtual Machines in the evaluation network, additional systems can be integrated fast and easily. Furthermore, the data stream can be anonymized and saved into a database in the evaluation network for, e.g., repeating experiments or optimizing IDS-parameters.

One of the advantages of our environment is the possibility of testing and optimizing new systems and configurations without an endangerment respectively disruption of the productive network. For example, when installing the Cisco IPS 4345 device, all rules for blocking traffic had been disabled. Even so, after putting the system into a TAP-link, it started to drop http-traffic. Incidents like this can have serious consequences when they interrupt systems and services in productive networks. In contrast, deployed inside the evaluation network and only working on the copied data stream, unwanted effects cannot influence the productive network and an evaluation of systems and their configurations is possible.

Another aspect is the system performance of IDS in real-world environments. As studies, e.g., by NSS-Labs have shown, often systems are not able to fulfil the specified performance. Until now, our evaluations produce similar results. Several of the considered systems have produced multiple unreported errors during runtime, not recognizable with their User Interfaces. In some situations, systems dropped up to 95% of the network packets or ended in a very high, incomprehensible use of system resources. We will investigate these phenomena in more detail as part of our current research because it is a crucial factor for the utility of IDS in real-world networks.

Even when the systems are running as expected, numerous False Alerts are hampering the use in today's networks. Therefore, our architecture can be used to operate multiple IDS in parallel without any interference or endangerment of the networks. The variety of systems can be used to develop and evaluate correlation strategies aiming for an improvement of detection- and false alarm rates.

At the moment, we are running multiple IDS for the evaluation of security incidents. By that, we have systems specialized and configured for four different doctrines:

- **External Attack Detection** on the Border Gateway
- **Insider Detection** on the Internal Network
- **Misconfiguration Detection** on all Networks
- **Data Leakage Detection** on all Networks

New algorithms and techniques for Alert Correlation are currently under our development. Figure 3 gives an overview of the system used for this purpose.

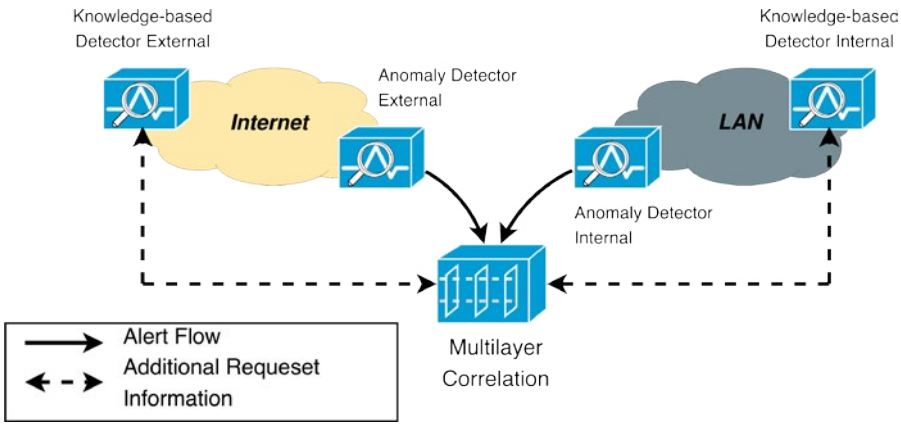

Figure 3. Multilayer-correlation of Sensor Information

The correlation between behaviour-based respectively knowledge-based sensors on different layers (e.g., internal and external network) on the one side and the correlation between behaviour-based and knowledge-based sensors on the same layers is of special importance for improving the detection accuracy and lowering FARs. Therefore, a central component placed in the evaluation network collects alert and sensor information from the different IDS of our network. Based on the doctrine under evaluation, alert and sensor information of different components is correlated and afterwards, further information is requested from other systems. For example, for the External Attack Detection, the behaviour-based alerts of the sensors in front of the border gateway and after the gateway inside the internal network are correlated. By that, chances are increased that a new, unknown attack, which cannot be detected by knowledge-based systems, can be found by the correlation of external and internal deviations and FARs can be reduced. On the other side, further information can be generated by the use of other sensor information, e.g., to narrow down external alerts. See Figures 4 and 5 for an example.

As illustrated, the typical situation, a higher number of anomalies in the external network and a lower number of anomalies in the internal is the case. Often, these anomalies are based on benign, but yet unknown user activities, therefore generating False Alerts. By the correlation of external and internal alerts, events of high relevance can be identified and checked.

For example, the calculation of standard deviations of characteristic traffic parameters can be used to manually identify low intensity anomalies. The IDS Flowmatrix gives a graphical representation, which enables the operator to visually identify anomalies (e.g., special patterns of higher deviations), which are below the regular alert thresholds of the IDS.

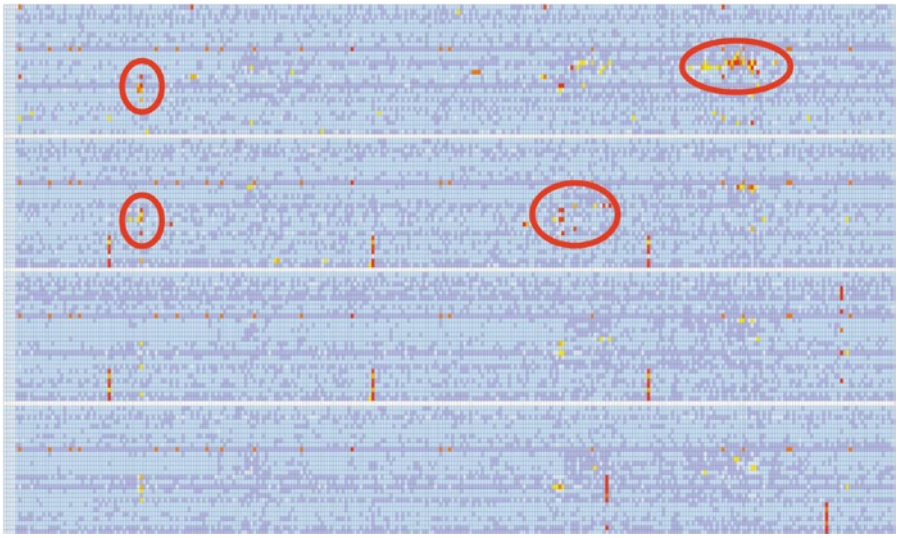

Figure 4. Degree of Standard Deviation for IP Addresses and Ports in the External Network. Clusters of interest are marked

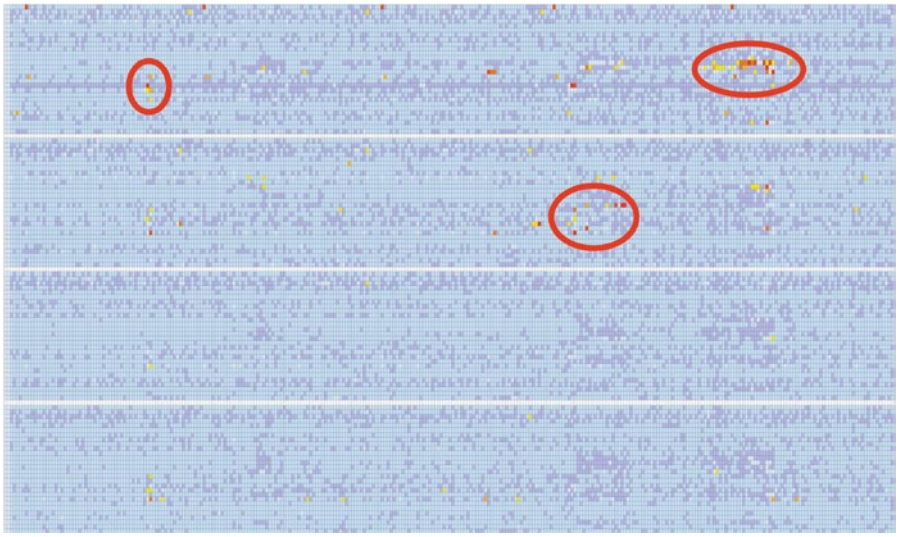

Figure 5. Degree of Standard Deviation for IP Addresses and Ports in the Internal Network. Clusters of interest are marked

Figures 4 and 5 present two screenshots of the analysis of a 300-minute timeslot done by Flowmatrix [17], for the external as well as the internal traffic. The warmer the colour, the higher the deviation of a parameter within a cluster; which is a sign

for an anomaly in the corresponding dataset. The next step of our ongoing work will be the development of correlation techniques, which are able to use such kind of information to reduce FARs on the one side and detect sophisticated attacks on the other.

Based on this knowledge about an event of interest, further data can be collected from other sensors. For example, based on the observation time and IP addresses, collected flow- and header-information can be analysed or knowledge-based systems can be checked for suspicious log-entries. If an identification of an attack is possible, the collected information can be used for a rapid and (semi-) automatic development of new patterns, which is a further goal of our research.

It is important to differentiate if and which alerts are seen in front and behind the border gateway: Typically, external and internal alerts will arise in case of a successful External Attack. Here, more events will be registered on the external network, but often with low intensity, e.g., a service scan. On the other side, after breaking into the system, the attacker will try to investigate the internal network inconspicuous. Because deviations are more significant in the controlled internal network, it is easier to filter our events of interest. These information can be used again for the selection of the relevant events in the external network, which otherwise don't exceed thresholds or decline in the background noise.

The need for intelligent new correlation techniques can be seen in Figure 6.

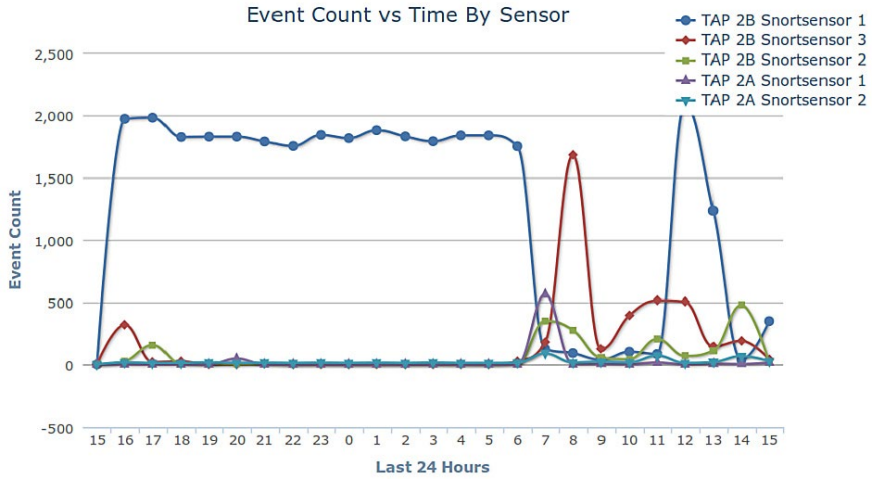

Figure 6. Events registered over 24 hours by different Snort-Sensor in an Academic Network

As depicted, the events produced by the different sensors are quite irregular. Even

throughout the night, a lot of alerts had been generated. Such a strongly irregular behaviour is difficult to learn by a behaviour-based system when the underlying model has to be created. Often, these alerts are harmless and caused by inoffensive actions. Therefore, it is necessary to reduce these alerts by our proposed multilayer-correlation. Based on that, this knowledge also can be used to optimize and control the learning phase of behaviour-based IDS, for example by the pre-filtering of unwanted traffic (which is otherwise learned as normal behaviour and therefore cannot be detected later on in the operational mode).

6. CONCLUSION

Network analysis in company networks is used frequently to discover potential attacks on the computer network. Traditionally, especially IDS are used to defeat these attacks. But often information system departments have difficulties to decide which of the offered IDS they should use. It is often an open question if the system will really fulfil its individual requirements such as different network structure, offered and used services, etc. and therefore companies tend to setup a test implementation before buying the product. One of the biggest problems is that a fair comparison (ceteris paribus) is not possible unless the test environment is equal. Today the well-known DARPA data-set is still used to compare IDS [3]. Because of multiple design errors, the data-set was often criticized and scientists disadvise using it any longer [18, 19]. Even if there are other data-sets available (e.g. MAWI Working Group [20], GEANT [21], ACM SIGCOMM [22]), up to now, none of them was able to prevail. Even more, it has been shown that there is often a strong difference between synthetic- and real-world data based evaluations of IDS (e.g., see [8]). Instead of using fixed and outdated data-sets, our concept shows the possibility to compare systems based on real world data. As described in the proof-of-concept section, our proposed architecture is inserted transparently into the productive network. Thus the architecture gives the opportunity to capture real-world data in real-time as well as the possibility to provide different IDS environments with the same data set for a fair evaluation as well as a sophisticated correlation.

Our architecture is transparent and allows several IDS environments to be implemented in parallel which can be used for configuration optimization, error checking, monitoring the learning phase of anomaly based IDS, etc.

The architecture has also withstood attacks when different security vulnerabilities became known (and exploitable), such as “Multiple Vulnerabilities in Cisco Firewall Services Module” [23] concerning the Cisco firewall module of one of the core switches used. Due to the multi-layered security, attacks in this case were already effectively blocked both by the ACLs of the access switches and the firewall.

The next step will be an integrated and common graphical user interface for the configuration, selection and surveillance of the IDS as well as the fast and easy specification of the rulesets used for the correlation of IDS events. Based on that, the advantages of different system architectures and capabilities can be combined and synergetic effects can be enabled, generating more reliable and secure IDS. Therefore, we are planning to include communication standards and mechanisms like, for example, the data exchange by the Common Intrusion Detection Framework (CIDF) [24], the Intrusion Detection Message Exchange Format (IDMEF) and its associated protocols (Intrusion Detection eXchange Protocol (IDXP), Intrusion Alert Protocol (IAP), Blocks Extensible Exchange Protocol (BEEP); [25, 26, 27]) or the Intruder Detection and Isolation Protocol (IDIP) [28, 29].

Even if a system is not able to provide one of these standard mechanisms, log and alert-files can be evaluated by the use of regular expressions in an efficient way, opening the possibility for integration as well.

Another important aspect of our future work is the conception and development of a correlation strategy for the integrated IDS. As already shown, numerous aspects must be taken into consideration when correlating alerts: For example, as shown by the evaluation of the prototype, some (correct) alarms are only raised by single systems, therefore a majority decision is not enough. Our test environment provides the basis for the development of required, sophisticated correlation techniques.

REFERENCES

[1] Winterfeld, S., & Rosenthal, R.: Understanding Today's Cyber Challenges. Manager, (703), 20151. Retrieved from http://www.tasc.com/news_media/white_papers/TASC_Cyber_Challenges_Study_May_2011_FINAL.pdf, 2011

[2] NSS Labs: Network IPS Group Test 2010. https://www.nsslabs.com/reports/network-ips-group-test-2010, 2010

[3] Lippmann, R., Haines, J.W., Fried, D.J., Korba, J., Das, K.: The 1999 DARPA Offline Intrusion Detection Evaluation. Tech. rep., Lincoln Laboratory MIT, 244 Wood Street, Lexington, MA, 2000

[4] Cuppens, F., Autrel, F., Miege, A., Benferhat, S.: Correlation in an intrusion detection process. Proceedings SEcurite des communications sur internet (SECI02) pp. 153-171, http://www.lsv.ens-cachan.fr/~goubault/SECI-02/Final/actes.pdf\#page=153, 2002

[5] Debar, H., Wespi, A.: Aggregation and Correlation of Intrusion-Detection. In: Recent Advances in Intrusion Detection, Springer 2001

[6] Valeur, F., Vigna, G., Kruegel, C., Kemmerer, R.A.: Comprehensive approach to intrusion detection alert correlation. In: Dependable and Secure Computing, IEEE Transactions on. . pp. 146-169. IEEE 2004

[7] Golling, M., Stelte, B.: Requirements for a Future EWS – Cyber Defence in the Internet of the Future, 2011 3rd International Conference on Cyber Conflict, IEEE, 2011

[8] Koch, R.: Towards Next-Generation Intrusion Detection, 2011 3rd International Conference on Cyber Conflict, IEEE, 2011

[9] Boggs, N., Hiremagalore, S., Stavrou, A., Stolfo, S.: Experimental Results of Cross-Site Exchange of Web Content Anomaly Detector Alerts, IEEE Conference on Technologies for Homeland Security, Boston, 2010

[10] Debar, H., Dacier, M., Wespi, A.: A revised taxonomy for intrusion-detection systems, Annals of Telecommunications 55(7), 361-378, 2000

[11] Network Working Group and others: IETF Policy on Wiretapping. RFC 2804, May 2000

[12] IEEE: 802.1 Q/D10, IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks, Copyright by the Institute of Electrical and Electronics Engineers. Draft, 1997

[13] Hucaby, D., McQuerry, S.: Cisco Field Manual: Catalyst Switch Configuration. Cisco Systems, 2003

[14] Cisco Systems. Inc: Configuring Isolated Private VLANs on Catalyst Switches. Cisco Systems http://www.cisco.com/image/gif/paws/40781/194.pdf, 2008

[15] Claise, B.: RFC 3954: Cisco Systems NetFlow Services Export Version 9. IETF http://www. ietf. org/rfc/rfc3954. txt, 2004

[16] Cowan, C.: Software security for open-source systems. In: Security and Privacy, IEEE Transactions on. . pp. 38-45. IEEE 2003

[17] Xharru Ltd. AKMA Labs: FlowMatrix - Network Behavior Analysis System, http://akmalabs.com/flowmatrix.php, 2012

[18] McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. Inf. Syst. Secur. 3(4), 262-294, dOI http://doi.acm.org/10.1145/382912.382923, November 2000

[19] Athanasiades, N. et al.: Intrusion detection testing and benchmarking methodologies. In: Information Assurance, 2003. IWIAS 2003. Proceedings. First IEEE International Workshop on . IEEE Computer Society, 63-72, March 2003

[20] MAWI Working Group Traffic Archive. Website, http://mawi.wide.ad.jp/mawi/, last seen on February 24th, 2012

[21] MoMe Cluster of European Projects aimed at Monitoring and Measurement. MOME Database. Website, http://www.ist-mome.org/database/, last seen on February 24th, 2012

[22] ACM SIGCOMM. Internet Traffic Archive. Website, http://www.sigcomm.org/ ITA/, last seen on February 24th, 2012

[23] Cisco Security Advisory: Multiple Vulnerabilities in Cisco Firewall Services Module, Cisco Systems, http://tools.cisco.com/security/center/content/CiscoSecurityAdvisory/cisco-sa-20121010-fwsm/, 2012

[24] Kahn, C., Porras, P., Staniford-Chen, S., Tung, B.: A common intrusion detection framework. Submitted to Journal of Computer Security, 1998

[25] Rose, M.: The Blocks Extensible Exchange Protocol core, 1–59. Retrieved from http://www.hjp.at/doc/rfc/rfc3080.html, 2001

[26] Corner, D.: IDMEF-Lingua Franca for Security Incident Management Tutorial and Review of Standards Development. SANS Institute, 2003

[27] Rose, M.: Mapping the BEEP Core onto TCP, 1–9. Retrieved from http://tools.ietf.org/html/3081, 2001

[28] Schnackenberg, D., Djahandari, K., Sterne, D.: Infrastructure for intrusion detection and response. In: DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings. vol. 2, pp. 3-11, IEEE 2000

[29] Kothari, P.: Intrusion Detection Interoperability and Standardization. SANS Institute. Retrieved from http://cs.uccs.edu/~chow/pub/master/sjelinek/doc/research/idmef.pdf, 2002

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Mission-Centricity in Cyber Security: Architecting Cyber Attack Resilient Missions

Gabriel Jakobson
Altusys Corporation
Princeton, NJ, U.S.A.
jakobson@altusystems.com

Abstract: Until recently the information technology (IT)-centricity was the prevailing paradigm in cyber security that was organized around confidentiality, integrity and availability of IT assets. Despite of its widespread usage, the weakness of IT-centric cyber security became increasingly obvious with the deployment of very large IT infrastructures and introduction of highly mobile tactical missions where the IT-centric cyber security was not able to take into account the dynamics of time and space bound behavior of missions and changes in their operational context. In this paper we will show that the move from IT-centricity towards to the notion of cyber attack resilient missions opens new opportunities in achieving the completion of mission goals even if the IT assets and services that are supporting the missions are under cyber attacks. The paper discusses several fundamental architectural principles of achieving cyber attack resilience of missions, including mission-centricity, survivability through adaptation, synergistic mission C2 and mission cyber security management, and the real-time temporal execution of the mission tasks. In order to achieve the overall system resilience and survivability under a cyber attack, both, the missions and the IT infrastructure are considered as two interacting adaptable multi-agent systems. While the paper is mostly concerned with the architectural principles of achieving cyber attack resilient missions, several models and algorithms that support resilience of missions are discussed in fairly detailed manner.

Keywords: *mission-centric cyber security, cyber attacks resilient missions, cyber terrain, impact dependency graphs, adaptable multi-agent systems*

1. INTRODUCTION

Traditionally, the success of cyber security has been measured by the level of cyber attack protection achieved for information technology (IT) infrastructure hardware and software components that are used as an operational resource by different time and space bound activities like military missions and enterprise business processes. Until recently the IT-centricity was the prevailing paradigm in cyber security. It was organized around achieving three main goals: confidentiality, integrity and availability of IT assets [1]. Despite of its widespread usage, the weakness of IT-centric cyber security became obvious with the deployment of large IT infrastructures, where it was economically unjustifiable to seek absolute protection for all IT components, and introduction of mobile tactical missions, where the IT-centric cyber security was not able to take into account dynamic behavior of the missions.

Initial changes in the cyber security paradigm were associated with the introduction of the notions of mission critical assets [2] and network-centricity [3, 4]. The essence of mission criticality in cyber security was 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 promoted by US DoD 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.

The concepts of mission critical assets and net-centricity were important steps in orienting IT security measures towards the real needs of mission security, however in both cases the missions were considered as static entities that at best were used for parameterization of the IT-centric security models. At the same time, protecting missions, not IT infrastructure components is the ultimate goal of cyber security. Of course, the protection of IT infrastructure components continues to play important, but still, the subordinate role in mission cyber security. In other words, the success of protecting IT infrastructure components should be measured by the success of missions that this IT infrastructure is supporting. We will call this mission-centric cyber security

In this paper we are introducing the notion of cyber attack resilient missions as an example of mission-centric cyber security systems. We will show that mission cyber attack resilience is achieved through emergent (collective and adaptive) behavior of IT infrastructure components and missions. The paper discusses several critical architectural principles of achieving cyber attack resilience of missions, including mission-centricity, resilience through adaptation, and synergistic mission C2 and mission cyber security management. In order to achieve the overall system

resilience under a cyber attack, both, the command an control of missions in the phusical space, and management of IT infrastructure components in s cyber cpace are considered as two interacting adaptable multi-agent systems. As such, the quality of those physical and cyber operations cannot be any more assessed as silos of two independent processes.

The rest of the paper is organized as follows. Section 2 discusses how the notion of resiliency is understood in different disciplines, provides a definition of resilient missions, and reviews relevant work. Section 3 describes the basic conceptual elements and architecture of a mission-centric cyber security. Section 4 describes the models of cyber terrain, impact dependency graph, and the tactical space and time bound missions that are used in the proposed approach. Section 5 provides a model of mission resilience that is reached via interactive adaptation of cyber terrain and missions, it describes how the process of mission adaptations can be implemented using an adaptable multi-agent system, and presents a sample set of mission adaptation policies. Section 6 draws some conclusions and refers to the future research directions.

2. CYBER ATTACK RESILIENT MISSIONS

In this section we'll review some origins of the notion of resiliency in complex systems and define the notion of a cyber attack resilient mission.

A. *UNDERSTANDING RESILIENCY*

Resilience as a fundamental feature of all complex systems, being them natural or artificial systems, has been an interest or study of many scientific disciplines. Dictionary.com defines resilience as the power or ability to return to the original form, position, etc., after being bent, compressed, or stretched; or as ability to recover readily from illness, depression, adversity, or the like. In social science resiliency is the ability of individuals, but also groups, to overcome challenges, like trauma, tragedy, crises, isolation, and bounce back stronger, wiser, and more socially powerful [5]. Psychological resilience is an individual's tendency to cope with stress and adversity. This coping may result in the individual "bouncing back" to a previous state of normal functioning, or simply not showing negative effects [6]. In engineering disciplines resilient systems are designed to anticipate and avoid catastrophic accidents, and survive and recover from natural disruptions and terrorist attacks [7]. A general framework for classifying system resilience is given in [8]. In [9] the resilience of a system or organization is understood as including at least two of the following capabilities: (a) anticipation and preparation before an adverse event; (b) survival during the event; and (c) recovery after the event.

Summarizing the different understandings of system resiliency, one can define the principal goal of resilient systems is the system's desire to survive, even if not any individual component of the system is surviving. In other words, the system resiliency is achieved through emergent (collective and adaptive) behavior of all components of the system. The emergent systems [10, 11] expose new global properties not as a mechanical sum of local properties of its components but as a qualitatively new feature that emerges from the inter-component interactions and adaptations.

B. DEFINING A CYBER ATTACK RESILIENT MISSION

Inspired by the definition of resilient computer networks given in [9] we define resilient missions as missions that in a given time window are able to reach their operational goals situation under the impact adverse events, like adversary attacks, human errors, disruptions in support services, and natural disasters. The concept of mission resilience assumes structural changes in mission task flows, adaptability of mission execution processes, and a graceful degradation of mission goals.

As applied to the domain of mission cyber security we define cyber attack resilient mission as resilient missions that are capable to:

a) Predict plausible impact of cyber attack situations **before** they occur;

b) Survive through adaptation and graceful degradation **during** the attacks;

c) Recover its operational capacities **after** the attacks;

As we already mentioned in the Introduction, we will consider a mission and its supporting IT infrastructure together as one synergistic interacting system. This is an important conceptual viewpoint – by adopting it we will show that cyber attack mission resiliency can be achieved by cross-mission and IT infrastructure interactions and adaptive behavior of all components of this synergistic system containing both the IT infrastructure components and mission components.

C. RELATED WORK

Over the last three decades significant research and development results have been reached in the area of cyber attack tolerant, survivable, and resilient IT systems [12-15]. A broad overview of resilient computer networking and related fields is given in [16], where the resilience is defined as the ability of the network to provide and maintain an acceptable level of service in the face of attacks, faults, natural disasters and other challenges to normal operation. Probably, Fraga and Powell were the first who used the terms of "fault tolerance" and "intrusion tolerance"

in 1985, when they described the capabilities of a fault and intrusion tolerant file system [17]. Since then the term fault tolerance is understood as a capability of the system to continue satisfactory operations in the presence of faults. Fault tolerance capabilities are built in almost every modern technological and infrastructure system, including communication networks, power grids, space systems and others. During the last three decades significant results in fault tolerance research were achieved by fault tolerant computing [18], including distributed fault-tolerant architectures, masking (hardware redundancies), models of graceful degradation, dynamic reconfiguration, fault detection by spatial and temporal event correlations, automatic recovery and response techniques, system vulnerability analysis, damage assessment and evaluation, and other methods. Since the start of research on intrusion tolerant systems almost two decades ago significant body of research and system development has been produced. A good overview of those results has been presented in [19].

A model of increasing mission survivability based on reinforcement learning was proposed in [20]. The paper defines the measure of mission survivability as a ratio between the successfully completed workflows of the mission to the total number of the workflows. The paper examines two core capabilities to increase mission survivability: redistribution of the network resources to ensure mission continuity, and learning of the attack patterns to estimate the level of vulnerability of other nodes. Both of these capabilities are concerning the resource network, while adaptation of the mission was not addressed.

In June 2011 The Defense Advanced Research Projects Agency (DARPA), the US Department of Defense's advanced research department, announced that it is working on a project called Mission oriented Resilient Clouds (MRC), which aims to build resiliency into existing cloud networks to preserve mission effectiveness during a cyber attack [21]. The MRC program will run an ensemble of interconnected hosts acting in concert. Loss of individual hosts and tasks within the ensemble is allowable as long as mission effectiveness is preserved. The MRC project will include redundant hosts and will be able to correlate attack information while switching around resources. The goal is to provide resilient support to the mission through adaptation. The MRC program looks on cyber attack resilient clouds that are adaptable towards mission needs, still adaptation of the missions themselves as in [21] is not defined in the program research topics.

3. ARCHITECTURE OF A CYBER ATTACK RESILIENT MISSION

Reference architecture of a cyber attack resilient mission is given on Figure 1. It contains two main interacting closed-loop processes: Cyber Security Situation Management (CSSM) process and the Mission Operations Situation Management (MOSM) process. The CSSM and MOSM processes interact through a common object of interest – the mission. As mission progresses in time CSSM receives IT service requests from the mission and provides the requested services back to the mission. Concurrently to this process, MOSM proceeds with the tasks of mission situation awareness, undertakes mission decision support functions, and transitions the mission into a new state. The new mission state might require renewed IT support services from CSSM. In order to achieve resiliency to withstand the impact of cyber attacks the above-described interaction between CSSM and MOSM requires of mutual adaptation of the cyber terrain and the mission, e.g. reconfiguration of dependencies among the cyber assets and services, replacing or upgrading certain assets, changing the logical or temporal order of mission tasks, or proceeding with a graceful degradation of the mission goals.

Figure 1 illustrates a tactical military mission conducted in an urban mission operational theater. The mission is conducted by two small military units against hostile agents. In addition to the cyber attacks, the mission must withstand physical impacts caused by natural forces and external mission disruptions. MOSM acts according to the mission model, and military tactical policies and rules. The MOSM includes two sub-processes, the Mission Situation Awareness (MSA) and the Mission Decision Support (MDS) processes. MSA and MDS themselves are fairly complex operations: MSA performs the tasks of (a) sensing and pre-processing of real-time data coming from sensors and human reports; (b) perception of the collected data and construction of the tactical situation model of the operational; (c) mission impact assessment caused by the actions and forces in the Physical Space; and (d) prediction of future plausible impacts on the mission caused by adverse events in the physical space. MDS performs the tasks of mission operations planning, mission adaptation and mission execution.

Like the MOSM process, the closed-loop CSSM process contains two major sub-processes, Cyber Security Situation Awareness (CSSA) and Cyber Security Decision Support (CSDS) processes. The CSSA process includes the following tasks: (a) real-time correlation of cyber attack alerts, and recognition of complex multi-stage cyber attacks; (b) cyber attack impact assessment on cyber assets that were directly hit by the attack, (c) propagation of the impact of the cyber attack through the inter-component dependencies in the Cyber Terrain, and

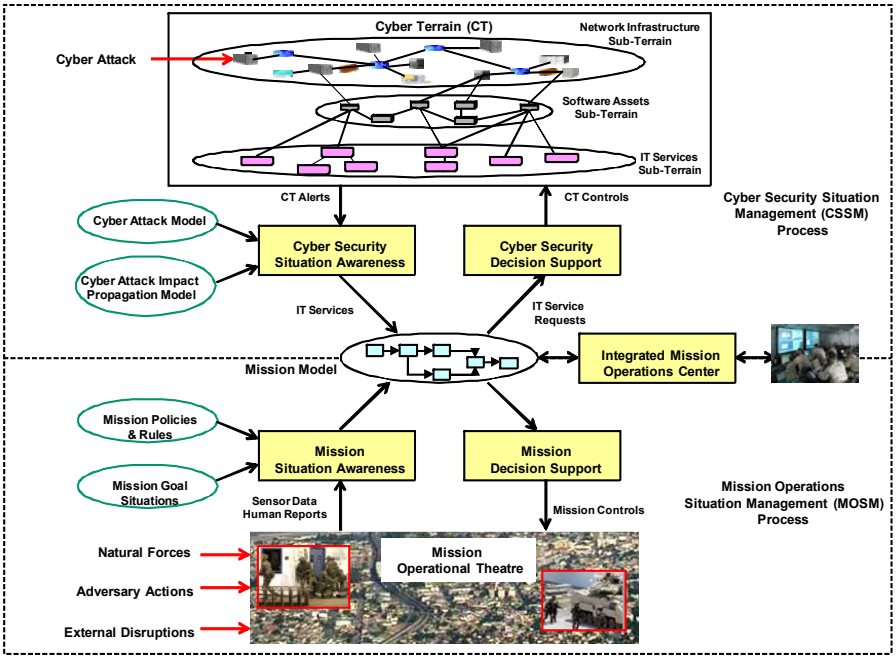

Figure 1. Synergistic mission cyber security and command & control management

(d) assessment of plausible future cyber attack impacts. The cyber attack impact propagates through the CT and reaches mission tasks that consume the cyber services provided by the CT. Through the fabric of mission task, sub-mission and mission dependencies this impact reaches the top level of a mission and might affect the success of the mission completion. The CSDS process contains the following tasks: (a) CT vulnerability scanning and preventive maintenance, (b) CT adaptation as response to the cyber attacks and as reaction to IT service requests from the missions, and (c) CT recovery actions.

For performing of the above-mentioned tasks the CSSA and CSDS processes need variety of data and knowledge sources. In this paper we will mention two of them, the Cyber Attack Model and the Cyber Attack Impact Propagation Model. The Cyber Attack Model is used for calculating the effect of the cyber attack on the operational capacity of the directly hit cyber assets, while the Cyber Attack Impact Propagation Model is used for calculating the indirect impact of the cyber attack on those assets that are tied by dependencies according to the structure of the CT.

The closed-loop CSSM and MOSM processes are conceptually built following the principles of Situation Management (SM), which is more in detail discussed in our earlier work [23].

4. MAIN CONCEPTUAL COMPONENTS OF THE APPROACH

In this section we will discuss several key elements of the proposed approach of building cyber attack resilient missions, including cyber terrain, tactical mission and impact dependency graph.

A. CYBER TERRAIN - MODELING IT INFRASTRUCTURE

Cyber terrain (CT) is a multi-level information structure that describes cyber assets and services, and their intra- and inter-dependencies [22]. As was already shown on Figure 1 it contains three sub-terrains: hardware, software, and service sub-terrains. The hardware (HW) sub-terrain is a collection of connected network infrastructure components like routers, servers, switches, firewalls, communication lines, terminal devices, sensors, cameras, printers, etc. All the dependencies between the components, like connectivity, containment, location, and other relations, represent the physical/logical topology of the HW sub-terrain. The software (SW) sub-terrain describes different software components, such as operating systems, middleware, applications, etc., and defines its own dependencies between the components. A software component in the SW sub-terrain might be characterized by different attributes like functional class of the component, vendor specification, release number, references to known vulnerabilities, etc. The service sub-terrain presents all the services and their intra-dependencies. Examples of typical services include database, file transfer, e-mail, GIS, universal time, and security services. The most common dependencies between two services include: enabling of one service by other and containment of one service within a package of multiple services.

As among the components of a sub-terrain, dependencies exist between the sub-terrains: a HW sub-terrain component may “house” SW sub-terrain components and a SW sub-terrain component may enable some services. CT is a dynamic information structure: its components and their inter-dependencies are a function of time.

While supporting the missions, the CT possesses certain “operational capacity”, i.e. the ability to provide resources and services to the missions with a certain level of quantity, quality, effectiveness, and cost to the missions. In this work we will introduce the operational capacity (OC) as a universal measure characterizing the operational quality of each of the component in the CT, being it a cyber asset or service. The operational capacity is measured in an interval [0, 1], which indicates to what level the asset or service was compromised under a cyber attack. Value 0 means that an component is totally compromised (not trustworthy, not operational) and value 1 means that the component is fully operational.

In a general attack situation, a software asset can be either directly hit by a cyber attack causing permanent damage to its operational capacity, or the operational capacity of an asset may be indirectly impacted by a remote attack via inter-asset dependencies. The operational capacity level of the directly hit asset stays unchanged as long as corrective actions are made to the asset. A sequence of direct attacks might reduce the operational capacity of an asset, or totally destroy the asset bringing its operational capacity to 0. Contrary to the effect of the direct attack, an indirect cyber attack does not cause permanent damage to the cyber asset. However, its operational capacity might be reduced because of its dependency on other assets that either suffer from direct attacks or are also indirectly impacted. To measure the impact of a permanent damage to the software assets, we will introduce the notion of permanent operational capacity (POC) that is applicable only to software assets.

B. MISSIONS

Military mission (aka military operation) is 0 coordinated order of space and time bound military actions to resolve political or military situations in the favor of the agent conducting the mission. Depending on the scope of developing situations, the size of the engaged military units, and the defined goal situations the military missions are considered at three main levels: strategic, operational and tactical levels. The strategic mission describes actions over large, often continental area of operations with national commitment to the mission. The operational level mission describes a subset of a strategic operation with specific military goals, while the tactical mission being part of an operational level mission is limited in time, space, the scope of objectives, and engaged military resources. In this paper we are focusing mostly on tactical missions.

Missions are modeled sequential or parallel flows of mission steps that in addition to the AND/OR logic, are controlled by temporal interval logic [25]. The content of the actions executed at a mission step is defined by a mission task. It is not excluded that the same tasks can be executed at several different mission steps, and a single task can be decomposed into a sequence of multiple steps, if of course, from the mission command control perspective such need arises. A mission step can be another flow, another mission, or mission task. Figure 3 illustrates a Mission X that has two parallel flows that are forked by an AND-node. The first branch contains another flow of three sequential steps (d1, d2, d3), while the second flow contains two sub-missions A and B. The Mission A represents itself two flows that are forked by an OR-node, while the second mission B represents a special case of an AND-node called “Cloud”. The AND-node requires that both branches of the flow should be executed, while the OR-node prescribes that at least one branch should be taken.

All missions and mission steps are formally described as interval events that have their start time, duration and end time. While the AND-nodes and OR-nodes specify only the logical conditions of executions of the mission flow branches, they do not identify the exact temporal order of missions/mission steps as events. For example, on Figure 2 the AND-node in Mission X specifies that both branches,

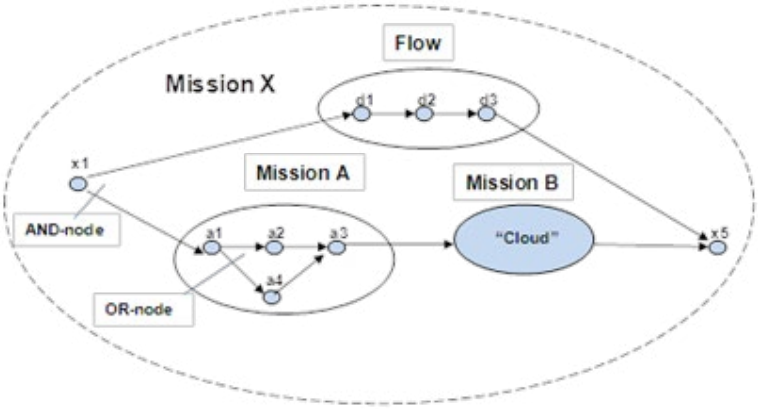

Figure 2. Mission Task Flows

Flow(d1, d2, d3) and Flow(Mission A, Mission B) should be taken, but the question in what temporal order remains open. In order to determine the order of execution of mission flows we will use temporal logical relations such as BEFORE, AFTER, STRICTLY-AFTER, etc. between the mission steps. In our earlier paper on temporal relations in event correlation [24] we used temporal interval logic proposed by John Allen [25]. In addition to those temporal relations we will introduce in this paper a temporal relation UNDEFINED that do not require any specific temporal relation to be identified between the events. The above-mentioned sub-mission B called "Cloud" is exactly described by the temporal relation UNDEFINED, namely we require that all steps from the "Cloud" should be taken, but in any arbitrary order.

The existence of temporal order between missions and mission steps, and the options to change the order, e.g. advance or delay the order of execution of mission flows, opens an opportunity to adapt mission so that to minimize the cyber attack impact on missions. Such method of mission adaptation will be discussed in the Section IV. As the embedded structure of missions unfolds during the mission execution process all mission steps will be ultimately turned into executable mission tasks.

C. IMPACT DEPENDENCY GRAPH

Formally, the cyber terrain and the missions and propagation of the impact through cyber terrain and the mission-submission structure is described by the impact dependency graph. Impact dependency graph (IDG) [22] is a mathematical abstraction of the domain semantics of assets, services, mission steps and missions and all of their dependencies. We consider assets, services, mission steps and missions as nodes of an IDG and their inter-dependencies as dependencies among the nodes of the IDG. In addition to the nodes of assets, services, mission steps and missions, IDG has two special nodes: AND-nodes and OR-nodes that represent logical dependencies among nodes in IDG. The AND-node defines that the parent node depends on all of its children nodes, while the OR dependency defines the required presence of at least one child node. The OR dependency is introduced to capture system redundancy or for alternative functionality, performance, cost, reliability or for some other reason. Figure 3 shows a sample impact dependency graph, which comparing with an IDG introduced in [22] has been extended with an Agent Pool.

As a result of a cyber attack against the cyber terrain, the cyber attack impact propagates through the IDG, and when the impact reaches an agent pool, the operational capacities (OC) will be calculated for all agents in the pool. The agent with the highest OC in the agent pool will be assigned to the corresponding mission task, and then the impact propagation process continues up to the top level mission node in the IDG.

During real-time mission monitoring, the impact of a cyber attack on a mission depends on two major factors: (1) what impact the attack has on steps of the mission, and (2) in what state - planned, ongoing, or completed state the mission steps are. For example, if the cyber attack can impact assets and services that support steps a, …, m, but those steps have been already completed (see Figure 4), then the impact of the attack should be irrelevant as far as these steps are concerned. Contrary, the ongoing steps during the cyber attack, like step x will be directly affected by the attack. The case for the steps that are planned for execution (steps p to s) at the moment when a cyber attack happens needs a special analysis.

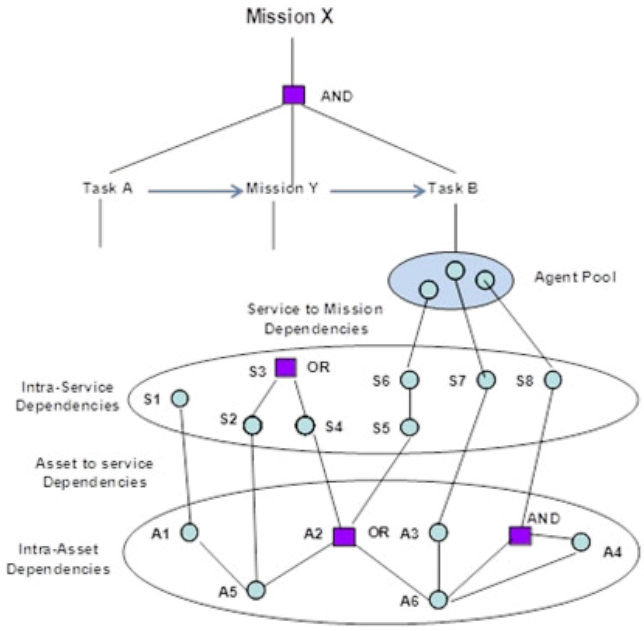

Figure 3. Impact Propagation Graph

First, since those steps have not yet been undertaken, their operational situation will not be accounted in the calculation of the operational situation of the overall mission. However, we are able to calculate a potential impact on those steps, which could happen. One practical action could be to reconfigure the cyber terrain or give a warning to the mission C2 commander.

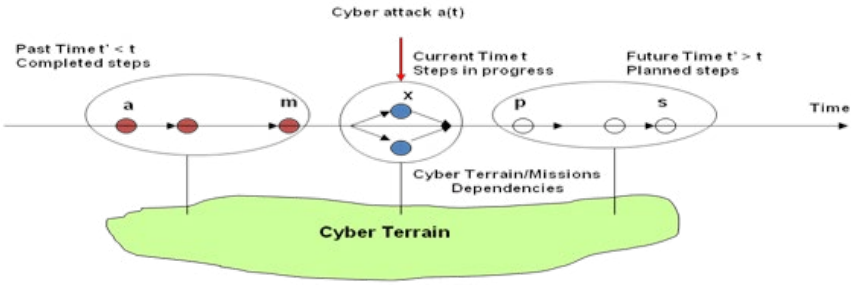

Figure 4. Time-dependent impact of cyber attacks on missions

5. MISSION RESILIENCE THROUGH ADAPTATION

A. *ADAPTATION IN MULTI-AGENT SYSTEMS*

In order to achieve mission resilience under a cyber attack, both, the missions and the CT are considered as two interacting adaptable multi-agent systems (MAS). In this section we outline some principles of CT and mission adaptation.

In many applications, including mission command and control, and mission cyber security management, operational components of the systems need to be flexible and adaptable to deal with dynamic environments. To address this need, there are several general requirements of the system architecture, including openness, self-awareness and the use of meta-knowledge to adjust the structural organization and behavior of the system according to the adopted policies. It is assumed that an adaptable system is capable of exhibiting autonomous run-time behavior without outside intervention. Often the following types of adaptation are considered:

- Structural adaptation – adaptation to internal structural changes, e.g. loss of inter-node connectivity, or loss of nodes
- Functional adaptation - detection of changes in the functions of nodes of the system,
- Resource adaptation - adaptation in the system internal resources, i.e. loss or corruption of physical memory, or loss of battery power

All these three types of adaptations are useful in adaptation of CT and missions to achieve mission survivability and they will be used through the framework of adaptable multi-agent systems. The paradigm of multi-agent systems has its roots in distributed artificial intelligence, object oriented systems and human team cognition. MAS is currently one of the most powerful approaches used in building distributed computing systems [26]. MAS have several important features which correspond to our specific interests, particularly:

- Adaptation: the ability to reorganize and improve behavior with experience
- Autonomy: goal-directedness, proactive and self-starting behavior
- Collaboration: the ability to work with other agents to achieve a common goal
- Inference: the ability to act on abstract task specifications
- Mobility: migration in physical or cyber space

A typical MAS solution to situation awareness, and consequently to the whole

process of command and control, is based on dividing situation awareness, command and control into several dedicated agents either across functional tasks, e.g. data detection, classification, visualization, etc., or across levels of abstraction of information, e.g. signal, data and semantic information levels. In this paper we will use BDI (Belief, Desire and Intension) agent model that was originally proposed in [27-29] and later advanced with adaptation capabilities [30, 31] as a main building block for MAS.

B. *MISSION ADAPTATION POLICIES*

Mission adaptation policies are rules that are used by an agent to modify the missions, its components and inter-dependencies between the mission components. From a mission execution viewpoint each task is implemented by an agent that is assigned to the task. As we talk about missions as objects of adaptations, two types of mission adaptation methods are considered, entity-level adaptation, relation-level adaptation. On entity-level adaptation each entity, a mission, task or an agent can be a subject for modification. For example, one can change the criticality index of a mission or a task, or operational capacity of a task or an agent. An important adaptation function is the election of an agent from a pool of pre-defined agents to implement a particular mission task. All these individual adaptation functions are undertaken within the constraints identified for each entity. The relation-level adaptation covers the functions of changing or modifying the structural, temporal, logical, or domain-specific relations between the entities. For example, adding or deleting a task, changing the AND-nodes and OR-nodes in a mission flow, changing the temporal order of tasks in a mission flow, delaying or moving up the start or the end time of a mission or its components. Below we will present a sample list of mission adaptation policies for ongoing mission tasks that are under execution at the time of the cyber attack:

1. For every currently active mission task select an agent from a corresponding agent pool that has the highest operational capacity that is equal or greater than the required operational capacity specified in the mission task. If no agent is found, use Policy #2.
2. Reduce incrementally the value of the task's required operational capacity from the current value to the lowest permitted level. For each incremental required operational capacity value perform the Policy #1. If no agent is found that matches the Policy #1, use Policy # 3.
3. Modify the mission task flow so that the tasks with no matching agents are moved for a later time of execution. Issue a CT reconfiguration order to replace/or repair the CT node with a low operational capacity.

4. Stop execution of those mission tasks, where (a) the stop execution permission is granted, and (b) no agent could be found with operational capacity that is at least equal to the required operational capacity of the task.
5. Select from the alternative mission flows (mission flows that are in OR condition among themselves) a flow where all tasks have the matching agents, whose operational capacities are greater than the required operational capacities in the corresponding tasks.
6. Select first those tasks from the "Cloud" in the mission flow that satisfy the required operational capacity condition. For the rest of the tasks issue CT reconfiguration order.

Our approach to mission cyber attack impact assessment, both to the current real-time impact when the cyber attack occurred during the execution of the mission, and assessment of the impact of plausible future cyber attacks is discussed elsewhere [22, 32].

6. CONCLUSIONS

In this paper we stressed the importance of a cyber security paradigm shift by moving towards mission-centricity in cyber security. We motivated this paradigm shift with several arguments, namely the fast increase in the scale of IT infrastructures and the practical inability to protect every component of the IT infrastructure, as well the high mobility and dynamics of modern battlefield and business processes. In this paper we proposed the notion of cyber attack resilient missions and how they should act before, during and after the cyber attacks. As proposing the architecture for building those missions, we presented several innovative solutions, including (a) synergistic adaptation of the cyber terrain and tactical missions implemented as two situation-aware adaptable BDI multi-agent systems, (b) the overall model of a cyber situation management system, (c) the model of cyber attack impact propagation through the impact dependency graph (IDG), and (d) modeling the dynamic behavior of missions by graphical flowcharts augmented with logical and temporal constraints.

We argued that only integrated approach that combines synergistic management of mission command and control, and mission cyber security can lead to resilient and survivable missions. Future work will include extending of the proposed principles to resilient and survivable missions that are oriented towards faults, human errors, and natural and technological disasters.

REFERENCES

[1] Aceituno, V., 2005, On Information Security Paradigms, *ISSA Journal,* September, 2005.

[2] US GAO, 2011, Critical Infrastructure Protection. Cybersecurity Guidance Is Available, but More Can Be Done to Promote Its Use", *USA GAO Report to Conressional Requesters GAO-12-92*.

[3] US DoD, 2012, "Department of Defense Net-Centric Data Strategy", http://dodcio.defense.gov/docs/net-centric-data-strategy-2003-05-092.pdf.

[4] Kerner, J., Shokri, E., 2012, Cybersecurity Challenges in a Net-Centric *World, Aerospace Crosslink Magazine*, Spring 2012.

[5] Cacioppo, J. T., Reis, H. T., Zautra, A. J., 2011, Social Resilience: The Value of Social Fitness with an Application to Military, *American Psychologist*, Vol. 66, No. 1, pp. 43-51.

[6] Reivich, K., Shatte, A., 2003, The Resilience Factor: 7 Keys to Discovering Your Inner Strength and Overcoming Life's Hurdles, *Random House*,.

[7] Jackson, S., 2007, A Multidisciplinary Framework for Resilience to Disasters and Disruptions, *Journal of Design and Process Science*, June 2007.

[8] Mostashari, A., 2010, Resilient Critical Infrastructure Systems and Enterprises, *Imperial College Press.*

[9] Westrum, R., 2006, A Typology of Resilience Situations, in (Eds. E. Hollnagel, D. Woods, D. Lelvenson) *Resilience Engineering Concepts and Precepts. Aldershot, UK: Ashgate*.

[10] De Wolf, T., Holvoet, T, 2004, Emergence and Self-Organization: a statement of similarities and differences, In: *Proceedings of the Second International Workshop on Engineering Self-Organizing Applications,* New York, USA , pp.96–110.

[11] Edmonds, B, 2004, Using the Experimental Method to Produce Reliable Self-Organized Systems, In Brueckner, S., Serugendo, G.D.M., Karageorgos, A., Nagpal, R., eds. *Engineering Self Organizing Systems: Methodologies and Application*s. Volume 3464 of Lecture Notes in Artificial Intelligence. Springer, 2004.

[12] Siewiorek, D., ed., 1995, *Fault-Tolerant Computing Highlights from 25 Years*, Special Volume of the 25th International Symposium on Fault-Tolerant Computing FTCS-25, Pasadena ,CA.

[13] Ellison, R. J., Fisher, D. A., Linger,R. C., Lipson, H. F., Longstaff, T. A., and Mead, N. R. 1999, An Approach to Survivable Systems, *Technical Report, CERT Coordination Center, Software Engineering Institute*, Carnegie Mellon Institute.

[14] Lipson, H.F., Fisher, D.A, 2000, Survivability—a New Technical and Business Perspective on Security, In: NSPW 1999: *Proceedings of the 1999 Workshop on New Security Paradigms*, pp. 33–39. ACM, New York.

[15] P. Smith, P., Hutchinson, D., Sterbenz, J. P. G, Scholler, M., Fessi, A., Karaliopoulos, M., Lac, C., Plattner, B., 2011, Network Resilience: A Systematic Approach, *IEEE Communications Magazine,* July, 2011, pp. 88-97.

[16] Sterbenz J. P. G., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer J. P., Schöller, M., Smith, P., 2010, Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines, Elsevier Computer Networks 54, pp. 1245-1265.

[17] .Fraga, J.S., Powell, D., 1985, A fault- and intrusion-tolerant file system, In *Proceedings of the 3rd International Conference on Computer Security*. 203–218.

[18] Rennels, D. A., 1999, Fault-Tolerant Computing, *Encyclopedia of Computer Science*, ed., Anthony Ralston, Edwin Reilly, and David Hemmendinger.

[19] Verıssimo, P., Neves, N., Correia, M., 2003, Intrusion-Tolerant Architectures: Concepts and Design. In Architecting Dependable Systems, R. Lemos, C. Gacek, A. Romanovsky (eds.), *LNCS 2677, Springer Verlag.*

[20] Carvalho, M. 2009, A Distributed Reinforcement Learning Approach to Mission Survivability in Tactical MANETs, *ACM Conference CSIIRW 2009*, Oak Ridge TN.

[21] Mission-Oriented Resilient Clouds, 2011, DARPA, Information Innovation Office, http://www.darpa.mil/Our_Work/I2O/Programs/Mission-oriented_Resilient_Clouds_(MRC).aspx.

[22] Jakobson, G., 2011, Mission Cyber Security Situation Assessment Using Impact Dependency Graphs, *Proceedings of the 14th International Conference on Information Fusion*, Chicago, IL.

[23] Jakobson, G., Buford, J., Lewis. L. 2007, Situation Management: Basic Concepts and Approaches, *Proceedings of the 3rd International Workshop on Information Fusion and Geographic Information Systems*, St. Petersburg, *Lecture Notes in Geoinformation and Cartography, Springer-Verlag Berlin Heidelberg.*

[24] Jakobson G., and Weissman, M., 1995, Real-Time Telecommunication Network Management: Extending Event Correlation with Temporal Constraints, the *4th IFIP/IEEE International Symposium on Integrated Network Management*, Santa Barbara, CA, pp. 290-301.

[25] Allen, J. F, 1983, Maintaining Knowledge About Temporal Intervals, *Communications of the ACM* 26 (11), pp. 832-843.

[26] Wooldridge, M., 2002, An Introduction to Multi-Agent Systems, John Wiley and Sons.

[27] Bradshaw. J. M. 1997, An Introduction to Software Agent, In: *Software Agents*, Menlo Park, Calif., AAAI Press.

[28] Norling, E., 2004. "Folk Psychology for Human Modeling: Extending the BDI Paradigm," *In International Conference on Autonomous Agents and Multi-Agent Systems.*

[29] Rao, A., and Georgeff, M., 1995, BDI Agents: From Theory to Practice, In *Proceedings of the First International Conference on Multi-Agent Systems.*

[30] Jakobson, G., Buford, J., and Lewis, L., 2008, Models of feedback and adaptation in multi-agent systems for disaster situation management, *SPIE 2008 Defense and Security Conference, Orlando, FL*.

[31] Maes, P., 1993, Modeling Adaptive Autonomous Agents, *Artificial Life*, vol.1, No 1-2, pp. 119-128.

[32] G. Jakobson. G., 2011, Extending Situation Modeling with Inference of Plausible Future Cyber Situations, *CogSIMA 2011*, Miami, FL.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Autonomous Intelligent Agents in Cyber Offence

Alessandro Guarino
StudioAG
Vicenza, Italy
a.guarino@studioag.eu

Abstract: Applications of artificial intelligence in cyber warfare have been mainly postulated and studied in the context of defensive operations. This paper provides an introductory overview of the use of autonomous agents in offensive cyber warfare, drawing on available open literature. The study supplies an introduction to the taxonomy and science underlying intelligent agents and their strengths and weaknesses: the technological elements that autonomous agents should include are explained, as well as the economics involved. The paper also aims to explore possible legal implications of the use of autonomous agents and shows how they could fit into the legal context of warfare. The conclusion of the study is that the use of AI technologies will be an important part of cyber offensive operations, from both the technological and the economical aspects; however, the legal and doctrinal landscape is still uncertain and proper frameworks are still to be developed.

Keywords: *artificial intelligence, autonomous agents, cyber warfare, cyber attack, international law*

1. INTRODUCTION

Cyber warfare is more and more a moving target: developments in the field are rapid, especially in the technological arena, and artificial intelligence (AI) techniques are more and more at the heart of applications. The concept of agents has been known for some time and software with some agent characteristics is already present and deployed, but in the near future we will probably see the birth of true autonomous agents, which will be a new breed entirely. This paper will propose a detailed definition of their capabilities and of what it will take to be considered truly an autonomous intelligent agent, as well as describing how their advent could fit into the international law of war. We are conscious that this subject can be considered highly speculative at the moment, and indeed it is; but until now, the discussion on international regulation of cyber warfare has been virtually nonexistent outside specialist circles and a debate on possible updates of international law to accommodate these new developments is sorely needed, especially considering that the very probable deployment of AI techniques could trigger an escalation in their use. This paper focuses on offensive activities in cyber warfare, i.e. 'cyber offence', including both Computer Network Attack (CNA) and Computer Network Exploitation (CNE), as defined in the U.S. Joint Publication 3-13. So, cyber offence includes 'actions taken via computer networks to disrupt, deny, degrade, or destroy the information within computers and computer networks and/or the computers/networks themselves' and 'actions and intelligence collection via computer networks that exploit data gathered from target or enemy information systems or networks'.

2. AUTONOMOUS AGENTS

A. CHARACTERISTICS

In the field of AI there exists a traditional division between two concepts: 'strong' and 'weak' AI: strong AI aims at creating nothing short of what the name implies, namely an intelligent being of a different species than humans but at least as capable, while weak AI assigns itself the more limited target of replicating single feats of intelligent behaviour, not surpassing human intelligence, e.g. computer vision systems, game-playing software, etc.

We leave aside the philosophical debate that the 'weak vs. strong AI' discussion entails, which is deeply interesting but alien to the aims of this paper. We concern ourselves here with 'autonomous (intelligent) agents': whatever their exact definition, surely they do not (yet) belong to the realm of strong AI and are rather applications

of various technologies mimicking natural intelligence. Autonomous intelligent agents can be purely software, or integrated into a physical system ('robots') – the difference lies mainly in the environment in which the agent operates: while purely software agents live in what we call 'cyberspace', robots can sense and interact with the same physical environment that we live in. The environment makes a lot of difference for autonomous agents, as we shall see, but the similarities between software agents and robots are relevant, given that even in a robot the embedded software – or firmware – is at the heart of its behaviour and capabilities.

Autonomous agents are a special kind of computer program, but what makes them special? Basically, every computer program is autonomous in a way – this is what computers are for, after all – so we need to develop a useful definition, especially for the field we are concerned with, cyber warfare. Franklin and Graesser [1] have given a convincing definition of agents and the ways in which they differ from other software:

> *An autonomous agent is a system 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.*

This formulation contains some very important points:

1. An agent is strictly associated with its environment: an autonomous agent outside the environment it was designed for can be useless, or not an agent at all.
2. An agent interacts with the environment, via appropriate sensors providing input from it and appropriate actuators allowing the agent to act and influence that environment.
3. An autonomous agent acts towards a goal, it has an 'agenda' in the words of Franklin and Graesser. In particular, an autonomous agent developed for warfare operations is assigned a target.
4. The activities of a truly autonomous agent are sustained 'over time', so it must have a continuity of action.

Some characteristics are probably missing here, which are required to round up our definition, even if it can be argued they are implicit in the above formulation. First, an autonomous agent needs to possess an internal representation of its environment, or what is called a 'belief state'. In their standard textbook *Artificial Intelligence – A modern approach,*[2] Stuart Russell and Peter Norvig introduce a classification of agents that ranges from simple reflex agents, where there is no internal model of the environment, through model-based agents and goal-based agents, ending with utility-based agents. In utility-based agents we find an internal representation of

the environment, a prediction of what it will be like and how near the goal it will be after the agent's actions as well as a measure of utility, i.e. a way of expressing preferences among various states of the environment (or the agent's 'world').

The utility function can be considered a measure of the performance of an agent and is the base for Russell and Norvig's definition of 'rational agent': an agent can be called 'rational' if – given the input from the environment and its internal knowledge – it will select the action or actions expected to maximize its performance measure, or utility function. The above is a paraphrase from the original.[3]

Finally, we should stress that, for an agent to be truly intelligent, the internal knowledge and the utility function itself should change over time responding to the experience acquired, or, in other words, an autonomous agent should learn from experience. This can even include modifications of the goal – the target – and can have deep ramifications for autonomous agents employed in cyber offence operations.

We should therefore round up the previous four points with two more, to achieve a complete definition of a truly autonomous intelligent agent:

5. An autonomous agent should possess an adequate internal model of its environment, including its goal – expressed possibly in terms of world-states – together with some kind of performance measure or utility function that expresses its preferences.
6. An agent must possess the capability to learn new knowledge and the possibility to modify over time its model of the world and possibly also its goals and preferences.

B. TAXONOMY

Artificial autonomous agents can first of all be divided into the two already mentioned classes of robotic and computational agents. Within the class of computational agents, i.e. purely software agents or 'softbots', we propose a further classification based on two coordinates that can usefully be incorporated into policies and strategic and tactical cyber operations doctrines.

Based on their role, autonomous agents can be employed in intelligence-gathering or in purely military operations: the main difference lying in the destructive nature of military operations, while usually intelligence-gathering does not cause damage to the targets and in fact tries to avoid detection in most instances. Based on architecture, autonomous agents can be either monolithic or decentralised. Monolithic agents are constituted by a single piece of software or else by strictly coordinated elements

without independent means of operation, for instance an executable file and some software libraries or data needed for it to work. Decentralised intelligent agents are systems where the intelligence is distributed among many simpler components, all similar or very similar, acting in concert, in a way similar to the artificial life 'flocks' developed by Craig Reynolds.[4] A decentralised agent can arguably be more resilient to disabling efforts and counterattacks: for instance, a botnet made of agents instead of conventional malware software would not present a central point of control that could be disabled.

Truly autonomous agents used in cyber warfare are not known at this time, at least in unclassified sources. It can be argued that, at the present stage, we are on the verge of seeing actual agents deployed, but for now the most advanced cyber weapon known – Stuxnet – falls short in many of the attributes we postulated for an autonomous (computational) agent:

- It does not possess any representation of its environment, for instance the topology of the network it is running on.
- Its action does not present a continuity in time.
- It does not have any learning capability.
- It does not perform an autonomous target evaluation or selection. It is true that it is capable of selecting systems according to a set of targeting criteria, but the set is fixed at programming time and not subject to expansion or modifications, exactly because no learning is involved.

C. THE ENVIRONMENT OF SOFTBOTS: 'CYBERSPACE'

Physical autonomous agents, or 'robots', operate in the normal physical environment, while computational agents operate in a unique environment, what is commonly called 'cyberspace'.

A valid definition of cyberspace is given in the White House Cyberspace Policy Review published in 2011. According to this document, cyberspace is 'the interdependent network of information technology infrastructures, and includes the Internet, telecommunications networks, computer systems and embedded processors and controllers in critical industries'.[5] Cyberspace is unique as an environment in many ways, but most of all because it is man-made, and not by a simple subject: it is constructed, maintained and operated by a plurality of stakeholders, public and private and with a multitude of somewhat conflicting interests and incentives.[6]

Following Russell and Norvig's characterisation of environments, we can list other peculiarities of cyberspace:

- Cyberspace is a partially observable environment, meaning that agents can have, at the best of times, knowledge of only a tiny fraction of it.
- It is a deterministic and discrete world, but of enormous complexity.
- For agents engaged in cyber warfare the environment is obviously adversarial, including enemy operators, enemy and foreign agents, and conventional security software like firewalls, Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs).
- Cyberspace is dynamic, meaning that the agent's environment can and will change during its operation.

3. CYBER ATTACK SCENARIOS FOR AUTONOMOUS AGENTS

Most of the public debate about cyber warfare policy and strategy in recent years, including academic production, has concentrated on defence. There are valid reasons for that, including: the obvious secrecy shrouding offensive tools and procedures; political reasons; access to information; and a bias towards defence in the West, always shy of appearing as the aggressor. In fact, in cyber warfare, not only does offence have a place in strategy, but an offensive stance – or at least an active defence – is probably preferable. Also, an offensive stance is easier to adopt if the perceived costs – in terms of casualties but also monetary and political costs – are very low compared to other forms of warfare.[7]

A. RECONNAISSANCE

All offensive operations begin with reconnaissance, and this first phase of a cyber attack will arguably provide an ideal arena for the deployment of autonomous agents in the near future, at least in two directions: automatic discovery of technical vulnerabilities in target systems or networks and, on another level, gathering of intelligence about them, for instance structure and topology and details of operating systems and applications, up to user details and access credentials.

The discovery of vulnerabilities in the target network, and the development of practical means of leveraging them (called '*exploits*'), are necessities; presently they are manually developed by skilled personnel or acquired on the market. A software autonomous agent will automatically reconnoitre the target, individuate vulnerabilities and develop means of exploiting them: while full automation of exploit development is still not widely available,[8] we can outline some scenarios for the use of agents incorporating such a capability:

- Software agents instructed to target a network. In this case, the operation will proceed beyond the information-gathering phase into actual infiltration; moreover, a target is already selected. The agents, however, will have no need for fixed, pre-programmed methods of penetrating the system, but will analyse the target, select autonomously the points of vulnerability and develop means to use them. The agent will then proceed to actual infiltration of the target system and – if it is operating as part of a decentralised agent – will share the information gathered with the other agents.
- Purely information-gathering agents. In these kinds of operation, the main objective of the agent is the acquisition of information, which will then be sent back to the command and control structure where it will be processed. Technical information about vulnerabilities present and exploits can be entered into a database that can be used by multiple operations.

B. INFILTRATION AND BEYOND

Autonomous agents, as defined here, will be able to 'remember' the information gathered during reconnaissance and use it to plan their infiltration path. One of the possible methods makes use of 'trees' – mathematical structures commonly used to represent AI problems – to model the possible alternatives in a cyber attack. [9] Future agents will conceivably be able to build *ad hoc* tree representations of possible infiltration routes on the fly, as opposed to manually, and apply techniques – some of them already very well established – to plan and execute the infiltration. Internal representation of the environment and possible threats from defenders will make it possible for autonomous agents to be much more 'persistent' than advanced persistent threats (APTs) known today, by allowing them to prevent and react to countermeasures: for agents tasked with intelligence-gathering, this will mean more time to do so, and agents tasked with disruption will enjoy much more flexibility in selecting specific targets (applications or systems) and means of attacking them. The selection of, for example, specific databases or documents to retrieve once the agent gains access would be achieved through AI techniques that can extract and process information even from unstructured data.

C. SWARMS

Decentralised agents, according to the taxonomy presented above, are sets of cooperating autonomous agents, that can form a whole new kind of botnet, where there is no need for a centralised command and control and individual agents can cooperate, amongst other things, by sharing information. A botnet of this kind would be very difficult to disable, because each single agent would be separately individuated and sanitised.

D. COMMAND AND CONTROL

From an operational standpoint, the difficult problem of command and control of autonomous agents should be stressed. On the one hand, the agent's goal and targets should be pre-programmed and precisely stated in order to facilitate the agent's task and stay as much as possible within legality. On the other hand, it will be tempting to leave free rein to an extent to agents, for instance regarding targets of opportunities. The above concerns the initial phase of developing and deploying an autonomous agent, but it should also be decided to what extent the agent could communicate with its 'base' during its mission, and if that communication should be monodirectional – intelligence originating from the agent, for instance – or bidirectional, i.e. if the command and control structure could issue 'orders' and instructions, including target selection and even self-destruct commands. It is obvious that a whole doctrine including detailed tactics, techniques and procedures (TTPs) for intelligent autonomous agents will have to be developed in order to be ready to integrate these new tools into a state's arsenal.

4. THE LEGAL LANDSCAPE

There has been general discussion on the application of international law to the case of cyber warfare, but a consensus on precise terms has not yet been reached: the possibility in the very near future of the emergence of true autonomous agents pushes the debate still further and shows even more clearly the limits of existing international law in the realm of cyber warfare. While it may be considered far-fetched at the least, and bordering on science fiction, to be discussing right now the legal implications of the use of intelligent autonomous agents as cyber weapons, information technology has a history of preceding the law by far and maybe it is not wrong to engage in such a debate earlier than usual.

The international body of law governing warfare basically consists of *jus ad bellum*, which regulates the resort to force by states; the International Humanitarian Law or *jus in bello*, which concerns itself with the actual conduct of armed conflicts; and the law of neutrality.

A. JUS AD BELLUM

The main source of the international *jus ad bellum* is the Charter of the United Nations, signed in San Francisco in 1945, and its successive amendments. Article 2(4) of the Charter concerns the use of force by Members: 'All Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state [...]'.

It is commonly accepted that this article of the Charter applies to cyber warfare too, the effects of which are 'comparable to those likely to result from kinetic, chemical, biological or nuclear weapons',[10] and in this regard autonomous agents are no different from any other tool or cyber weapon employed. Where the advent of true autonomous agents could really require new interpretations or new formulation is in the question of agency, i.e. 'the attributability of individual conduct to a state'.[11] An autonomous agent will act – up to a point – independently from its developers, at least as regards the details of its actions: we should ask ourselves if the notion of an individual acting as a 'state agent' should be extended to autonomous agents. The notion of states and their sovereignty is central to the Charter and all international laws of war, even if, post-9/11, this is somewhat less true and non-national actors have been accepted as, for instance, capable of waging war, as in the case of terrorist networks. Theoretically, a true autonomous agent could exceed its assigned tasks and engage in what could legally be defined as 'use of force': in this case, should the nation state behind the agent's creation be deemed responsible? The problem of attack attribution, so important for cyber warfare in general ,is even more problematic for attacks realised by an autonomous agent, if only for the fact that its creators themselves possibly would not have known in advance the precise technique employed, or even the precise system targeted, because of autonomous decisions taken by the agent during its operations. In other words, command and control of a true autonomous agent, especially a purely computational one, can be hard to achieve and would have to translate chiefly in precise specifications of the agent's target and objectives – the goals – or, in military terms, in precise briefings before any mission.

Another question that should form part of the debate on the legal aspects of autonomous agents in cyber warfare is whether they can be considered *per se as* 'state agents'. Again, discussion of a similar concept can seem far-fetched at this time, but we are not far from the deployment of real agents and policy-makers should be made aware of the implications. A true intelligent autonomous agent, as defined above, would possess the capability to make decisions based on its belief state of the moment and its assigned tasks, so it is reasonable to consider it a 'state agent' in a legal sense. If so, it seems reasonable to argue that a way should be found to distinguish whether a software agent is to be considered civilian or military, in a technical and a legal sense. In the case of the postulated physical agents, it is easy to assume that – as in the case of remotely controlled 'drones' – they would sport national identification marks, but what about software bots? If ever an international treaty on cyber warfare is signed, it should contain provisions for the identification of autonomous agents, maybe through mandatory signatures or watermarks embedded in their code.

The United Nations Charter does not clearly forbid the use of force in any case, but implicitly admits it in the case of self-defence, in its Article 51: 'Nothing […]

shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member […]'. While, customarily, only attacks by national actors were deemed to be covered by the provisions of this formulation, after the events of September 11, 2001 a new concept emerged, heralded obviously by the US, whereby a nation-state is allowed, in self-defence, also to use force against non-state actors, such as the Al-Qaeda network. In order to be lawful, however, the use of force in self-defence should be governed by some principles, primarily necessity and proportionality. 'Proportionality' in this instance means that the level of force used in self-defence should be no more than what is necessary to repel the threat. Again, in the case of a true autonomous agent, if used as a weapon in self-defence, care should be taken in the command and control function to clearly state the agent's targets and build in appropriate safeguards.

Concerning the individuation of true 'armed attacks', as defined in Art. 51 of the U.N. Charter, and in particular their distinction from lesser instances of use of force ('border incidents'),[12] what is relevant is to determine the actual intent of the state operating the autonomous agent. In this case, the theme of independent behaviour of agents returns. True autonomous agents certainly offer military leaders the broadest possible extent of plausible deniability and it would be difficult to make all actions by an agent the responsibility of its creators.

B. *JUS IN BELLO (INTERNATIONAL HUMANITARIAN LAW)*

This body of law strictly governs actions and activities that happen during actual armed conflicts. Its main sources are the Geneva Conventions of 1949, the Additional Protocols of 1977 and the much earlier Hague Convention of 1907. Together with this written corpus, there is also a rich tradition of customary humanitarian law dating back at least to the Roman *jus gentium*. The dates of the applicable conventions should be enough to emphasise how problematic it is to reconcile the realities of cyber warfare with this established legal landscape, even more so because up to this point we have seen very few instances of actual cyber warfare and none involving a true autonomous agent, whether in international armed conflicts or non-international ones. The recently published *Tallinn Manual*[13] provides a much-needed guide on how international law applies to cyber warfare, even if its scope extends – by the authors' choice – only to existing law (*lex lata*) and how it is applicable to fully-fledged conflicts.

For a cyber operation to be considered an attack under international humanitarian law (IHL), it needs to occur in the context of an armed conflict (hostilities), or, in other words, to have a nexus to a conflict. In the existing law, only in this case is IHL applicable. As the definition implies, IHL is mainly concerned with the protection from violence that should be guaranteed to entities not involved in the

conflict, be they persons or physical assets. So military actions, to be lawful, should avoid – or at least minimise – what can be defined as 'collateral damage': attacks must not be indiscriminate and targets must be carefully selected. Also, the means used (weapons and tactics) should aim to avoid, as far as possible, unnecessary victims and damage to civilian assets. This principle goes under the name of 'proportionality', a somewhat different concept from the principle of the same name present in *jus ad bellum*. Precise targeting of an autonomous agent should not be difficult for computational agents, where, for instance, it can be assured either by protocol parameters such as addresses defining the boundaries of a network or by profiling beforehand what constitutes the agent's target in terms of technical characteristics, application discovered, or types of data.

As in the case of drones, where many have expressed concerns over a so-called 'PlayStation syndrome', the remoteness of the command and control operators could result in less restraint in attack operations. For cyber autonomous agents, a 'remoteness' in time is present, in addition to a distance in space similar to that for the pilots of unmanned aircraft systems, and appropriate doctrines should be developed for the planning, command and control of cyber operations involving autonomous agents. If some form of built-in identification mark is introduced for cyber weapons, this should infuse more responsibility in their planning and operation.

5. CONCLUSION

While, currently, true autonomous agents probably do not yet exist, the technological preconditions for their development are in place and active research is ongoing: it is useful, therefore, to reflect beforehand on the all-round implications of their use in warfare, considering both the technical foundations and the legal implications. The taxonomy proposed here can be considered as a basis for future works, including the place of such agents in doctrine. Agents relying on AI techniques and operating independently would be considered a force enhancer and would make offensive operations even more attractive; anonymity would be enhanced by the independent way in which such agents would operate, conducting attacks without supervision or contact with a command and control structure and for a prolonged period of time; the amount of information needed to launch an attack would be far less than with conventional cyber weapons, because the agent itself would develop its own intelligence, for instance analysing vulnerabilities and developing exploits for them. Also, an autonomous agent, especially if decentralised, would be much more resilient and able to repel active measures deployed to counter it, while conventional malware is somewhat more fragile in this regard.

If – or when – actual autonomous agents are developed and deployed, existing international law will be truly strained and will need adjustments, for instance regarding the possibility of considering a software agent as a 'state agent' as conceived in international law. Work on cooperation and agreements is necessary in order to avoid something similar to an arms race, mainly because the capabilities postulated here for autonomous agents will render cyber offence activities even more attractive than they already are.

Further research is needed both on the technical and the legal side. The author is working on a possible implementation of an autonomous agent using AI techniques and also on further developing the modes in which such agents could be deployed. On the legal side, more work will need to be done on how AI agents would fit into contexts other than a fully-fledged armed conflict, for instance in peacekeeping operations.

REFERENCES

[1] S. Franklin and A. Graesser. 'Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents', in *Proceedings of the Third International Workshop on Agent Theories*, Springer-Verlag, 1996.

[2] S. Russell and P. Norvig. *Artificial Intelligence: a modern approach* 3rd edition, Pearson, 2010.

[3] S. Russell and P. Norvig, op.cit., page 37.

[4] C. Reynolds. 'Flocks, Herds, and Schools: A Distributed Behavioral Model', in *Computer Graphics*, 21(4) (SIGGRAPH '87 Conference Proceedings)' pages 25-34.

[5] White House, *National Security Presidential Directive 54/Homeland Security Presidential Directive 23*.

[6] N. Melzer. *Cyber Warfare and International Law* - U.N. UNIDIR 2011.

[7] R. Anderson, *Why Information Security is Hard – An Economic Perspective*, University of Cambridge 2001.

[8] J. DeMott, R. Enbody and W. Punch, 'Towards an Automatic Exploit Pipeline', in *Internet Technology and Secured Transactions (ICITST)*, 2011.

[9] A. Moore, R. Ellison and R. Linger, *Attack Modeling for Information Security and Survivability* - Carnegie Mellon Software Engineering Institute 2011.

[10] N. Melzer, op.cit.

[11] N. Melzer, op.cit.

[12] N. Melzer, op.cit.

[13] M. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare* - Cambridge University Press 2013.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Autonomous Decision-Making Processes and the Responsible Cyber Commander

Jody M. Prescott
Senior Fellow, West Point Center for the Rule of Law
Adjunct Professor, Department of Political Science, University of Vermont
Burlington, Vermont, USA
jody.prescott@us.army.mil; 01jpresc@uvm.edu

Abstract: With cyber operations conceivably moving at near light speed, commanders in cyber warfare will likely need to rely extensively upon autonomous decision-making processes (ADPs) to be effective. For commanders to meet their obligations under the Law of Armed Conflict (LOAC) and complementary Rules of Engagement (ROE), these ADPs must function in a manner compliant with both. To better understand how such ADPs might be effectively used, it is important to consider the operational challenges cyber commanders face in conducting cyber warfare, the different options available to cyber commanders to decrease the time frame required for making effective, LOAC-compliant decisions, and how ethical ADPs might be created. To that end, this paper focuses on the development of programme architecture and procedures that will be necessary to meet LOAC and ROE requirements, rather than the applicable law or potential ROE.

Keywords: *law of armed conflict, commander, autonomous decision-making processes*

1. INTRODUCTION

From the perspective of the law of armed conflict (LOAC), a cyber commander's battle-space resembles a game of three-dimensional chess in important respects. On the level in which the effects of cyber operations might ripple into the geophysical world and injuries to people and damage to tangible objects may be reasonably foreseen, LOAC will likely apply just as it does during traditional kinetic warfare.[1] This would include criminal responsibility for a cyber commander whose actions or inaction resulted in LOAC violations.[2] On a second level, the one in which the effects of cyber actions remain in cyberspace and result at most in the manipulation or deletion of non-critical data, more cyber annoyance than cyber attack, LOAC would not appear to apply at all.[3] Instead, the governing authorities are likely national rules of engagement (ROE). There is likely a third level in between these two; the one in which the direct effects of cyber actions remain in cyberspace but very serious indirect effects register in the geophysical world as a result of the degradation or destruction of critical national cyber infrastructure. U.S. cyber strategy documents[4] and statements of Department of Defense (DOD) officials[5] suggest that LOAC-like principles embedded in ROE might be part of the decision calculus governing whether and how to respond to such cyber activities.

Thus, although each level is in play simultaneously, unlike a game of three-dimensional chess, different rules apply to each level. Further, unlike the deliberative pace of chess, the operational tempo of cyberspace action is much more intense and capable of moving at almost the speed of light.[6] As a matter of operational necessity, cyber commanders will need to rely extensively upon autonomous decision-making processes (ADPs) that conduct cyber response activities and operations as the

---

1 Tallinn Manual on the Law of Cyber Warfare, Rule 13, para. 6, 55; Rule 29, para. 1, at 91; Rule 30, para. 5, 93.

2 *Id*., Rule 24, at 80.

3 *See id*., Rule 13, para. 6, 55 ("acts of cyber intelligence gathering and cyber theft, as well as cyber operations that involve brief or periodic interruption of non-essential cyber services, do not qualify as an armed attack.").

4 Department of Defense, Cyber Policy Report Pursuant to Section 934 of the NDAA of FY 2011 (Nov. 2011), 3-8, *available at* http://www.defense.gov/home/features/2011/0411_cyberstrategy/docs/NDAA%20Section%20934%20Report_For%20webpage.pdf [hereinafter "Cyber Report"].

5 *See* Ellen Nakashima, *Pentagon proposes more robust role for its cyber-specialists*, washingtonpost.com (Aug. 9, 2012), *available at* http://www.washingtonpost.com/world/national-security/pentagon-proposes-more-robust-role-for-its-cyber-specialists/2012/08/09/1e3478ca-db15-11e1-9745-d9ae6098d493_story.html; Amber Corrin, *Cyber warfare: New Battlefield, new rules*, FCW.com, Jul. 9, 2012, *available at* http://fcw.com/articles/2012/07/15/feat-inside-dod-cyber-warfare-rules-of-engagement.aspx; William J. Lynn, Remarks on the Department of Defense Cyber Strategy, speech made in Washington, D.C., (July 14, 2011) *available at* http://www.defense.gov/Speeches/Speech.aspx?SpeechID=1593.

6 Thomas C. Wingfield et al, *Optimizing Lawful Response to Cyber Intrusions*, 2 (2005) (unpublished paper), *available at* http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA464203 .

product of computer-driven decision cycles lasting perhaps no more than fractions of a second. There appear to be misgivings in general about the use of ADPs when human targets are involved.[7] From a strictly operational perspective, however, the risk of not utilising ADPs to conduct cyber operations might be unacceptable, and the temptation to accelerate the pace of decision-making by reducing the role of the human commander might be very strong.[8] As defensive and offensive cyber measures become more sophisticated, the demarcation between the two might become more blurred,[9] and the issues regarding the propriety of these different uses of force more intertwined.

To better understand how cyber commanders might use ADPs in a LOAC-compliant manner, this article first identifies the operational challenges faced by current cyber commanders. Second, the different options available to compress the time frames within which cyber commanders must make their decisions are explored, and the relative advantages and disadvantages of each assessed. Third, on-going research in a related field – the development of autonomous geophysical robot weapons – is then examined to highlight the challenges involved in seeking to embed LOAC and ROE principles, prohibitions and permissions into cyber ADPs. In conclusion, this paper suggests that in combination with other options to compress a cyber commander's decision timeframe, LOAC- and ROE-compliant ADPs could constitute an effective and necessary means by which the obligation of command responsibility is met in cyber operations, and that steps should be taken immediately to ensure that LOAC principles are incorporated into ADP design processes.

2. OPERATIONAL CHALLENGES FACING CYBER COMMANDERS

In the most comprehensive study of its kind available in the public domain to date, a senior U.S. naval officer surveyed a number of senior U.S. officers who had cyber operations experience and who served on the staff of the U.S. Chairman of the Joint

[7] *See* Ronald C. Arkin, Governing Lethal Behavior in Autonomous Robots, 52-55 (2009) (in a 2008 survey of 430 roboticists, military personnel, members of the general public, and policy makers, over half of the respondents found the taking of human life by an autonomous robot unacceptable). 66% of those surveyed believed that the robot should be held to higher ethical standards than soldiers. *Id.* at 55.

[8] Thomas K. Adams, *Future Warfare and the Decline of Human Decisionmaking*, 31 Parameters 57, 66 (Winter 2001/2002).

[9] Christopher Ford, *Cyber Operations: Some Policy Challenges*, newparadigmsforum.org (June 3, 2010), *available at* http://www.newparadigmsforum.com/NPFtestsite/?p=270 (last visited Dec. 16, 2012).

Chiefs of Staff. The survey included 15 men and six women,[10] who had on average 22 years of military service, over one year of cyber warfare decision-making experience, and almost two and one-half years' of service on the Joint Staff.[11] All of these officers had master's degrees, over half held two or more master's degrees, and 14% had professional degrees.[12] On the basis of their experiences, the study participants identified several significant concerns they had with conducting cyber operations. Their concerns included the uncertainty that results from the complexity of cyber operation response processes, the technical challenges in discriminating between military objectives and civilian objects, the difficulty in applying LOAC and ROE to cyber operations, and importantly for purposes of this article, a sometimes troubling perception of the duality of virtual agents with their geophysical personae.

A. COMPLEXITY

The officers surveyed believed that the uncertainty they had experienced in responding to cyber-attacks resulted in part from the complexity of the response processes they used, and that understanding the response processes required "an in-depth mastery of cyber warfare tactics, techniques and procedures."[13] This complexity led to ambiguity in lines of authority and actionable thresholds of adversary activity, or "red-lines,"[14] in determining a proper response.[15] This internal "fog of war" in the decision-making process was exacerbated by the lack of scalable response options,[16] which the officers believed limited the ability to respond to the wide range of cyber-attacks.[17]

---

[10] Daryl L. Caudle, Decision-Making Uncertainty and the Use of Force in Cyberspace: A Phenomenological Study of Military Officers, 221, DTIC X (Oct. 14, 2010) (Ph.D. dissertation, University of Phoenix), *available at* www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA534888. The author is grateful for CAPT Caudle's insightful comments and suggestions on this paper.

[11] *Id.* at 225.

[12] *Id.* at 226.

[13] *Id.* at 253.

[14] Accordingly, cyber response decisions are complicated by the assessment of tradeoffs between operational gain and intelligence loss. *Id.* at 261. *See* Ellen Nakashima, *Dismantling of Saudi-CIA Web site illustrates need for clearer cyberwar policies*, washingtonpost.com (Mar. 19, 201), *available at* http://www.washingtonpost.com/wp-dyn/content/article/2010/03/18/AR2010031805464_pf.html (conflict between U.S. Army and the CIA on whether to take down a joint Saudi intelligence-CIA web site used to collect information on potential jihadists, but also used by jihadists to plan operations against U.S. Army units).

[15] Study participants assessed that the planning and conduct of cyber operations is hampered by unwieldy targeting processes and competing interagency interests, a lack of transparency among other government agencies, and unnecessarily classifying information at too high a level compound these problems. Caudle, *supra* note 10, at 255, 261, 263.

[16] *Id.* at 253.

[17] *Id.* at 254.

The study officers also noted the negative impact of an external aspect of complexity in cyberspace; chaos. Chaos describes the phenomenon observed in dynamic, complex systems in which small variations among initial inputs into the systems lead to large variations among the results of these inputs, often in a seemingly random manner.[18] These systems are deterministic, however, and "normally achieve equilibrium around a confined region of space, called a *strange attractor*, where the system permanently resides,"[19] i.e., the precise result cannot be predicted, but it will be of a reasonably foreseeable nature. In practical terms, the impact of chaos on the battlefield has long been recognised.[20] For a cyber commander, chaos means that the same response actions within very similar operational contexts will not always have the same effects, and this uncertainty as to unintended effects further complicates decision-making.[21]

B. TECHNICAL CHALLENGES

"Because cyber warfare is an emerging, non-kinetic, asymmetric warfighting discipline that occurs in a virtual domain, leaders lack experience; moreover, physical effects from the actions they take are not always observable."[22] This likely makes it more difficult in advance of an attack for a responsible cyber commander to make an accurate assessment of the battle-space which would be necessary to support a proper analysis of proportionality and military necessity.[23] Likewise, battle damage assessment is more complex than in the geophysical world,[24] and this likely has a feedback effect on cyber commanders' ability to learn from their experiences in terms of making future assessments of proportionality and military necessity.

Creating effective and integrated hardware and software that would allow a cyber commander to visualize the area of cyber operations accurately will likely require the automatic analysis of multiple sources of data, and the combination of analysis

---

18 *Id.* at 131.

19 *Id.*

20 Alan Beyerchen, *Clausewitz, Nonlinearity and the Unpredictability of War*, 17 Int'l Sec. 59, 70 (1992).

21 *See* Simon R. Atkinson & James Moffat, The Agile Organization: From Informal Networks To Complex Effects And Agility, 77-84 (2005) (Networked decision-making by Allied forces in the Battle of the Atlantic more capable of dealing with intricacy and uncertainty than German forces), *available at* http://www.au.af.mil/au/awc/awcgate/ccrp/atkinson_agile.pdf.

22 Caudle, *supra* note 10, at 76.

23 Neil C. Rowe, *The Ethics of Cyberweapons in Warfare* (2009), *available at* http://faculty.nps.edu/ncrowe/ethics_of_cyberweapons_09.htm (last visited Dec. 29, 2012).

24 *Id.* Participants in the study believed that "undefined information valuation standards (i.e., clear, consistent, and generally accepted expression of the worth of information)" impeded the conduct of battle damage assessment after a cyber-attack. Caudle, *supra* note 10, at 256.

and data from these multiple and diverse sources.[25] These systems would also need to be able to manage the sensors that collect information about the cyber battle-space, and must be essentially defect-free in the performance of their gathering, managing, and analysis functions. Further, the integrity of the data upon which the systems rely must be defended. Such reliability would allow cyber commanders to trust these systems in making their own decisions, rather than second-guessing the systems.[26]

The surveyed officers were also very concerned with the technical challenges they encountered in discriminating between those things that appeared to be valid military targets and those which were protected because of their civilian nature.[27] Identifying a cyber attacker both accurately and quickly enough to allow for an effective response might be one of the most ambiguous and difficult hurdles to overcome in waging LOAC-compliant cyber warfare.[28] Additionally, difficulty in discerning patterns in attacks as to source and potential severity increased the cyber officers' concerns as to causing unintended, higher level effects, and led to their decision-making cycles being prolonged.[29] More pointedly, the officers believed "the level of attribution certainty required to respond to a cyber-attack [was] arbitrary and unrealistically high."[30]

C. LOAC AND ROE APPLICATION

The study participants acknowledged the applicability of treaties, laws and policy directives to decision-making following a cyber-attack, but found these authorities contributed to the uncertainty in formulating the appropriate response.[31] In particular, the officers "found applying the conventional rules of warfare in

---

[25] Computational Methods for Decision-making, Special Notice 12-SN-0009, Special Program Announcement for 2012 Office of Naval Research, 1-2 (2012) (request for proposals), *available at* http://www.onr.navy.mil/~/media/Files/Funding-Announcements/Special-Notice/2012/12-SN-0009.ashx (last visited Dec. 28, 2012) [hereinafter "Computational Methods"]. General Alexander, the NSA Director and U.S. Cyber Command commander, has stated that it is difficult to obtain a common operational picture of the relevant portions of cyberspace in real time to support operations. Ford, *supra* note 9.

[26] *Id*. at 1-2. Conflicts in equities between agencies regarding cyber actions may be compounded by the difficulty in having stakeholders able to "visualize cyber war and cyber damages." *See* Caudle, *supra* note 10, at 256

[27] Caudle, *supra* note 10, at 254.

[28] Matthew M. Hurley, *For and from Cyberspace: Conceptualizing Cyber Intelligence, Surveillance, and Reconnaissance*, 26 Air & Space Power J. 12, 19 (2012).

[29] Caudle, *supra* note 10, at 254. Surveyed officers noted that decision-making uncertainty "is strongly influenced by the 'fog of war' created in cyberspace resulting from advanced deception capabilities and methods." *Id.* at 255.

[30] *Id.* at 260.

[31] *Id.*

cyberspace to be challenging and not straight forward."[32] Although they appeared familiar with literature suggesting that extant LOAC was applicable when "cyber attack [could] be represented by equivalent kinetic attack characteristics," the officers did not believe the current legal framework addressed "sovereignty challenges, jurisdiction boundary problems (e.g., cloud computing and transborder data flows), non-state actors, severity thresholds, and the technical nuances of cyber attack."[33] The officers noted significant concerns with the ROE under which they had operated, which they assessed as "nascent and generally untested,"[34] and in particular they found that "the lack of practical definitions for hostile intent and hostile act in cyberspace" made consistent cyber responses difficult.[35]

D. VIRTUAL AND GEOPHYSICAL DUALITY

For purposes of this article, perhaps one of the most interesting findings of this study was the cyber warfare officers' perception of the "complex duality of [cyberspace's] virtual and physical nature."[36] This both complicated the understanding of higher-order effects in decision-making, and the deconfliction of "traditional military activities from intelligence gathering activities."[37] This duality is further reflected in their perception of their adversary human counterparts being virtually coupled with their digital agents.[38] To improve decision-making, the respondents believed that policies and ROE that dealt with cyber actions conducted solely within cyberspace were necessary to "depersonalise" cyber-attacks, and that this depersonalisation would reduce uncertainty and make cyber responses quicker.[39]

E. SUMMARY

The scope of the challenges identified by the cyber warfare officers illustrates just how very different cyber conflict is from geophysical conflict, as well as the clash between these differences and the perfectly human desire to understand cyber conflict in terms consistent with, or at least analogous to, geophysical human

[32] *Id.* at 255. Study participants described the current legal framework as "antiquated and inadequate to support military operations in an effective manner." *Id.* at 261.

[33] *Id.* at 262.

[34] *Id.* at 260.

[35] *Id.* at 262.

[36] *Id.* at 266. As one writer has noted, "[i]n the virtual world, when we refer to an enemy or an opponent, we may actually be referring to what really are the second and third order effects of the actual activity of our opponent, or even beyond." The Basics of Cyber Warfare: Understanding the Fundamentals in Theory and Practice, 67 (Steve Winterfield & Jason Andress, eds.,2012).

[37] Caudle, *supra* note 10, at 266.

[38] *Id.* at 256.

[39] *Id.* at 258.

experiences. This suggests that there will likely be a heavy burden upon system designers to accommodate the human aspect of decision-making in the development of hardware and software intended to support cyber commanders and operators. In particular, the officers' concerns as to the perceived duality of cyberspace actors suggests that proposed solutions must be mindful of the need to deliberately and explicitly differentiate between cyber agents and their geophysical personae when it would increase operational efficiency, and to foster this duality when it would have the same effect.

3. BUYING TIME: ALLOWING THE COMMANDER TO BE RESPONSIBLE

Given the concerns detailed by the surveyed officers as to the conduct of cyber operations, and the misgivings of many regarding the use of autonomous systems waging war, it is useful to consider alternative methods of compressing the temporal aspect of cyber decision cycles while ensuring the responsible cyber commander remains in the decision loop. These methods include improved cyber intelligence, surveillance and reconnaissance (ISR) capabilities, innovative staffing techniques, enhanced human-computer interfaces (HCIs), and possibly even brain-computer interfaces (BCIs).

A. *CYBER ISR*

In light of the current emphasis on defensive and offensive cyber operations in the U.S. national and military doctrine and policy, one writer has assessed the ISR aspect of cyber operations as particularly needing "doctrinal, educational, and organizational concepts that forcefully emphasize the centrality and operational nature of cyber ISR."[40] Cyber ISR can take many forms, with different levels of intrusiveness into potential adversaries' cyber infrastructure. At one end of the intrusiveness spectrum, "honeypots" or "honeynets" can be emplaced in a cyber system's defences to lure intruders to penetrate them instead,[41] thereby providing a cyber commander with advanced warning of potential attacks. At the other end of the spectrum is the use of "active defence" mechanisms that operate within an adversary's cyberspace, "exploit[ing] [its] cyberspace vulnerabilities while gaining

[40] Hurley, *supra* note 28, at 20-21. Enhanced cyber ISR would likely need to include robust indications and warnings (I &W) intelligence processes to be effective. *See* Chairman, Joint Chiefs of Staff, Joint Publication, 2-0, Joint Intelligence , I-16-17 (Jun. 22, 2007) (I & W intelligence is "very time-sensitive" forewarning of adversary actions or intentions).

[41] *See* Lance Spitzner, Honeypots: Tracking Hackers, 50-71 (2002) (discussing the operational value of honeypots and honeypot networks).

a deeper understanding of the enemy's decision cycle and defensive weaknesses."[42] Theoretically, this too should allow more time for responsible cyber commanders' to make decisions.

Such measures, however, inevitably lead to effective countermeasures.[43] Rather than compressing decision cycles through the provision of better informational inputs, the continuous ISR race between cyber adversaries might in the end only result in maintenance of the decision cycle status quo. Further, the use of active defence measures could result in a cyber adversary interpreting such probing as an indication of hostile intent, or a hostile act,[44] and decide to engage in an unexpected, forceful counter-response it might otherwise not have conducted.

B. STAFFING TECHNIQUES

A second approach would be the acceleration of cyber commanders' decision-making processes through innovative staffing techniques. This could include the use of multiple planners and multiple commanders with cyber weapons release authority working within a cyber operations centre, each supported by teams of technical advisers (TEKADs), political advisers (POLADs), and legal advisers (LEGADs).[45] The use of teams of commanders would allow multiple emergent situations to be dealt with simultaneously. Although weapons release authorities and processes are not currently structured this way for kinetic operations, multiple commanders could possibly be tiered, so that increasing levels of likely incidental cyber or geophysical damage or injury could be handled by progressively senior commanders. The commanders' reaction times could likely be reduced if they were supported by dedicated adviser teams with whom they habitually trained and operated, particularly if they were using clear ROE.[46] Unfortunately, whilst such staffing measures could possibly reduce a cyber commander's decision cycle by minutes, the operational flow within cyberspace might be moving too fast for such a reduction to make a meaningful difference.

---

[42] Caudle, *supra* note 10, at 77.

[43] *See, e.g.*, Neil C. Rowe & Han C. Goh, *Thwarting Cyber-Attack Reconnaissance with Inconsistency and Deception*, Proceedings of the 2007 IEEE Workshop on Information Assurance, 151, 151-58 (U.S. Military Academy, West Point, NY, June 20-22, 2007).

[44] *See* Timothy L. Thomas, *China's Electronic Long-Range Reconnaissance*, 87 Military Review 47, 47-54 (Nov./Dec. 2007) (discussing suspected Chinese cyber intrusions in the context of Chinese cyber strategy and reconnaissance's role in the Chinese concept of "active offense").

[45] *See* Tallinn Manual, *supra* note 1, Rule 52, para. 6, 158 ("mission planners should, where feasible, have technical experts available to assist them in determining whether appropriate precautionary measures have been taken.").

[46] General Alexander has suggested that clear ROE might speed up cyber decision-making. Ford, *supra* note 9.

C. HCIs

A third approach, heightening the intimacy of the connections between cyber commanders and their computer systems,[47] could occur in two primary ways: enhanced human-computer interfaces (HCIs) and brain-computer interfaces (BCIs). As to HCIs, the software that creates the operational picture for the commander might be tailored to that specific commander's personality traits. Research has shown that students using computer interfaces that recognized their individual learning styles showed higher learning results.[48] Further, technologies such as deep-learning programmes, which use artificial neural nets to imitate the way the human brain learns, offer the promise of computers that could both recognize patterns in large amounts of information and then communicate this to humans via speech.[49] Such programmes have already displayed what appears to be the capability to learn as they recognize patterns.[50]

The use of HCIs that allow interactions similar to how human-human interactions occur, so that computers could potentially understand and or anticipate human intentions,[51] would conceivably allow for a commander to react more quickly to emergent situations in cyberspace. Through enhanced communication, these systems could shorten cyber commanders' decision cycles by presenting cyberspace visualisations specifically attuned to particular commanders' problem-solving and interaction styles. As with staffing innovations, however, the decreases in time might simply not add any operational advantage.

D. BCIs

At one level, given the physical invasiveness of some BCI techniques, such connections resemble science-fiction, but recent advances in this field have been remarkable. For example, a user has been able to move an automated prosthetic arm and grasp items simply through thought, as her brain activity was registered

---

47 Adams, *supra* note 8, at 66.

48 Edmond Abrahamian, Jerry Weinberg, Michael Grady, and C. Michael Stanton, *Is Learning Enhanced by Personality-Aware Computer-Human Interfaces?,* Proceedings of I-KNOW '03, 226, 228-29 (Graz, Austria, July 2-4, 2003).

49 John Markoff, *Scientists See Promise in Deep-Learning Programs*, NYTimes.com (Nov. 23, 2012), *available at* http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html?_r=0.

50 John Markoff, *How Many Computers to Identify a Cat?* 16,000, NYTimes.com (June 25, 2012), *available at* http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all.

51 Hayrettin Gürkök & Anton Nijholt, *Brain Computer Interfaces for Multimodal Interaction: A Survey and Principles*, 28 Int'l J. Human-Computer Interaction 292, 292-93 (2012).

by microelectrodes implanted in her brain.[52] Certain techniques do not require physical contact between users' brains and the computer systems, but instead monitor brain activity through contact sensors on the users' scalps.[53] The potential melding of multimodal interaction techniques, in which computers use multiple sensors to gather physical information about interacting human partners (e.g., cameras to watch hand movements and eye gaze, microphones to register spoken commands) with BCIs that directly track human brain activity[54] offers a possible future means to further speed up a commander's decision-making.

On balance, however, ethical and technological challenges suggest that this approach is likely to be of limited use in cyber conflict in the near term.[55] Further, invasive interfaces would appear to raise the possibility of directly targeting the human operator through cyber attack. Creating a vulnerability that allows the specific targeting of a highly trained commander makes little operational sense.

E. SUMMARY

Use of these different approaches, possibly in combination, could yield significant improvements in the response times of cyber commanders while ensuring their actions remain LOAC-compliant. Currently, however, it does not appear that any single approach or combination of approaches would satisfy the operational imperative to be able to respond to emergent cyber situations as quickly as ADPs could. Therefore, it is important to explore how ADPs might be constructed and employed in cyber operations to provide commanders means of effective engagement.

4. BUILDING LOAC-COMPLIANT ADPs

The increasing use of military unmanned aerial vehicles (UAVs) operated by human controllers has triggered constructive discussion regarding the ethics, legality, and practicality of such weapon systems becoming autonomous.[56] The issues raised

---

[52] Jennifer L. Collinger et al, *High-performance neuroprosthetic control by an individual with tetraplegia*, thelancet.com, 6-7 (Dec. 17. 2012), *available at* http://dx.doi.org/10/1016/S0140-6736(12)61816-9.

[53] Alessandro Pressaco et al, *Neural decoding of treadmill walking from non-invasive, electroencephalographic (EEG) signals*, J. Neurophysiology, 5-6 (July 13, 2011), *available at* http://jn.physiology.org/content/early/2011/07/11/jn.00104.2011.full.pdf+html.

[54] *See* Gürkok, *supra* note 51, at 303-04.

[55] *See* Jens Clausen, *Moving minds: Ethical Aspects of neural motor prostheses*, 3 Biotechnology Journal 1493, 1496-98 (2008), *available at* http://www.yorku.ca/lsergio/Clausen_MovingMindsBTJ2008.pdf (medical complications, interference with personality and personal identity, and responsibility for malfunctions are among the ethical issues raised by BCI).

[56] Arkin, *supra* note 7, at 21-25.

therein are directly applicable to the proper relationship between cyber conflict ADPs and human commanders. Consideration of the advantages and disadvantages of using such systems, the potential architecture of LOAC-compliant ADPs, and the human-robot interaction (HRI) aspect of the operation of these ADPs all suggest that the role of the responsible cyber commander could be appropriately factored into their design and use.

A. *POSSIBLE ADVANTAGES OF ADPs*

ADP proponents argue that these systems could enhance the observance and application of LOAC and complementary ROE in conflicts.[57] First, human combatants endure physical pressures in conflict that degrade human perception and the rational decision-making based upon it.[58] These stressors generate negative emotions that further degrade both perception and cognition.[59] ADPs would not be subject to these physical stressors, and the programmed responses would not be clouded by emotion.[60] Second, because of their potential to receive and integrate vast amounts of information quickly from multiple sensor systems, the decisions made might be based on a more complete picture of the cyber operational area than a human could comprehend.[61] Further, this operational picture could be evaluated without "the human psychological problem of 'scenario fulfilment'" in which "humans use new incoming information in ways that only fit their pre-existing belief patterns.'"[62] Third, ADPs could serve as independent witnesses of the cyber action, whose decisions to record and report potential violations of LOAC and ROE are not subject to concerns of disloyalty to fellow soldiers.[63] Fourth, one human operator might be capable of simultaneously overseeing multiple ADPs.[64]

---

[57] *Id.* at xv-xviii.

[58] Helmet-Mounted Displays: Sensation, Perception and Cognition Issues, 675-749 (Clarence E. Rash et al, eds., 2009).

[59] *See* Arkin, *supra* note 7, at 33-36 (robots would not engage in irrational thinking that tends to dehumanize adversaries and excuse their lethal engagement on the basis of genocidal, penal, or utilitarian rationales).

[60] Ministry of Defence, Joint Doctrine Note 2/11, The UK Approach to Unmanned Aircraft Systems, 5-11 (Mar. 30, 2011), *available at* http://www.mod.uk/NR/rdonlyres/F9335CB2-73FC-4761-A428-DB7DF4BEC02C/0/20110505JDN_211_UAS_v2U.pdf. [hereinafter "UK Approach"].

[61] Arkin, *supra* note 7, at 29-30; Gary E. Merchant et al, *International Governance of Autonomous Military Robots*, 12 Colum. Sci. & Tech. L. Rev. 272, 279-280 (2011).

[62] Arkin, *supra* note 7, at 29-30; Merchant, *supra* note 61, at 279-80.

[63] *Id.*

[64] UK Approach, *supra* note 60, at 5-10.

B. *POSSIBLE DISADVANTAGES OF ADPs*

Critics of ADPs believe that their use would actually result in fewer LOAC-compliant decisions. First, they suggest that human emotions are actually a safeguard against killing,[65] because currently only humans have the ability to "bring empathy and morality to complex decision-making,"[66] and that the human ability to factor emotion into assessments of hostile intent is crucial when decision-makers are dealing with human behaviour.[67] Second, human operators might experience "automation bias," and be unwilling to challenge an ADP's assessment or action.[68] Third, ADPs cannot be made sophisticated enough to make the context-dependent assessments that human commanders make on the basis of incomplete information, such as proportionality.[69] Similarly, ADPs will not have sufficient capability to distinguish between civilians and combatants.[70] Even assuming sufficiently sophisticated software could be developed to allow ADPs to make such decisions, perhaps using artificial intelligence,[71] "[c]omputer programs do not behave as predictably as software programmers would hope."[72] Complex systems are subject to malfunctions, and '[p]ortions of programs may interact in unexpected, untested ways."[73] Complexity itself might generate non-programmed and unanticipated emergent behaviours.[74] Even an ability to learn, which would appear desirable from the viewpoint of creating a system that could adapt to novel situations, "raises the question of whether it can be predicted with reasonable certainty *what* the [ADP] will learn."[75] Further, critics believe the use of ADPs will lead to a "responsibility gap," because there is no fair and effective way to hold humans responsible for the effects of automated decision-making when they had no direct control over the decision-making process.[76]

---

[65] Human Rights Watch & International Human Rights Clinic, Losing Humanity: The Case against Killer Robots, 37 (2012) [hereinafter "Losing Humanity"].

[66] *See* UK Approach, *supra* note 60, at 5-11

[67] Losing Humanity, *supra* note 65, at 31-32; Merchant, *supra* note 61, at 283.

[68] Losing Humanity, *supra* note 65, at 13.

[69] *Id.* at 42; Merchant, *supra* note 61, at 285.

[70] Losing Humanity, *supra* note 65, at 20.

[71] UK Approach, *supra* note 60, at 5-4. "[S]uch operations would present a considerable technological challenge and the software testing and certification for such a system would be extremely expensive and time consuming."

[72] Merchant, *supra* note 61, at 284.

[73] *Id.* at 283-284.

[74] *Id.* at 284.

[75] *Id.* (emphasis in original). This is particularly of concern if the robot operates in an unstructured environment.

[76] Losing Humanity, *supra* note 65, at 42.

C. ARCHITECTURE

Professor Arkin has posited that there are three primary requirements that must be met for an ADP to respond to a situation in conformance with ethical parameters, such as LOAC and ROE. These are the ability to perceive the operational environment correctly, the inclusion of content which identifies the specific types of acts permitted or prohibited under these parameters, and the appropriate representation of this content within the decision architecture.[77] Arkin advocates programming which essentially errs on the side of caution so that the context and nuance upon which a human commander would rely becomes non-relevant in the ADP. For example, a certain continuing level of quality in the situational awareness upon which the ADP relies could be required before it could respond. Requiring this threshold to be met would enhance target discrimination, and prevent the engagement of civilians, civilian objects or friendly forces throughout the course of an engagement.[78] One component of this threshold could be the requirement to have consistent information from multiple sensors.[79]

As to possible ADP responses, Arkin suggests first that the decision architecture should be designed so that ethical responsibility is segregated within it.[80] Arkin sees four separate functions as being necessary to ensure conformance with ethical standards, the first of which would be an "ethical governor," which would "conduct an evaluation of the ethical appropriateness" of any ADP-proposed lethal response.[81] The ethical governor would be complemented by "ethical behavior controls," which would only allow the system to propose responses consistent with LOAC and ROE,[82] and by "ethical adaptors," which would monitor on-going cyber responses and essentially call "cease fire" if certain thresholds were exceeded.[83] With UAVs in the geophysical world, this requirement is satisfied by means of near real-time video feeds that provide commanders and their advisers with an understandable operational picture of the target site.[84] The fourth component would be a "responsibility advisor," which would be "part of the human-[computer]

---

[77] Arkin, *supra* note 7, at 70-91.

[78] *Id.* at 119.

[79] *Id.* at 120-21.

[80] *Id.* at 126.

[81] *Id.* at 127. The governor essentially serves as a cross-check on the response proposed by the ADP. *Id.* at 125.

[82] *Id.* at 133. Further, actions deemed to violate LOAC requirements as programmed would never be undertaken, nor would actions permitted under ROE but in violation of LOAC. *Id.* at 212.

[83] *Id.* at 138.

[84] *See* UK Approach, *supra* note 60, at 5-1, n.2 (UAVs in Afghanistan use the same ROE and targeting guidance as manned aircraft, "but they have the persistence to check and re-check, possibly via legal advisers, that they are compliant" with the ROE).

interaction component" that is used to secure permission from the commander for the ADP to engage in the mission, and to allow commander overrides of ADP decisions.[85]

From an engineering perspective, the legal framework for operating the ADP could essentially be treated the same as other technical and operating requirements at the beginning of the design, so that it could be referenced in the "specification and design of various subsystems, as well as informing the concept of employment."[86] There are different models for ethical decision-making in the context of LOAC and ROE that could be utilised, and this suggests that LEGADs should be part of the software development team.[87] Legal review of the information that would be considered by the ADP would likewise be required, so that the achievable level of situational awareness can be understood[88] in the context of LOAC compliance.

D. HRI

HRI is a relatively new field that addresses in a multidisciplinary manner how people work or play with robots rather than computers or tools, and "[t]his large multi-disciplinary mix presents a very different mindset from traditional engineering design, interface development or ergonomics."[89] Research into HRI so far largely replicates findings from human-human research, and interestingly, shows that "humans expect unmanned systems to meet expectations of a team member with known competences."[90] This is perhaps in its own way a reflection of the cyber/geophysical duality that the surveyed cyber officers noted in their perceptions of cyberspace actors. To meet these expectations, "[m]odels of what operators or decision-makers need to know about the system or state in order to maintain trust in the predictable outcomes from using the system"[91] would need to be developed.

Perhaps because HRI is so new, it is not clear that designers of ADPs fully recognize how important human-friendly perception of the battle-space in which the ADPs operate is for humans to be able to work effectively with the ADPs.[92] As one report

---

85 Arkin, *supra* note 7, at 143.

86 UK Approach, *supra* note 60, at 5-2.

87 Arkin, *supra* note 7, at 95-113.

88 *See* UK Approach, *supra* note 60, at 5-3.

89 Defense Science Board, Department of Defense, Task Force Report: The Role of Autonomy in DoD Systems, 44 (July 2012), *available at* http://www.acq.osd.mil/dsb/reports/AutonomyReport.pdf [hereinafter "Role of Autonomy"].

90 *Id.* at 46.

91 *Id.* at 49.

92 *See id.* at 23 ("For the operator, autonomy is experienced as human-machine collaboration, which is often overlooked in design.").

assessing current autonomous military systems noted, this crucial aspect "is largely ignored and instead erroneously treated as a computer display problem; however, a display cannot compensate for a lack of sensing."[93] Further, "[m]ultisensor integration, either for increased sensing certainty or more comprehensive world modelling, appears to be ignored."[94] As a result, areas of deficiency in human-system collaboration include the lack of "natural user interfaces enabling trusted human-system collaboration and understandable autonomous system behaviors."[95] This could be remedied through the creation of "perceptually oriented interfaces and sensor placement designed around the psycho-physical attributes of the human perceptual system," such as enabling dialogue between humans and ADPs "using natural human interaction modes, especially natural language and gestures."[96]

Not properly addressing the need for optimal human interface in design architecture causes commanders and operators to lack confidence that the systems will operate as they are supposed to, and this lack of trust[97] in turn likely limits the systems' usefulness and the speed at which decisions can be made.[98] This gap in confidence could possibly be remedied by an emphasis on "natural user interfaces and trusted human-system collaboration, perception and situational awareness to operate in a complex battle-space, large-scale teaming of manned and unmanned systems, and test and evaluation of autonomous systems."[99] These improvements would provide the human partner sufficient visibility of the system's activities and how they related to the mission objectives,[100] and would also likely help normalise the legal review process as well.[101]

Concerns of ADP critics to the contrary, the proper assignment of responsibility for the use of ADPs is likely easily resolved – it will lie "with the last person to issue the command authorising a specific activity."[102] Such authorisations can be reliably recorded in a log for audit trail purposes; for example, as part of the preparation process for a cyber commander assuming a watch in an operations centre.[103] In fairness, however, for the authorising commander to be held responsible there would

93 *Id.* at 36.

94 *Id.* at 37.

95 *Id.* at 48.

96 *Id.*

97 *Id.* at 2.

98 *See id.* at 1 (misperceptions as to the meaning and implications of autonomy are limiting its adoption in the military).

99 *Id.* at 8-9.

100 *Id.* at 48.

101 UK Approach, *supra* note 60, at 5-3.

102 *Id.* at 5-5.

103 *Id.* at 5-6.

need to be an underlying "assumption that a system will continue to behave in a predictable manner after commands are issued."[104] Reliance upon this assumption would become more problematic "as systems become more complex and operate for extended periods" without human intervention.[105] Fostering the level of trust in the operation of ADPs necessary for a commander to decide affirmatively to be responsible for their effects would likely require an extensive, holistic development of complementary education, realistic training, and user-friendly ADP hardware and software.[106]

E. SUMMARY

At this point in time, it does not appear to be technologically feasible, nor is it necessary, for an ADP to attempt to quantify the moral and emotional differences between responses to a potential targeting problem. The practical effect of Professor Arkin's proposal for programming ADPs is to have them unable to make the close calls, and therefore unable to engage on their own unless quantifiable criteria which have been established by erring on the side of caution have been met. The close calls, and the potential use of emotion and morality, are reserved for the human member of the team.

5. CONCLUSION

Ensuring that commanders remain responsible in the course of cyber conflict will first require significant investment in the sensors, machines, and software that are used to provide the operational visualisation of cyberspace and the geophysical world upon which the commanders would rely when making use of force decisions. This would include the ability to map the relevant portion of cyberspace to identify those points at which cyber action might reasonably be expected to ripple into the geophysical world. Commanders must be confident that the situational information and the analysis derived from these systems are accurate, and they must have the ability to access geophysical surveillance and reconnaissance assets that could provide them near real-time awareness of the likely ripple points.

Second, ADPs must be developed that quickly assess this situational information and analysis in a conservative fashion so that commanders are alerted when proposed cyber responses could be reasonably expected to cause injuries to humans or damage

[104] *Id.* at 5-5.

[105] *Id.* at 5-5.

[106] *See* Caudle, *supra* note 10, at 259, 273, 280 (new doctrine and training required to optimise cyber commander performance).

to tangible objects; or violate pre-set red-lines in terms of significant damage to data in targeted systems. Such systems might employ expanded consideration of indirect effects to reassure commanders that they are being provided a satisfactorily holistic assessment, and as a preventive measure to keep either geophysical or cyberspace thresholds from being crossed automatically or inadvertently. This would give cyber commanders confidence that their decisions to engage in cyber actions were not made in too narrow a fashion.

Third, if instances of human injury or damage to geophysical objects could be reasonably expected, then a commander at some level, rather than an ADP, would need to make the affirmative decision to engage after satisfying LOAC and ROE requirements. If no geophysical injury or damage was reasonably expected, but red-lines regarding cyber infrastructure were reasonably likely to be approached within a certain margin of uncertainty, a cyber commander at some level would also need to decide whether and how to respond within the prescribed ROE, which might themselves contain LOAC-like decision factors. If the effects of the proposed cyber action would occur and remain solely within cyberspace, then ADPs could conceivably operate without human intervention using default settings based on approved ROE. For a commander to remain confident that ADPs were behaving as expected, though, continued monitoring of the cyber and geophysical aspects of the battle-space would be necessary to ensure human override thresholds were not reached until the cyber action was completed.

Presumably, since the vast bulk of cyber action would occur within cyberspace and the effects would remain there, keeping cyber commanders responsible and compliant with LOAC should be achievable. Proposed cyber actions that could be reasonably expected to ripple into the geophysical world and cause human injury or damage to objects might likely prove to be the exception rather than the rule. Cyber actions that might violate cyberspace ROE red-lines regarding cyber infrastructure would likely be more common, but these activities would not entail potential criminal violations of international law at this point, and their effects might be both reversible and quickly terminated. For ADPs to be used properly, however, military organisations must begin investing in the education and training curricula and opportunities that that will be required to groom young cyber operators for their future roles as effective and responsible cyber commanders.[107]

[107] *Id.* at 279-80.

Chapter 5.

Cyber Conflict – Politics, Semantics, Ethics and Moral

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Divided by a Common Language: Cyber Definitions in Chinese, Russian and English

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

William Hagestad II
Red Dragon Rising
Bayport, Minnesota USA
hagestadwt@red-dragonrising.com

Abstract: During 2012, both the US and UK have signalled increased willingness to engage with Russia and China on cyber security issues. But this engagement will be extremely difficult to achieve in the absence of commonly agreed definitions, and even concepts, for what constitutes cyber security.

Russian and Chinese doctrine and writing emphasise a very different set of security challenges to those which normally concern the US and UK. There is the additional complication of direct translations of specific terms from Russian and Chinese which resemble English-language terms, and therefore give the misleading impression of mutual understanding, while in fact referring to completely different concepts.

A number of states including Russia and China, which do not subscribe to the Euroatlantic consensus on the nature and future of cyberspace, have already achieved a commonality in their views and language; while this language sometimes has no equivalent in English and is therefore imperfectly understood.

This paper examines these distinctions, comparing and contrasting terms and concepts in English, Russian and Chinese. This will illustrate the dangers involved in attempting to reach a consensus - or at the very least confidence and security building measures - with states with widely differing views on cyber security without first establishing a baseline of common definitions. Examples will show how previous attempts at doing so have been counter-productive and set back mutual understanding.

Keywords: *Russia, China, doctrine, terminology*

1. INTRODUCTION

At the end of 2012, a series of international events brought years of private dissension over the nature and future of the Internet into very public view. At the Budapest Conference on Cyberspace in October, and the World Conference on International Telecommunications (WCIT) in Dubai, a Euroatlantic consensus on an international space for free exchange of information and views clashed with an alternative model backed by Russia, China and other states, advocating national control of information space and an entirely different approach to managing content. Debates which until that point had been conducted bilaterally or through such fora as the United Nations Group of Government Experts were aired in public, leading to at times acerbic exchanges. In Budapest, on 3-5 October, European nations stressed the human rights aspects of cybersecurity, based on their understanding of internet freedom as a fundamental right (Budapest, 2012), leading an exasperated Chinese representative to ask whether he was at a conference on cybersecurity or on human rights (Samuel, 2012). And in Dubai, a proposed new set of International Telecommunication Regulations (ITR) struggled to gain the support of many of the 151 delegate nations, after strong opposition from Euroatlantic states led by a formidable US delegation (ITU, 2012).

The failure to reach agreement on fundamental principles affecting cyberspace was indicative of the fact that despite increased willingness during 2012 by the USA, UK and other nations to engage with Russia and China on cyber security issues, this engagement remains extremely difficult in the absence of commonly agreed concepts of what constitutes cyber security.

The UK's Cyber Security Strategy, issued in November 2011, states that "we will work internationally to develop international principles or 'rules of the road' for behaviour in cyberspace (UK Government, 2012) - language not dissimilar to that used by Russia and China when proposing an "International Code of Conduct" for information security (UN, 2011). But as well documented previously (Giles, Russia's Public Stance on Cyberspace Issues, 2012) (Thomas, 2001), Russian and Chinese doctrine and writing emphasise a very different set of security challenges to those which normally concern the US and UK, a disconnect which has thus far stymied progress toward mutual understanding.

Yet even before addressing divergences in attitude and threat perception, there is the more basic problem of absence of a common terminology between the major players in cyberspace. The definitions of such terms as cyber conflict, cyber war, cyber attack, cyber weapon, etc. used by the UK, USA, Russia and China do not coincide - even where official or generally recognised definitions exist in each respective language. Furthermore, direct translations of specific terms from Russian and

Chinese which resemble English-language terms, and vice versa, can complicate matters further by giving the misleading impression of mutual understanding, while in fact referring to completely different concepts.

This paper will seek to illustrate fundamental incompatibility between terms and concepts subscribed to in these four countries, by examining a number of Russian and Chinese concepts and by including reference to and comparison with US and UK policy statements. The intention is to point to the dangers involved in attempting to reach a consensus - or at the very least confidence and security building measures - between states with widely differing views on cyber security without first establishing a baseline of common definitions, and show how at least one previous attempt at doing so has been counter-productive and set back mutual understanding.

2. A DIFFERENT VIEW

The existence of this fundamental disconnect between the Euroatlantic view of information security and the Russia and Chinese approaches has long been recognised among the expert communities dealing with both countries. In the Russian case, one main distinction is the holistic approach to information security, as opposed to a siloed focus on cyber issues. As pointed out by Tim Thomas in a 2001 comparison of Russian and US information security definitions from official sources, "Thus, differently than the U.S., Russia views both the mind and information systems as integral parts of its concept of information security." (Thomas, 2001)

More recently, this consciousness has spread beyond subject matter experts to be generally accepted by policy-makers - including public recognition by senior UK figures that the lexicon of foreign counterparts is based on a fundamentally different conceptual approach to the nature of information, and thus of information security (GSF, 2012).

3. FINDING COMMON GROUND

Initiatives seeking harmonisation between Russian and English terminology appear mainly to come from the Russian side, at least in public. At a 2007 NATO-Russia workshop aimed at developing a common vocabulary to deal with information security issues, leading security official Anatoliy Streltsov stated that Russia hopes for the "development of [a] multilingual conceptual framework that will allow both politicians and specialists working in the field[s] of legislation, law enforcement and prosecution, to have a common approach to legal regulation." (Streltsov, 2007) Yet the stated objective of this harmonisation may serve as a deterrent in some cases:

the same speaker continued that:

> "The creation of such [a] conceptual framework will contribute to forming necessary conditions for harmonizing national legislations and for developing international agreements aimed to regulate relations in the field of providing information security of a single state and [of the] international community as a whole."

- language which could have been calculated to trigger neuralgia among those states who do not subscribe to the notion of national information space, or international treaties regulating information security.

An initiative by the EastWest Institute, confusingly labelled a "Russia-US Bilateral", sought to break this deadlock by introducing "a joint effort between American and Russian experts to seek consensus definitions around three key cluster areas of cybersecurity terminology". (EWI, 2011)

This laudable effort appeared at first sight to make ground-breaking progress in establishing a baseline of common understanding. Regrettably, this progress proved illusory, since the agreed definitions in each language did not actually match up with each other, leaving each side under the impression that consensus had been achieved but in fact remaining as far apart as ever.

For example, the English-language definition of "Cyber Warfare" reads:

> *Cyber Warfare is cyber attacks that are authorized by state actors against cyber infrastructure in conjunction with a government campaign*.

Whereas the Russian version below it reads:

> *Combat actions in cyberspace are cyber attacks carried out by states, groups of states, or organised political groups, against cyber infrastructure, which are part of a military campaign*.
>
> (*Боевые действия в киберпространстве - кибератаки, проводимые государствами (группами государств, организованными политическими группами), против киберинфраструктур, и являющиеся частью военной кампании*.) (EWI, 2011)

The differences between the supposedly harmonised definitions, and their implications, are clear enough to require little further elaboration. The difference between a "government campaign" and a "military campaign" when defining warfare is problematic enough; but the mention of "organised political groups", present in one language but absent in the other, would cause serious difficulties if an attempt were made to apply it to determining whether the online activities of

Russian state-sponsored groups such as *Nashi* in fact constituted undeclared "cyber warfare".

Nevertheless, the task of finding common ground between Russian and US experts on this topic should not be underestimated. EastWest Institute's attempt in Brussels in November 2011 to follow up the initial 20 terms with a further range of agreed definitions stalled on the inability to reach a common understanding of the fundamental term "information".

4. SOURCING

If seeking to compare and baseline terminology between languages, the question arises of where precisely to seek the "official" definitions espoused by each nation. Russia's Information Security Doctrine was issued in 2000 but is still the key public document governing official information (including cyber) policy. The doctrine lists threats and challenges but avoids precise technical definitions of the key terms used (Russian Government, 2000). In this, the document is not unique to Russia: the UK Cyber Security Strategy 2011, referenced above, does precisely the same. So in some cases a direct comparison of the interpretation of key terms from foundational documents is not possible, and inferences have to be drawn from usage and second-line documentation. In fact, in the absence of officially and publicly approved definitions, allowance must be made for usage of terms remaining in flux even within individual nations - one of the immediately noticeable changes between the initial 2009 version of the UK Cyber Security Strategy and the most recent version at the time of writing, issued in November 2011, was a graduation from the phrase "cyber space" to the word "cyberspace". This, while hardly a noteworthy change in itself, was indicative of the fact that even the most basic terms have yet to evolve into a settled and universally accepted vocabulary even in individual countries.

Fortunately, there is no shortage of official pronouncements and documentation from which to derive interpretations of key terms, as well as to establish that in addition to the difficulties of mismatched interpretations, Russia, China and the Anglosphere use a number of terms which denote important information security concepts in the home language, but which simply have no easily comprehensible equivalent when translated.

In the case of Russia, it should be possible to source and interpret many of these terms from those Russian documents which are intended for international consumption, given the persistent efforts over a number of years to promote the Russian view of information security to the world and gather supporters. One source of definitions which can be treated as representing the official view is the "Draft Convention on

International Information Security", which outlines Russia's desired end state for international agreement on governance of cyberspace as a subset of information security overall (Russian Government, 2011). This document has already been analysed in detail in a joint Russian-British commentary, which noted linguistic complications in its interpretation (CSRC, 2012). The case studies below examining specific points of lexical contention will further compare individual terms from the Russian document with their Chinese and English-language equivalents, where these exist.

5. CASE STUDIES – SPECIFIC TERMS

The table below gives the English, Chinese and Russian renderings of common information security terms. Yet as can be seen from the detailed examination of each term that follows, these literal translations are potentially misleading, since the concepts and assumptions that lie behind them vary so widely.

Table I. Key Cyber Security Terms

English	Chinese	Russian
information space	信 息空間 xìnxī kōngjiān	информационное пространство *informatsionnoye prostranstvo*
information warfare	信 息战争 xìnxī zhànzhēng	информационная война *informatsionnaya voyna*
information weapon	信息武器 xìnxī wǔqì	информационное оружие *informatsionnoye oruzhiye*
information security	信息安全 xìnxī ānquán	информационная безопасность *informatsionnaya bezopasnost*
cyber warfare	網絡戰爭 wǎngluò zhànzhēng	кибервойна *kibervoyna*
cyberspace	網絡空間 wǎngluò kōngjiān	киберпространство *kiberprostranstvo*
cyber security	網絡安全 wǎngluò ānquán	кибербезопасность *kiberbezopasnost*
network warfare	網絡戰 wǎngluò zhàn	сетевая война *setevaya voyna*

A. *"INFORMATION SPACE"*

Both Russia and China refer to "information space", a concept which is much less well established in the Anglosphere. In Russia's Draft Convention, "information

space" (информационное пространство, *informatsionnoye prostranstvo*) is defined as "the sphere of activity connected with the formation, creation, conversion, transfer, use, and storage of information and which has an effect on individual and social consciousness, the information infrastructure, and information itself" – although subsequent usage within the Convention shows that this definition itself is subject to flux. In Chinese, the equivalent phrase is 信息空間, rendered in PinYin as "Xìnxī kōngjiān". The Chinese definition of this phrase includes the following: "The main function of the information space for people to acquire and process data... a new place to communicate with people and activities, it is the integration of all the world's communications networks, databases and information, forming a "landscape" huge, interconnected, with different ethnic and racial characteristics of the interaction, which is a three-dimensional space." (Wasuo, 2000)

Thus the Chinese view "information space" as a domain, or landscape, for communicating with all of the world's population. This chimes with the Russian view of this space including human information processing, in effect cognitive space. This factor is key to understanding the holistic Russian and Chinese approaches to information security as distinct from pure cybersecurity, a fundamental difference from the Euroatlantic approach to the subject. As expressed by Timothy Thomas, "differently than the U.S., Russia views both the mind and information systems as integral parts of its concept of information security... China appears more like Russia than the U.S. in its understanding of information security, with its emphasis on the mental aspect of information security and its extended use of the term itself." (Thomas, 2001)

B. "CYBERSPACE"

By contrast, Russian and Chinese official references to "cyberspace" occur primarily in translations of foreign texts or references to foreign approaches. According to a US military definition, "Cyberspace...is the Domain characterized by the use of electronics and the electromagnetic spectrum to store, modify, and exchange data via networked systems and associated physical infrastructures"; and consequently, "Cyberspace Operations [is the] employment of cyber capabilities where the primary purpose is to achieve objectives in or through cyberspace. Such operations include computer network operations and activities to operate and defend the Global Information Grid." (US DoD, 2010) But the Russian rendering киберпространство, *kiberprostranstvo*, and the Chinese 網絡空間, Wǎngluò kōngjiān, are merely subsets of "information space" and inseparable from it, unlike in Western treatment where "cyberspace" continues in some writing to be treated almost as a separate domain. Meanwhile, the natural Chinese term which comes closest to what English-language readers might understand as "cyberspace" is 虛擬

主機 , Xūnǐ zhǔjī, which could simply be translated as virtual host – no more than the necessary components for connecting a machine to a network for the specific purposes of communicating via protocols such as HTML, email and so on.

C. *"CYBER WARFARE"*

A similar pattern pertains with the phrase "cyber warfare". Unsurprisingly, this phrase is well defined in US terminology. The Joint US Military definition for "cyber warfare" is "an armed conflict conducted in whole or part by cyber means. Military operations conducted to deny an opposing force the effective use of cyberspace systems and weapons in a conflict. It includes cyber attack, cyber defense, and cyber enabling actions." (US DoD, 2010) The US definition is further elaborated with defensive and offensive capabilities in the cyber warfighting domain[1] – a distinction from other areas of the information space which has yet to find expression in public Russian writing on the subject, for example.

The difficulties encountered by EastWest Institute in attempting to harmonise Russian and English definitions of "cyber warfare" have been described above. In part this derives from the fact that in Russia and China, similarly to "cyberspace", the phrase "cyber warfare" is used primarily to denote potential US and allied activity (Giles, 'Information Troops' – a Russian Cyber Command?, 2011). Russia's Draft Convention does not make any reference at all to "cyber warfare". Meanwhile China's People's Liberation Army (PLA) uses the term 網絡戰, Wǎngluò zhàn, as a necessary vocabulary item to render "cyber warfare" specifically for understanding the way the Western world defines conflict in this new domain. Operations in the cyber realm are further defined as 網絡作戰, Wǎngluò zuòzhàn, network warfare

[1] The Joint Terminology for Cyberspace Operations; 2010-11 defines Defensive Counter-Cyber (DCC) and Offensive Counter-Cyber (OCC) operations.
Defensive Counter-Cyber (DCC) are "All defensive countermeasures designed to detect, identify, intercept, and destroy or negate harmful activities attempting to penetrate or attack through cyberspace. DCC missions are designed to preserve friendly network integrity, availability, and security, and protect friendly cyber capabilities from attack, intrusion, or other malicious activity by pro-actively seeking, intercepting, and neutralizing adversarial cyber means which present such threats. DCC operations may include: military deception via honeypots and other operations; actions to adversely affect adversary and/ or intermediary systems engaged in a hostile act/ imminent hostile act; and redirection, deactivation, or removal of malware engaged in a hostile act/ imminent hostile act." Offensive Counter-Cyber (OCC) are "Offensive operations to destroy, disrupt, or neutralize adversary cyberspace capabilities both before and after their use against friendly forces, but as close to their source as possible. The goal of OCA operations is to prevent the employment of adversary cyberspace capabilities prior to employment. This could mean preemptive action against an adversary." The Joint U.S. Military definition of Offensive Cyberspace Operations (OCO) is "Activities that, through the use of cyberspace, actively gather information from computers, information systems, or networks, or manipulate, disrupt, deny, degrade, or destroy targeted computers, information systems, or networks. This definition includes Cyber Operational Preparation of the Environment (C-OPE), Offensive Counter-Cyber (OCC), cyber attack, and related electronic attack and space control negation."

operations, and offensively, 網絡戰攻擊, Wǎngluò zhàn gōngjí, cyber warfare attacks (Zaiyao, 2006).

In all cases, as in the case of "cyberspace" described above, the phrase "cyber warfare" in Russian and Chinese writing describes foreign concepts and activities – denoting the foreign notion that information conflict could be restricted to the cyber domain as opposed to encompassing other areas of the "information space".

D. "INFORMATION WEAPON"

"Information weapon" is another phrase which is not in common usage in the Anglosphere, but used as a current term in Russian discourse – as, for example, in a presentation by Anatoliy Streltsov to the International Information Security Research Consortium on 2 October 2012 detailing Russian proposals for confidence building measures in cyberspace, specifically:

> "The adoption [of] international legal instruments, emerging norms of international humanitarian law, international security law and law of war as they apply to the use of the 'information weapon' in interstate conflicts"

In keeping with the broader Russian understanding of "information space", the term "information weapon" has an impressively broad application. The definition given in Russia's Draft Convention – "information technology, means, and methods intended for use in information warfare" – is in fact misleading, since it appears close to the English-language concept of a cyber weapon, whereas in fact usage both in this document and elsewhere makes it very clear that "information weapons" can be used in many more domains than cyber, crucially including the human cognitive domain. For instance, only one of the three following examples maps to the concept of a cyber weapon:

> "Propaganda carried out using the mass media is the most traditional and most powerful general-purpose information weapon... Information weapons are being actively developed at the present time based on programming code... Information weapons also include means that implement technologies of zombification and psycholinguistic programming." (Fedorov & Tsigichko, 2001)

E. "INFORMATION WARFARE"

In common with "information weapons", it is crucial to understand that "information warfare" itself in Russian and Chinese usage carries meaning which is specific, broad, holistic, and not rendered by the direct translation into English.

Western definitions of "information warfare" are varied but broadly speaking semantically equivalent. One uncontroversial definition dating from the 1990s reads:

> "Information warfare is the offensive and defensive use of information and information systems to deny, exploit, corrupt, or destroy, an adversary's information, information-based processes, information systems, and computer-based networks while protecting one's own. Such actions are designed to achieve advantages over military or business adversaries." (Arquilla & Ronfeldt, 1993)

However, more recently English-language military terms and concepts for cyberspace operations have almost eclipsed mentions of "information warfare" as a whole, whose components have to be sought under the separate headings of disciplines such as psychological operations, Influence operations, strategic communications and more.

Meanwhile, Russian and Chinese writing on the subject has more explicitly retained the more holistic and integrated view of information warfare as a distinct, but unified and complete discipline – as pithily described by Sergey Rastorguyev, a Russian writer on information theory and information warfare with a useful line in animal metaphors:

> "...the tortoise never understood, and now never will, that information war is the deliberate teaching of your enemy how to remove his own shell." (Rastorguyev, 2006)

This conceptual gap has been well documented elsewhere, and is well recognised among US and UK practitioners. It is important also to recognize that discussion of the subject among Chinese and Russian military academics has a particularly long and well-established history. The basis for Chinese information warfare doctrine is derived from earlier Chinese military doctrine up to and including Sun Tzu's "Art of War" and Sun Ping's "Military Methods" in the 6th and 4th centuries BC respectively. Modern Chinese military cyber strategists use these ancient military annals as a guiding tenet for modern-day cyber and information warfare military strategy.

The evolution of Chinese information warfare in the digital age begins notably with People's Liberation Army General (PLA) Major General Wang PuFeng in 1995. General Wang is considered by many in the Western world to be the founding father of Chinese information warfare theory. At the same time, PLA Senior Colonels Wang Baocun and Li Fei of the Academy of Military Science, Beijing, were examining and studying the United States military tenets of information warfare,

including the current writings on the digitized battlefield and informatisation of the military (Baocun & Fei, 1995). The eventual result was the decision by the Central Military Commission in late 2003 on building computerized armed forces and winning the new strategic goal of information warfare (Zhuangzhi, 2012).

The early and mid 1990s also saw Russian recognition that existing concepts of information warfare needed to adjust to new digital realities. As noted by information warfare theorist Vitaliy Tsygichko and others in 1995, "the development of a [US] national, and then an international, information superhighway" would "create new conditions for the effective employment of information weapons" and furthermore that "the prototype of this superhighway already exists. That is the Internet, a worldwide association of computer networks".

Tsygichko went on to warn that:

> Although we live in an era of global information systems and we understand that economic vegetation awaits the country if it is not connected to the world information space, we must precisely imagine that Russia's participation in international telecommunications and information exchange systems is impossible without the comprehensive resolution of the problems of information security. (Smolyan, Tsygichko, & Chereshkin, 1995)

6. CHINESE AND US INFORMATION SECURITY POLICY

This last quotation reminds us that the differences in definitions and understandings of key information security and cyber warfare terms between Russian, Chinese and English are more than an academic problem presenting a stimulating translation challenge. Since they form the underpinning for entire national approaches to the subject by major players in the cyber domain, it is important to understand how they affect policy and how conceptual differences extend into distinct policy approaches. The Russian approach to information security has been described, and contrasted with the Euroatlantic view, in previous work (Giles, Russia's Public Stance on Cyberspace Issues, 2012). The following section will describe and contrast Chinese and US information security policy, in order further to illustrate the conceptual gap and consequent challenges for mutual understanding.

In 2012 the State Council of Central People's Government of the People's Republic of China mandated that the security and protection of information technology would be a national Chinese priority (Gu Fa, 2012). The State Council's information security mandate states that the Council will "vigorously promote" development of various forms of information technology while ensuring the protection and importance of information security.

The importance assigned to information security in the official view from the State Council is not that dissimilar to the situation in the United States. At the same time, the incongruence between the United States cyber security order issued by the White House (available, at the time of writing, in draft form) and that of the Chinese is actually startling when compared directly. The US Executive Order on cyber security directs all US federal entities to develop their own guidelines for cyber security to protect national critical infrastructures (US Government, 2012).

Meanwhile, China's State Council mandate reflects an overarching concern for *all* information technologies, suppliers and infrastructures both civilian and governmental, including the People's Liberation Army. The State Council proclaimed that the country will "Improve the security and management, information security and protection of key areas…" through a series of specific improvement programmes (Gu Fa, 2012).

The first mandate of improvements includes a focus on all critical information systems and infrastructure with particular attention being paid to the security of information networks. Thus in this way the State Council is giving very specific official intent, rather than guidance, to Chinese civilian and government leaders regarding what they must protect and the importance of the role this plays in the overall State Council plan. Further classifications and definitions are detailed within the critical information systems ecosystem, including but not limited to national and private telecommunications systems, radio networks, and the internet. From this overarching taxonomy the State Council further delineates required areas to be secured, including basic information networks such as energy, transportation, financial and other related industries where a cyber attack would cause a detrimental effect to the People's Republic of China's civilian economy.

The US cyber security order offers no distinction between wholesale protection of conjoined US Federal and commercial infrastructure. Indeed, businesses within the United States must rely heavily on a self-educated information security profession to protect themselves from the vagaries of attacks delivered by or through cyber means. Conversely, the People's Republic of China dictates and assigns responsibility to all levels of both governmental and commercial entities to share the duty of protecting a holistic realm of national critical infrastructure.

The Chinese State Council continues to demonstrate a national sense of ownership by providing amplifying instructions as their commander's intent for securing national information systems. The Council specifies distinct actions to be taken going so far as to personify these actions by using the pronoun 您, (Nín) which is the Mandarin formal word for "you", thus rendering them a direct order. The actions the state and commercial leaders within China are to take include, but remain not limited to, information security planning which must be coordinated and synchronised. Within the synchronised operation of security facilities, "you"

must strengthen against and prevent the impact and negative effect of cyber-attacks. Information security management must include continued implementation and improvement of information security measures such as cyber-attack defeat systems, including countering attacks from the web, hardware, and software. Increased resilience of "anti-attack", tamper-proof, "anti-virus, anti-paralysis and anti-theft capabilities" is also specified.

The second definitive State Council action mandates the strengthening of governmental and classified information security systems. This particular statement also includes further amplification regarding the use of cloud based information systems, data centre facilities, and the prohibition of unauthorised software installation. The State Council further expects the establishment of a government website set up to perform audits, monitor and report. Chinese Government agencies will reduce the number of points at which they are connected to the internet, and strengthen information security and confidentiality protection monitoring, as well as implementing "a hierarchical system of protection of classified information systems, strengthening also the review mechanism of classified information systems." (Gu Fa, 2012)

The third element of State Council combined and coordinated information security guidance addresses the protection and security of industrial control systems (ICS). ICS security and protection must be achieved and maintained at Chinese facilities involved in the nuclear, aerospace, advanced manufacturing, petroleum and petrochemical, oil and gas pipelines, power systems, transportation, and water conservancy industries, urban facilities and what the State Council refers to explicitly as "the Internet of Things applications." The State Council also mandates a digital city construction safety and management policy, including regular safety checks, security audits and risk assessments. Regulation is to be strengthened, especially on those ICS that may endanger the safety of life and public property (Gu Fa, 2012).

The fourth information security mandate concerns the safeguarding of Chinese citizens' personal information, stating that "the protection of personal information is a necessary condition for the overall welfare of the People's Republic of China in the Information Age. Geographic, demographic, legal, statistical and other basic information resources will be afforded the utmost in digital protection and management. Similarly the protection of sharing information resources and the interoperability of security information systems is paramount." Clear sensitive information protection requirements are to include the strict regulation of all Chinese businesses, institutions, in order to "protect user data and national basic data throughout the entire information network of economic activity in the People's Republic of China". In relative terms, United States federal policy on cyber security is not prescriptive on protection of personal information, simply mandating that

commercial enterprises which fail to follow basic guidelines for the protection of personal information will be penalised monetarily.

In summary the People's Republic of China takes a proactive and holistic approach, directed from above, to protecting its overall national information security including both Chinese commercial enterprises and governmental entities. In contrast, the United States by and large gives direction only to federal entities to ensure awareness on what is vulnerable to cyber-attacks. Commercial organisations in the United States are not issued prescriptive instructions on ensuring their own protection, and are subjected to relatively light touch regulation in this field, trusted to protect their own digital interests.

Besides reflecting the approaches of the respective governments to centralised versus decentralised command, this distinction in approach between the two states provides a clear illustration of another disconnect between concepts of the internet which hinders international understanding: the Euroatlantic view of a free and open space, effectively self-governed by a broad range of stakeholders, as opposed to the state-centric view espoused by Russia, China and like-minded nations where it is the national government which carries responsibility for the domestic "information space". (CSRC, 2012)

7. CONCLUSION - OPTIONS FOR PROGRESS

The "UK non-Paper on Global Cyber-Security Capacity Building" presented at the Budapest Conference on Cyberspace noted that "it is crucial to develop the capacity and trust to cooperate internationally", and included in its list of "key dynamics" and "potential ways forward" a note that:

> "our response is often limited by the legal and political boundaries of our states or the boundaries of commercial interests. In many cases, our state- or organisation-based response is insufficient to counter the threat: **effective response depends on working collectively**." [emphasis in original] (UK Government, 2012)

Yet it included no indication or suggestion of any path towards achieving a meaningful dialogue with those international actors who do not share the UK vision of cyberspace. This is an indication that although broad dialogue continues bilaterally as well as in fora like the Organisation for Security and Cooperation in Europe (OSCE) and the UN, agreement or even mutual understanding are still distant. As noted in the introduction to this paper, public exposition of the two opposing views has only properly begun in the past year. Thus, the fundamental difference between these two views is only now achieving broader recognition.

Bilateral dialogue is particularly challenging in the case of the United States and

China. Attempts at engagement by the U.S. on cyber issues have been hampered by the need to address persistent reporting, including by commercial information security firms in the U.S., that hostile activity including cyber espionage and hacking is conducted by PLA units. China calls these claims "false", "unlawful" and "without merit" (Bloomberg, 2013). Official statements by the Communist Party of China (CPC) in conjunction with the PLA claim that first, the CPC does not condone hacking of any kind, and in fact has cracked down on unlawful criminal usage of computers since the creation and implementation of its municipal cyber police (China Police); and that second, there are no information warfare units active in the PLA (Japan Times, 2013) (LeClaire, 2013). Meanwhile, any action proposed by the U.S. against China is portrayed by China as further evidence that the US is seeking global escalation of information warfare (Jian, 2013).

In some cases, venues which might provide an opportunity for engagement between the opposing views do not succeed in doing so. In advance of the Budapest conference, lists of "International Cyber Documents" and "National Cyber Strategies" were provided for reference by delegates and the public on the conference website. Significantly, these lists of official national and multilateral statements only included those documents which subscribed to the Euroatlantic view of cyberspace: documents published or endorsed by Russia, China and like-minded states were either omitted entirely, or simply not submitted by those states in the first place (Budapest, 2012). Subsequently, according to delegates, much of the Russian address to the plenary session did not survive the interpreting process and therefore was lost on non-Russian speaking attendees - just as at the preceding London Conference on Cyberspace in November 2011, where the illusion of consensus was created by key caveats being omitted from the translation of the speech by Communications Minister Igor Shchegolev.

Despite the impression created by some public statements from the US and UK, Russia and China are not isolated in their view of information security: there are a large number of other states which share their views, and their concerns over hostile content as well as hostile code. At present, given the congruence between Russian and Chinese approaches and concepts, terminology and policy, it is far easier for Russia, China and like-minded nations to find common ground than it is for English-speaking nations to engage constructively with them. Those holding an optimistic view on the prospects for relations between Russia, China and the West would argue that this process of engagement has the potential to provide opportunities for productive dialogue on other topics - especially in an environment where some representatives of Russia and the USA in particular have repeatedly voiced the desire to find any possible areas for cooperation. What is certain is that in the absence even of a mutually comprehensible lexicon for describing the concepts within information security, any potential for finding a real commonality of views on the nature and governance of cyberspace remains distant.

REFERENCES

Arquilla, J., & Ronfeldt, D. (1993). Cyberwar is Coming! *Comparative Strategy , 12* (2), 141-160.

Baocun, W., & Fei, L. (1995). *Information Warfare*. Retrieved from Federation of American Scientists: http://www.fas.org/irp/world/china/docs/iw_wang.htm

Bloomberg. (2013, March 17). *Li Rejects U.S. Hacking Allegations Against China as Groundless*. Retrieved from Bloomberg: http://www.bloomberg.com/news/2013-03-17/li-rejects-u-s-hacking-allegations-against-china-as-groundless.html

Budapest. (2012). Budapest Conference on Cyberspace. Budapest.

Budapest. (2012). *International Cyber Documents*. Retrieved from Budapest Conference on Cyberspace: http://www.cyberbudapest2012.hu/international-cyber-documents

China Police. (n.d.). *Public Information Network Security Supervision*. Retrieved from Ministry of Public Security of the People's Republic of China: http://www.mps.gov.cn/English/menu_1_4_1.htm

CSRC. (2012, April). *Russia's 'Draft Convention on International Information Security' - A Commentary*. Retrieved from Conflict Studies Research Centre: http://conflictstudies.org.uk/files/20120426_CSRC_IISI_Commentary.pdf

EWI. (2011, April). *Russia-U.S. Bilateral on Cybersecurity – Critical Terminology Foundations*. Retrieved from EastWest Institute: http://www.ewi.info/cybersecurity-terminology-foundations

Fedorov, & Tsigichko. (2001). *Information Challenges to National and International Security*. Moscow: PIR Centre.

Giles, K. (2011). 'Information Troops' – a Russian Cyber Command? Tallinn: CCDCOE.

Giles, K. (2012). Russia's Public Stance on Cyberspace Issues. Tallinn: CCDCOE.

GSF. (2012, November 21). Cyber Security: Meeting The Challenges, Combating The Threats? *Global Strategy Forum seminar*. London.

Gu Fa. (2012). *State Council vigorously promotes the development of information technology and to effectively protect the information security*. Retrieved from http://www.gov.cn/zwgk/2012-07/17/content_2184979.htm

ITU. (2012, December 14). *New global telecoms treaty agreed in Dubai*. Retrieved from http://www.itu.int/net/pressoffice/press_releases/2012/92.aspx#.UOse-G9WySo

Japan Times. (2013, March 10). *China foreign minister denies hacking claims*. Retrieved from Japan Times: http://www.japantimes.co.jp/news/2013/03/10/asia-pacific/china-foreign-minister-denies-hacking-claims

Jian, Y. (2013, March 8). *Yang Jian: the United is promoting global "network arms race"*. Retrieved from People's Daily: http://military.people.com.cn/n/2013/0308/c1011-20718565.html

LeClaire, J. (2013, February 20). *China Denies Its Army Is Behind Hack Attacks*. Retrieved from Newsfactor: http://www.newsfactor.com/story.xhtml?story_id=111003TU8EJ9

Rastorguyev, S. (2006). *Information War. Problems and Models*. Moscow.

Russian Government. (2011, October 28). *Draft Convention on International Information Security*. Retrieved from Embassy of Russia to the UK: http://rusemb.org.uk/policycontact/52

Russian Government. (2000, September 9). *Information Security Doctrine of the Russian Federation*. Retrieved from Russian Ministry of Foreign Affairs: http://www.mid.ru/ns-osndoc.nsf/1e5f0de28fe77fdcc32575d900298676/2deaa9ee15ddd24bc32575d9002c442b?OpenDocument

Samuel, C. (2012, October 11). *Some takeaways from the Budapest Conference on Cyberspace*. Retrieved from Institute for Defence Studies and Analyses: http://www.idsa.in/idsacomments/SometakeawaysfromtheBudapestConferenceonCyberspace_csamuel_111012

Smolyan, G., Tsygichko, V., & Chereshkin, D. (1995, November 18). A Weapon That May Be More Dangerous Than a Nuclear Weapon: The Realities of Information Warfare. *Nezavisimoye voyennoye obozreniye* .

Streltsov, A. A. (2007). Legal Groundwork for Information Security and Conceptual Framework. In J. van Knop, *A Process for Developing a Common Vocabulary in the Information Security Area*. IOS Press.

T'ung, M. T. (1967). On Protracted War. In *Selected Works of Mao Tse-tung* (pp. 113-114). Peking: Foreign Languages Press.

Thomas, T. L. (2001, July). *Information Security Thinking: A Comparison Of U.S., Russian, And Chinese Concepts*. Retrieved from Foreign Military Studies Office: http://fmso.leavenworth.army.mil/documents/infosecu.htm

UK Government. (2012). *The UK Cyber Security Strategy - Protecting and promoting the UK in a digital world*. Retrieved from http://www.cabinetoffice.gov.uk/sites/default/files/resources/uk-cyber-security-strategy-final.pdf

UK Government. (2012). *UK non-Paper on Global Cyber-Security Capacity Building for the Budapest Conference on Cyberspace*. Retrieved from Budapest Conference on Cyberspace: http://www.cyberbudapest2012.hu/download/f/31/00000/UK_NON_PAPER_ON_CAPACITY_BUILDING%20-%20BUDAPEST_CONFERENCE.pdf

UN. (2011). International code of conduct for information security. *Annex to the letter dated 12 September 2011 from the Permanent Representatives of China, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General*.

US DoD. (2010). *Joint Terminology for Cyberspace Operations*. US Department of Defense. Memorandum For Chiefs Of The Military Services Commanders Of The Combatant Commands Directors Of The Joint Staff Directorates.

US Government. (2012, September 28). *White House Cyber-Security Order*. Retrieved from http://www.securitydefenceagenda.org/Contentnavigation/Library/Libraryoverview/tabid/1299/articleType/ArticleView/articleId/3251/categoryId/64/White-House-cybersecurity-order.aspx

Wasuo, H. B. (2000). *Information Space*. Shanghai: Translation Publishing House.

Zaiyao, H. B. (2006). Honeypot technology architecture network warfare training virtual shooting range environment. Construction of a virtual target circumstances for cyber war training by honeypot technology. *Journal Of Huazhong University Of Science And Technology (Nature Science)* .

Zhuangzhi, X. (2012, July 29). *10 years, China's national defense and army building to achieve a historic leap*. Retrieved from Xinhua News Agency: http://www.gov.cn/jrzg/2012-07/29/content_2194324.htm

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)

Towards a Cyber Conflict Taxonomy

Scott D. Applegate
Center for Secure Information Systems
George Mason University
Fairfax, Virginia
sapplega@gmu.edu

Angelos Stavrou
Center for Secure Information Systems
George Mason University
Fairfax, Virginia
astavrou@gmu.edu

Abstract: This paper seeks to create a practical taxonomy to describe cyber conflict events and the actors involved in them in a manner that is useful to security practitioners and researchers working in the domain of cyber operations. The proposed Cyber Conflict Taxonomy is an extensible network taxonomy organized as a plex data structure. Subjects of the taxonomy are entered as either Events or Entities and are then categorized using the categories and subcategories of Actions or Actors. Each of these categories is further subdivided into increasingly specific subcategories used to describe the defining characteristics of each subject and labeled lateral linkages are used to illustrate the associative relationships between Entities and Events. The categories are organized in both a hierarchical and associative manner to illustrate the relationships between subjects and categories. A prototype of this taxonomy was developed and tested using a test set of recent cyber conflict events and used to explore the relationship and connections between these events and the states, groups or individuals that participated in them. Furthermore, this taxonomy can potentially identify actors across different events based on their similar method of operation, toolsets and target sets.

Keywords: *Cyber Conflict, Cyber Operations, Taxonomy*

1. INTRODUCTION

This paper seeks to construct a practical and comprehensive taxonomy to describe cyber conflict events and the actors involved in them in a manner that is useful to security practitioners and researchers working in the domain of cyber operations. Our aim is to provide an organized formal model that can be used to measure the impact of attacks and different defense strategies both in specific scenarios and in large-scale cyber conflicts. To study a subject effectively, one must have some means of organizing the knowledge related to that subject. A taxonomy provides a logical organizational framework for doing this and can act as a tool to assist users in visualizing relationships and classifying data in a useful manner. The military strategist Carl von Clausewitz discussed the importance of the "coup d'oeil" which he roughly described the ability for a military leader to be able to see and immediately grasp the implications of a military situation with one "cast of the eye" [1]. With this in mind, this project attempts to create a Cyber Conflict Taxonomy that will give the security practitioner a coup d'oeil of cyber conflict related events.

The use of the term Cyber Conflict Taxonomy versus a Cyber Warfare Taxonomy in this project seeks to recognize the fact that other entities beyond states, such as non-state actors, hacktivists groups and even private individuals, are playing a role in the ongoing hostile, politically motivated actions that are taking place in cyberspace. It is therefore important that a taxonomy designed to describe these events and actors take that fact into account, hence, the proposed taxonomy will attempt to describe not just events that take place solely between nation-states, but also events undertaken by non-state entities directed at other competitor states for political, nationalistic or ideological purposes.

To further this effort, a review of previously developed taxonomies was undertaken to give the paper a logical starting point and to determine what previous works were relevant to this work. To date, no one has undertaken a taxonomy specifically geared towards classifying and understanding cyber conflict, but numerous taxonomies have been created that address cyber threats and other aspects of cyber security.

2. SURVEY OF PREVIOUS RELEVANT TAXONOMIES

A great deal of previous work has been done in the area of classifying threats and vulnerabilities. Early taxonomies such as Bishop's 1995 work focused on categorizing security vulnerabilities in software to assist security practitioners in maintaining more secure systems through an understanding of these vulnerabilities [2]. John Howard extended this idea in his 1997 work in which he analysed and

classified 4299 security related incidents on the internet. Howard's work was notable because he included attackers, results and objectives as classification categories expanding threat taxonomies beyond the technical details of an attack to include more intangible factors such as an attacker's motivation for conducting an attack [3]. Hansman and Hunt created a unique taxonomy in 2004 which was designed to be used by information bodies to classify new attacks. This taxonomy was based on four dimensions but was also designed to be extensible in that additional dimensions, some of which the authors suggested, could be added to the taxonomy as needed [4].

The vast majority of threat taxonomies are designed as attacker-centric frameworks which categorize attacks from the perspective of an attacker's tools, motivations and objectives. Killouri, Maxion and Tan created a taxonomy in 2004 designed to be defense-centric based on how an attack manifested itself in the target systems. Based on a test set of 25 attacks, this taxonomy was able to predict whether or not the defenders detection systems would be able to detect a given type of an attack [5]. In a similar effort, Mirkovic and Reihner created a taxonomy of Distributed Denial of Service (DDoS) Defenses which categorized DDoS defense mechanisms based on activity level, degree of cooperation and deployment location [6]. These two taxonomies are among the few that classify threats or security incidents from a defensive viewpoint and show the importance of addressing such issues from different perspectives to gain a more holistic view of security issues.

Another approach towards classifying cyber-attacks is to look at the actors involved versus the actual attacks. Kjaerland's 2005 study categorized cyber intrusions based on four categories; (1) method of operations, (2) impact of the intrusion, (3) source of the intrusion and, (4) target [7]. This study examined the likelihood of attacks against different kinds of targets and the likelihood of various kinds of attacks occurring together on a given target. It proved very valuable to this project in that it examined relationships between targets and the impact of attacks on those targets. In 2005, Rogers was one of a number of researchers who attempted to classify the actual attackers themselves. The Rogers' study modeled its taxonomy using a modified circular order circumplex which classified eight levels of hackers across two principal dimensions of skill and motivation [8].

Researchers at the University of Memphis created a cyber-attack taxonomy called AVOIDIT in 2009 which described attacks using five, extensible classifications: Attack Vector, Operational Impact, Defense, Informational Impact, and Target [9]. This taxonomy was created as a network plex taxonomy which, unlike previous efforts, allowed the classification of blended attacks. Additionally, it also allowed for the classification of attacks by both operational and informational impacts and was designed to help educate defenders by looking at attacks' various impacts,

vectors or target types. While this taxonomy focused exclusively on cyber-attacks, its structure and style were very useful in designing the proposed taxonomy in this paper, especially the ability to view and categorize attacks from different taxonomic perspectives.

In recent years, a number of researchers have begun to look at creating taxonomies specifically addressing SCADA systems. In 2010 Fovino, Coletta & Masera created a comprehensive taxonomy describing SCADA architecture, vulnerabilities, attacks and countermeasures [10]. In 2011 Zhu, Joseph, & Sastry highlighted the difference between what they termed standard information technology (IT) systems versus SCADA systems and focused on systematically identifying and classifying attacks against SCADA systems [11]. Neither of the papers presented a taxonomic view describing relationships between the areas they addressed and both focused on attacks while excluding many other relevant details such as actors, impact of the attacks or characteristics of the attacks such as attack vectors.

Moving outside the realm of traditional IT threat taxonomies, Cebula & Young created taxonomy of operational cyber security risks in 2010 which categorized risks into four classes: (1) actions of people, (2) systems and technology failures, (3) failed internal processes, and (4) external events. A valuable aspect of this taxonomy was its insight into the fact that risks can cascade and "that risks in one class can trigger risks in another class" [12]. This insight demonstrated the difficulty in trying to quantify events in a mutually exclusive manner when dealing with complex interactions in cyber security risk. This insight also holds true when trying to identify and classify the complex interactions involved in cyber conflict and was a contributing factor to the development of a network plex topology for the proposed taxonomy in this paper

3. REASONS TO CREATE A CYBER CONFLICT TAXONOMY

As the preceding section demonstrates, there are a number of previously developed taxonomies that address various aspects of cyber threats. While almost any cyber-attack can be categorized and described using these taxonomic frameworks, none of these previous frameworks are capable of illustrating the complex interactions between attacks, actors and other potentially related events and connecting them through logical links that formally describe their relationships. Previous taxonomies are valuable in classifying technical threats and vulnerabilities, but will fall short when it comes to linking actors with different methodologies, goals and patterns of behavior. For security practitioners operating in the realm of cyber conflict, understanding these interactions and the relationships between various aspects

of cyber conflict events can be critical in developing strategy and doctrine. For cyber operations practitioners who must develop doctrine and strategy, the ability to classify and study conflict related events from various taxonomic perspectives can give them unique insights that are not supported by previous works.

To address these issues, the proposed taxonomy has been developed to give users the ability to classify events and expose logical connections and links between different actors, types of attacks and vectors used and various types of impacts associated with each event. Once data is entered into the taxonomy, users can also look at cyber conflict events from discrete taxonomic perspectives such as looking at all events related to a particular actor or all attacks which use a social engineering vector, etc. and then explore the relationships between events and actors to look for commonalities that an operator could act upon.

4. PROPOSED TAXONOMY

The proposed Cyber Conflict Taxonomy is an extensible network taxonomy organized as a plex data structure. Each node in the taxonomy below the four primary category and subject headings can have more than one parent and any secondary or below level item in the plex structure can be linked to any other item based on defined relationships and classifications. This serves to organize the taxonomy into both hierarchical and associative categories which are useful in illustrating the many relationships that can exist between various nodes. The taxonomy is divided into categories and subjects. Categories are the taxonomic classifications that are applied to subjects and are further subdivided into subcategories. Subjects represent the real world events classified as cyber conflict and the real world entities such as individuals, groups or governments that participate in these events. Because cyber conflict involves interactions between states, non-state actors, and other competing entities, it is necessary to have a taxonomy that incorporates both events and entities and applies taxonomic classifications to them both in order to properly understand the complex relationships involved. The initial categories and subjects used in this taxonomy are defined below, however, since this taxonomy is designed to be extensible, additional categories and subjects may be added in the future as necessary.

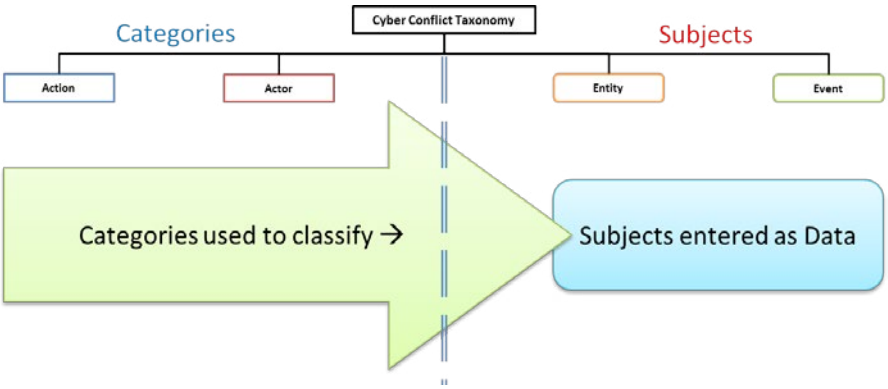

Figure 1. Cyber Conflict Taxonomy

A. SUBJECTS

Subjects are the actual real world cyber conflict related events and the individuals, organizations or states that participated in those events. Subjects represent the data objects that this taxonomy was meant to classify and are divided into Events and Entities. Subjects will always be linked to at least one category or subcategory and more than likely will be linked to multiple subcategories in order to provide accurate and discrete classification of the characteristics of the subject in question. Further subdivision of subjects, beyond Events and Entities, is not necessary for the taxonomy although specifications of subjects can be employed by the user to create logical groupings that may be useful when users wish to create groupings not covered by the actual classification scheme of the taxonomy.

Entities. The Entities subject heading is used to organize and list the actual, real world individuals, groups, organizations or governments that initiated, were targeted or took part in cyber conflict events. Entities will be classified using the Actors category of the taxonomy and will also be laterally linked to the specific Events in which they participated or in which they have suspected involvement. Entities can also be laterally linked to other entities with which they have a defined relationship. An example would be two entities which are directly politically opposed to each other.

Events. The Events subject heading is used to organize and list the actual, real world cyber conflict incidents which will be described in this taxonomy. Events will be hierarchically classified using the Actions category and subcategories of the taxonomy and will also be laterally linked to the specific Entities that participated in these events. Currently, Events are only organized by the specification Year in the

prototype, but no subdivision of Events is actually required by the taxonomy and this specification was added for the author's purposes.

- Year. The Year specification is an optional subdivision used in the prototype that allows a user to organize events temporally by the year or years in which they occurred. Many events related to cyber conflict span multiple years and it may be valuable for a user to be able to view events from this perspective

B. CATEGORIES

Categories represent the various forms of taxonomic classification used to describe the subjects of this taxonomy. The two primary parent categories in this taxonomy are Actions and Actors which are divided into subcategories as necessary to provide discrete and accurate descriptions of subjects. Subcategories are arranged hierarchically but are applied associatively to subjects so that any given subject will be described by multiple subcategories.

Actions. The Actions category is used to describe cyber conflict events and the characteristics of those events in a manner that is useful for researchers and operators. Actions are subdivided into attack and defense related subcategories.

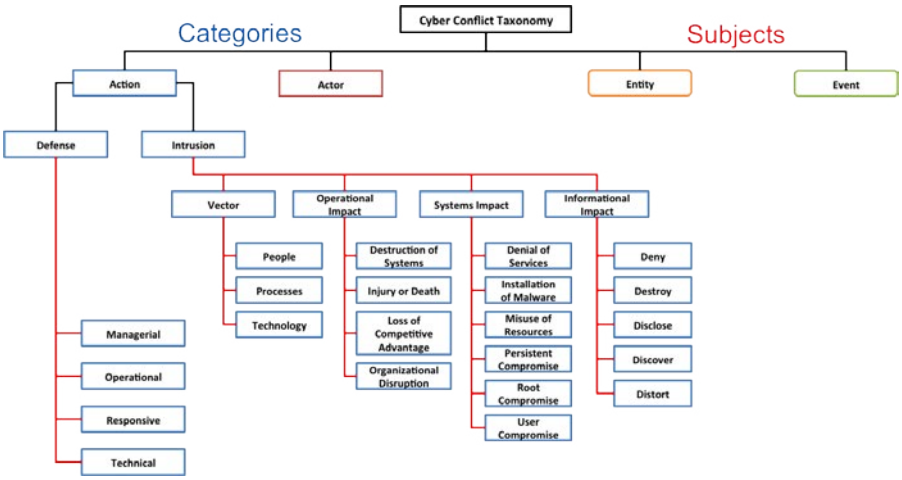

Figure 2. Actions Category of Cyber Conflict Taxonomy

- Intrusion. The Intrusion subcategory describes aggressive actions taken by one actor to affect other actors. Intrusions can be further divided into as many descriptive subcategories as necessary to describe said aggressive action. A

single intrusion may have many characteristics that must be classified in order to accurately classify the event in a complete and useful manner.

- Vector: This subcategory describes the path or means by which an attacker attempts to gain access to information resources or systems. This subcategory has been further divided into vectors which target people, processes or technology. Each of these subdivisions could be further subdivided into increasingly specific and discrete vectors as well.
  - People: This subcategory describes a vector based on the manipulation of people. An example would be the use of social engineering to gain credentials.
  - Process: This subcategory describes a vector based on the manipulation of flawed organizational processes. An example would be an organization that allows a visitor to hand carry their security credentials rather than mandating that the credentials be verified directly with the issuing source. An attacker might exploit this flawed process to illegitimately gain legitimate credentials to a system.
  - Technology: This subcategory describes a vector based on the manipulation of technology and technical processes. An example would be exploiting a vulnerability in a software program.
- Informational Impact: This subcategory describes the impact an intrusion has directly on the victim's information. This subcategory has been further divided into five additional child subcategories.
  - Deny: Denying legitimate users access to information within their own systems or networks.
  - Destroy: Destruction of information, usually through the permanent deletion of files, on a target system or network.
  - Disclose: Illegitimate access to or disclosure of sensitive, confidential or classified information.
  - Discover: Discovery of information previously unknown to an attacker which could potentially give the attacker additional advantages during follow on operations.
  - Distort: Distorting or changing information in a target system in a way that disadvantages the legitimate users of that information and or provides advantages for the attacker.
- Operational Impact: This subcategory describes the impact of an intrusion on the victim's operations. The term operational should not be misconstrued

to mean the operational level of war; it is used in this context to indicate the effects of an intrusion on the personnel, business processes and operations of the victim or victim organization.

- Destruction of Systems: Impact of an intrusion, which results in actual physical damage or the destruction of systems. The systems in question may be the actual information systems or other types of systems attached to or controlled by information systems. An example of this would be the damage to centrifuges that resulted from the Stuxnet attack.
- Injury or Death: Impact of an intrusion, which results in actual physical injury or death. This subclass could be further subdivided to differentiate between injury or death to human beings versus injury or death to non-human life. For example, a cyber attack which causes the injury or death of wildlife or livestock.
- Loss of Competitive Advantage: Impact of an intrusion which results in a victim organization losing its competitive edge due most likely to disclosure of plans, proprietary information, classified information or confidential technical data. An example would be a competitor state stealing data from a defense contractor related to a classified technology which enables it to reverse engineer this technology for its own use.
- Organizational Disruption: Impact of an intrusion, which causes the disruption of operations within an organization. An example would be altering information in a supplier database system to reroute critical supplies to the wrong destinations.

○ Systems Impact: This subcategory describes the impact of an intrusion on the actual information systems of the victim organization.

- Denial of Service: Denying a victim access to information resources or system services.
- Installation of Malware: The installation of malicious software onto the target host or system beyond what is required for the initial compromise of the system in question.
- Misuse of Resources: An unauthorized use of system resources. This may consist of any system related function that requires certain elevated privileges and those privileges are then converted into abusive action [9].
- Persistent Compromise: Gaining a persistent foothold on a particular host or within a particular network that goes undetected for an extended period of time. This type of compromise may remain undetected for months or even years and is usually used to facilitate other actions.

- Root Compromise: Gaining unauthorized root or administrative privileges on a particular host or system.
- User Compromise: Gaining unauthorized use of a non-administrator's user privileges on a particular host or system.

- Defense. The Defense subcategory describes actions taken by an actor to protect their information systems from attacks. Defense is divided into Managerial, Operational, Responsive and Technical subcategories, which can be further subdivided into more specific subcategories as is necessary for the user. Three of these subcategories roughly align to the security controls advocated by the National Institute of Standards and Technology [13]. The fourth subcategory, Responsive defenses, expands on the NIST standard to account for more active responses such as counter-attacks which would not be seen in a commercial setting but which could certainly be used in a cyber conflict setting.
  - Managerial: Defensive techniques and methods, normally addressed by management, regarding an organization's computer security strategy.
  - Operational: Defensive strategies based on policies and procedures implemented and executed by people, as opposed to systems, to improve the security of a system or group of systems.
  - Responsive: Direct responses to a malicious intrusion targeting the source of the intrusion. Examples could include counter-attack or counter-reconnaissance.
  - Technical: Defensive tools or strategies executed by automated systems to improve the security of individual systems or a group of systems.

Actors. The Actors category classifies the entities participating in cyber conflict by type. Currently, this category is divided into two subcategories; Non-State Actors and State Actors. These subcategories may be further divided down as needed.

- Non-State Actors: The Non-State Actors subcategory describes entities participating in cyber conflict events, which have no known ties to government entities.
- State Actors: The State Actors subcategory describes governments, government organizations or government sponsored entities that participate in cyber conflict events.

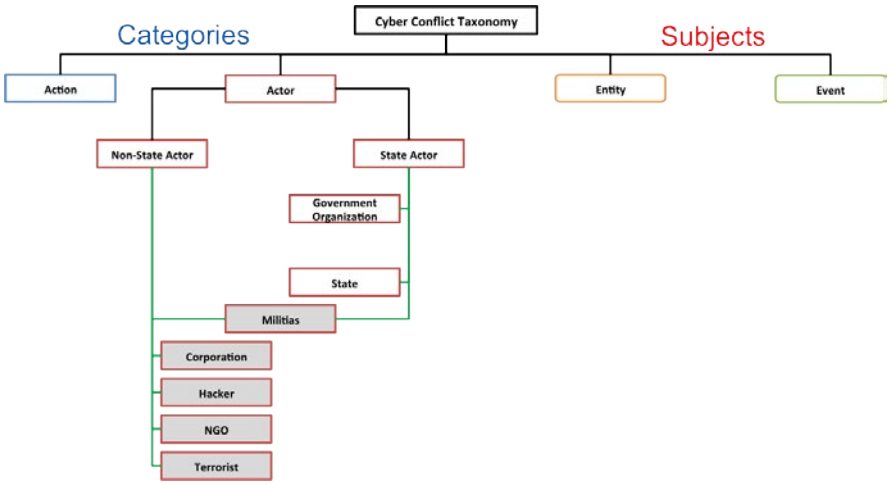

Figure 3. Actors Category of Cyber Conflict Taxonomy

C. TYING IT ALL TOGETHER

In order to begin testing the usefulness of the proposed taxonomy, two prototypes were developed. The first prototype was modelled using mind-mapping software called *The Brain*. Version 7 of this software was used for the development of the initial prototype. This software was used to rapidly build and visualize the proposed taxonomy. This first prototype provides the ability to show multiple child- and parent-relationships hierarchically in a network plex and to laterally link related entities and events together depicting the causal relationship between various subjects. The prototype also allows the user to define the different types of relationships that link nodes together throughout the taxonomy and to color-code, tag and categorize both nodes and links. This allows the user to search or filter the taxonomy based on key words, node types or even relationship types.

A sample set of a ten real world events was entered into the taxonomy as Events and then classified using the categories and subcategories previously described. Additionally, more than fifty entities were additionally entered into the taxonomy based on their relationship to the previously entered events. These entities represented the actors involved in these events, including those suspected of involvement in cases where definitive attribution (i.e. most cases) could not be established. This prototype proved to be very useful in developing classification categories and in visualizing the data entered into the taxonomy. The main limitation of this prototype, based primarily on the software package used to develop it, was the need to manually link each subject entered into the taxonomy to the various categories

and subcategories that would apply to it. For a large data set, this would be a very tedious task prone to omissions and errors. Ideally, a fully automated and polished version of the taxonomy would include simple drop lists with all the categories from which the user could select multiple classifications simultaneously to describe the subject. Additionally a similar list of subjects would be available to simultaneous select related or causal subjects as well.

A second prototype of the proposed taxonomy was modelled using *Protégé* version 4.1. *Protégé* is a free, open-source platform that provides a suite of tools to construct domain models and knowledge-based applications with ontologies using Web Ontology Language. Use of *Protégé* for the second prototype allowed for more formal and rigorous definitions of the relationships between entities and categories and provided a platform capable of more easily identifying trends in the knowledge base. In defining relationships, *Protégé* allows for the specification of domains and ranges for each relationship. It allows additional facets of such relationships to be specified such as transitive, functional, symmetric, asymmetric and reflexive properties. Additionally, due to the open source nature of the software, it would be easier to alter this platform to provide for easier data entry due to the availability of the original source code.

5. APPLICATION EXAMPLES

To demonstrate the use of the proposed taxonomy three examples are shown below all related to the same event, Operation Shady RAT. This event is shown from three different taxonomic perspectives; one view with the event as the central node in the taxonomy, one from the perspective of one if the event's systems impacts, and finally, a view from the perspective of its suspected initiator. Each view shows different characteristics of the event and illustrates the potential relationships between this event and other entities or events. It should be remembered that in the examples below, only a limited data set of ten events was entered into the prototype.

A. OPERATION SHADY RAT – TAXONOMIC VIEW OF AN EVENT

Operation Shady RAT was a targeted set of intrusions into more than 70 global companies, governments and non-profit organizations that took place from 2006 to 2011 [14]. When entered into the prototype taxonomy (see Fig. 4), the result shows links to the actors which were targeted, the suspected initiating actor, the years over which the event took place, and the various types of impacts. Additionally, other events are shown which took place during the same time frame, which had similar types of impacts, or which were related to the actors listed.

This initial view gives an operator a starting point to begin studying related events in order to look for trends or patterns in the data such as, for example, looking at other events which involved the installation of malware on targeted systems.

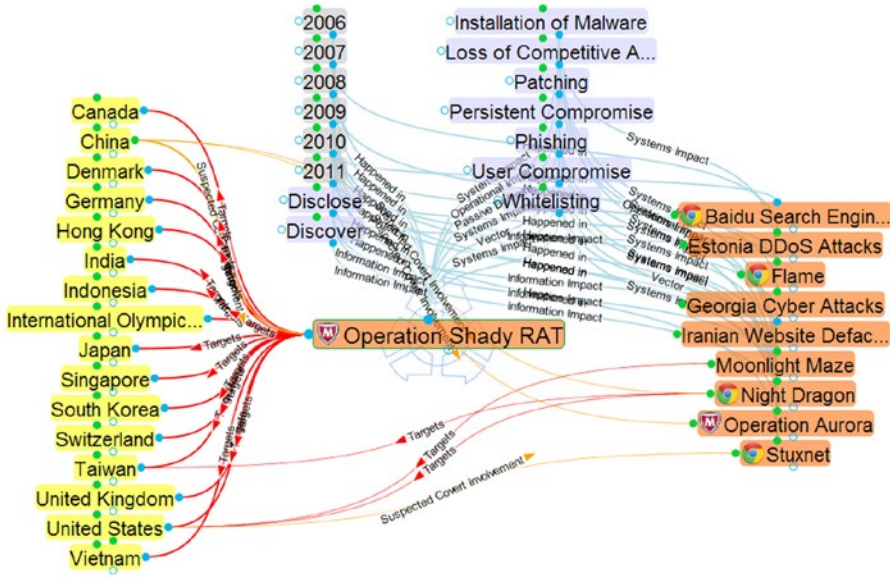

Figure 4. Taxonomic View of Operation Shady RAT

B. INSTALLATION OF MALWARE – TAXONOMIC VIEW OF A SYSTEMS IMPACT

To view this event from a different taxonomic perspective, an operator can simply select one of the categories by which the event was characterized such as the Systems Impact – Installation of Malware. As can be seen in Fig. 5, this view shows the user other events which shared this same systems impact. Additionally it show links from these other events to additional systems impacts they exhibited allowing the operator to compare impacts of similar events.

C. CHINA – TAXONOMIC VIEW OF AN ENTITY

To view Operation Shady RAT from the perspective of the suspected initiating actor, the operator can select the State – China (see Fig. 6). This perspective shows other events in which China is suspected to have been involved and also displays which other actors were targeted by these events.

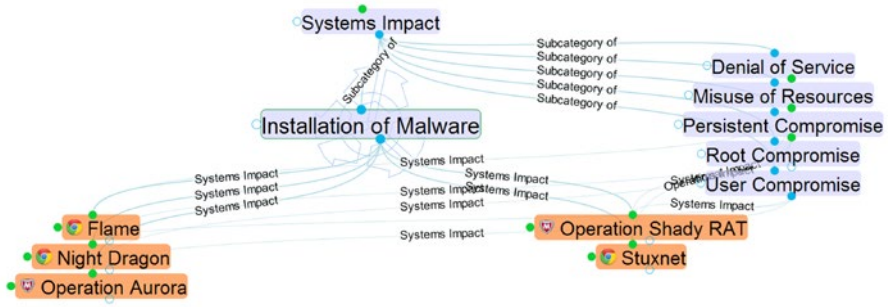

Figure 5. Taxonomic View of Systems Impact: Installation of Malware

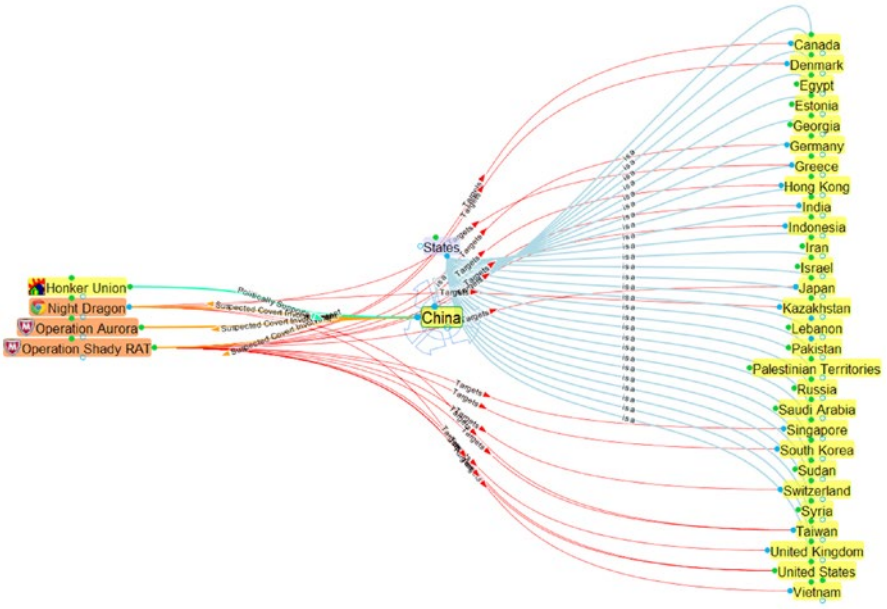

Figure 6. Taxonomic View of Actor: China

If the network plex is expanded by one additional level of connectivity the complexity of the events and interactions related to China becomes apparent. Relationships that are separated by 3 or 4 degrees of separation can now be illustrated and users can look for insightful patterns of behavior or similar methodologies. This expanded plex shows other state and non-state actors involved in similar events, the targets of these events, the time frame of these events and other related information such as the political allegiances of various non-state entities illustrated by the extended connectivity.

D. COMPARISON OF OTHER RELAVANT TAXONOMIES

Using Operation Shady RAT as a case study, the proposed taxonomy in this paper was studied in a side-by-side comparison with two other taxonomic systems previously discussed above. Howard's Computer & Network Attack Taxonomy classifies attacks using five classification categories: Attacker, Tools, Access, Results and Objective [3]. Table I shows the result of classifying Operation Shady RAT using this Taxonomy. While this taxonomy does provide some important information about this attack, it lacks a couple of important characteristics such as vector, defensive actions and the specific actors involved.

Table I. Classification of Operation Shady RAT using Howard's Taxonomy

Howard's Taxonomy								
Name	Attacker	Tools	Access				Results	Objective
Shady RAT	Spies	Toolkit	Design & Config Vulnerabilities	➡	Unauthorized Use Unauthorized Access	➡	Files	Compromise of Information Disclosure of Information	Poloitical & Finaicial Gain

The AVOIDIT Taxonomy also classifies attacks using five classification categories: Attack Vector, Operational Impact, Informational Impact, Defense and Target. Table II shows the result of this classifying this attack using the AVOIDIT Taxonomy. While this taxonomy does improve on Howard's in some key areas such as attack vector and defensive strategy, it still lacks specificity when it comes to identifying actors involved in this attack.

Table II. Classification of Operation Shady RAT using AVOIDIT Taxonomy

AVOIDIT Taxonomy					
Name	Attack Vector	Operational Impact	Informational Impact	Defense	Target
Shady RAT	Spear Phishing	Installed Malware: Trojan	Discovery Disclosure	Remediation: Patch System, Whitelisting	Network

Classifying Operation Shady RAT using the proposed taxonomy, the first thing that becomes apparent is the inclusion of all the actors involved in this event (see Table III.). A compressed list was used for this paper as the original attack targeted more than 70 organizations across 14 nation-states. This taxonomy also differentiates between Systems Impact and Operational Impact while the AVOIDIT Taxonomy only highlights the technical impact of attacks on systems and excludes the impact of attacks on the target's operations. All information from the AVOIDIT Taxonomy is accurately captured in the proposed taxonomy and all information from Howard's taxonomy, with the possible exception of the vulnerability portion of Access, are also captured.

Table III. Classification of Operation Shady RAT using Cyber Conflict Taxonomy

Cyber Conflict Taxonomy								
Name	Vector	Informational Impact	Operational Impact	Systems Impact	Defense	Actors		
Shady RAT	Spear Phishing	Discover Disclose	Loss of Competitive Advantage	Installation of Malware; Persistant Compromise	Passive - Whitelisting, Remediation - Patching	**Targets:** Denmark Hong Kong Indonesia Japan South Korea Taiwan Unites States	Canada Germany India IOC Singapore Switzerland United Kingdom Vietnam	**Source:** China (Suspected)

An important feature of the proposed taxonomy that is not addressed in all of the previous taxonomies is the ability of this taxonomy to identify related subjects (both entities and events). Looking back at Fig. 4, a group of related events appears on the right hand side of the image (the 9 items which are circled). These events all share some of the characteristics of Operation Shady RAT. They may use the same vector, target the same states or organizations, or may have just happened in the same timeframe. Three of the nine events identified share a high degree of similarity with Operation Shady RAT and could potentially be related to this event. Given that this prototype had a very limited test-set, it is easy to see how this capability would be useful for researchers and planners working in the cyber operations domain. This capability can assist a researcher in attributing an anonymous event to a specific actor based on similarities in methodology, impacts and target sets.

Each of the above taxonomic frameworks can provide useful information; however, the proposed taxonomy provides the most robust classification scheme and provides the ability to identify related subjects. This improvement on previous taxonomic frameworks and the focus on cyber conflict events at an operational level make this proposed taxonomy a useful tool for both security researchers studying cyber conflict and for planners and operators working in the domain of cyber operations.

6. LIMITATIONS AND FUTURE RESEARCH

Over the course of this research, a number of limitations were identified in relation to the use of a taxonomy to evaluate cyber conflict events. Introducing such a taxonomy to classify the events and entities involved in cyber-conflict is important and offers a good first approximation of what a security analyst can derive and potentially plan for when it comes to cyber operations. However, there are inherent limitations that stem from the use of a taxonomy, which is a hierarchical categorization of entities within a domain. A taxonomy does not allow for any formal or empirical

relationships among the entities beyond parent-child relationships. To capture most, if not all possible relationships and characteristics between different actors and events, a more formal mechanism such as an ontology is needed. Unlike a taxonomy, an ontology allows for the formal description of multiple relationships between entities in an empirical manner. The creation of the second model using Protégé and OWL constituted the first step in this process and will be used in future research to expand the scope of this project. Once this second model has been more extensively defined and tested, a larger data set will be used to validate the model's ability to identify commonalities between related events.

7. CONCLUSION

This paper presents a taxonomy for classifying cyber conflict events and the entities involved in these events. All data are entered into this taxonomy as subjects and then classified according to the categories and subcategories used to describe the characteristics of these subjects. A prototype was developed which demonstrated that the proposed Cyber Conflict Taxonomy is useful in categorizing and describing events and entities involved in cyber conflict in a manner that would be beneficial to researchers and operators. All events and actors entered into the prototype were fully describable using the proposed categories. Even with a limited data set, the ability to study linkages between related subjects demonstrates patterns and provides researchers with insights into commonalities between different events and entities and would be useful when developing doctrine and strategy. This feature is unique to this taxonomic model and is an improvement on previous frameworks. It can potentially allow an operator to identify actors across different events based on their similar method of operation, toolsets and target sets.

Finally, this taxonomy is designed to be extensible so that users can categorize the characteristics of cyber events or entities using increasingly discrete descriptions. This allows this framework to be as specific as necessary for various purposes. For future work, a much larger data set should be created and empirical studies undertaken to validate the taxonomy's ability to identify commonalities between related events.

Acknowledgements

The authors would like to gratefully acknowledge the efforts of LTC André Abadie, COL Jody Prescott (Ret.), and Dr. Duminda Wijesekera who assisted in the editorial review of this paper. Portions of this project were conducted using the Protégé resource, which is supported by grant LM007885 from the United States National Library of Medicine.

REFERENCES

[1] Lambe, P. (2006, April 18). Defining Taxonomy. Retrieved from Green Chameleon: http://www.greenchameleon.com/gc/blog_detail/defining_taxonomy/

[2] Bishop, M. (1995). A Taxonomy of UNIX System and Network Vulnerabilities (University of California at Davis No. Report CSE-95-10). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.5712

[3] Howard, J. D. (1997). An Analysis of Security Incidents on the Internet 1989-1995 (Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA, 1997). Retrieved from www.cert.org/archive/pdf/JHThesis.pdf.

[4] Hansman, S., & Hunt, R. (2004). A taxonomy of network and computer attacks. Computers & Security, 24(1), 31-43. http://dx.doi.org/10.1016/j.cose.2004.06.011

[5] Killourhy, K. S., Maxion, R. A., & Tan, K. M. C. (2004). A Defense-Centric Taxonomy Based on Attack Manifestation. Presented at the International Conference on Dependable Systems & Networks, Florence, Italy.

[6] Mirkovic, J., & Reiher, P. (2004). A Taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication Review, 34(2), 39-53. http://dx.doi.org/10.1145/997150.997156

[7] Kjaerland, M. (2006). A taxonomy and comparison of computer security incidents from the commercial and government sectors. Computers & Security, 25(7), 522-538. Retrieved from http://dx.doi.org/10.1016/j.cose.2006.08.004

[8] Rogers, M. K. (2006). A two-dimensional circumplex approach to the development of a hacker taxonomy. Digital Investigation, 3(2), 97-102. Retrieved from http://dx.doi.org/10.1016/j.diin.2006.03.001

[9] Simmons, C., Ellis, C., Shiva, S., Dasgupta, D., & Wu, Q. (2009). AVOIDIT: A Cyber Attack Taxonomy. Retrieved from http://issrl.cs.memphis.edu/files/papers/CyberAttackTaxonomy_IEEE_Mag.pdf

[10] Fovino, I. N., Coletta, A., & Masera, M. (2010, March). Taxonomy of security solutions for the SCADA Sector, Deliverable: D 2.2, Version: 1.1. A European Network For The Security Of Control And Real Time Systems.

[11] Zhu, B., Joseph, A., & Sastry, S. (2011). A Taxonomy of Cyber Attacks on SCADA Systems. IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing. DOI 10.1109/iThings/CPSCom.2011.34

[12] Cebula, J. J., & Lisa, R. Y. (2010). A Taxonomy of Operational Cyber Security Risks (Carnegie Mellon University / Software Engineering Institute No. CMU/SEI-2010-TN-028). Retrieved from http://www.sei.cmu.edu/library/abstracts/reports/10tn028.cfm

[13] National Institute of Standards and Technology (2009). NIST Special Publication 800-53 Revision 3: Recommended Security Controls for Federal Information Systems and Organizations. National Institute of Standards and Technology. United States Department of Commerce. Gaithersburg, MD.

[14] Alperovitch, D. (Vice President, Threat Research, McAfee). (2011). Revealed: Operation Shady RAT. McAfee. Retrieved from http://www.mcafee.com/us/resources/white-papers/wp-operation-shady-rat.pdf.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

Cyber Attack: A Dull Tool to Shape Foreign Policy

Emilio Iasiello
iSight Partners
iasiello@aol.com

Abstract: This paper examines how cyber attacks, if indeed conducted by nation states, have been unsuccessful in supporting states' foreign policy objectives. By analyzing three prominent case studies, I show that as a result of geopolitical tensions, cyber attacks were implemented to further nation state objectives in support of foreign policy considerations and failed to achieve their respective outcomes despite successful deployment against their intended targets. The three case studies, hypothetical scenarios because attribution has not been confirmed, include: (1) the October 2012 distributed denial of service attacks targeting the U.S. banking sector; (2) the 2012 Stuxnet attack against Iran; and (3) the 2007 cyber attacks against Estonia. I work with the assumption that nation states were orchestrating the attacks through proxies, or else were actual participants, based on intent, motive, and a plethora of circumstantial evidence presented in each scenario. Data has been collected from newspapers, information technology security periodicals, and expert analysis. This paper challenges the notion that states can use the threat of cyber attack to influence an adversarial nation state's behavior, much the same way the threat of nuclear weapons holds other states in check.

Keywords: *cyber attack, DDOS, Stuxnet, Estonia, Iran, foreign policy*

1. INTRODUCTION

In 2007, the Internet security company McAfee published a report citing that approximately 120 countries already possessed or were developing capabilities to conduct offensive cyber operations.[1] Ostensibly, such capabilities would enable a nation state to perform cyber attacks or conduct cyber espionage at home or abroad against adversaries and allies alike, depending on the intent. Since the publication of that report, there have been several world events that have demonstrated a different, more strategic purpose behind cyber attacks: use as a tool to exert control and gain political influence. In particular, the 2007 distributed denial-of-service attack against Estonian government networks; the 2012 unofficial acknowledgement by a senior government official of the United States involvement in deploying Stuxnet and Flame (and possibly Duqu and Gauss) against a network controlling Iranian nuclear centrifuges, as well as other Middle Eastern networks;[2] and the 2012 DDOS against U.S. financial institution networks hint of nation state direction and/or involvement based on intent, target selection, and the desired effect created. Why these events are significant is that a deeper motivation lurked beneath the perpetrators' intent to just disrupt, deny, degrade, or destroy information systems or the information resident on them; these attacks were designed to create opportunities and influence events to gain political advantage. However, while it can be argued that these cyber attacks succeeded in accomplishing their tactical missions, they ultimately failed in their strategic objectives. At present, authoritarian governments such as China and Iran have achieved some modicum of success using offensive cyber operations to monitor and censor hostile information from reaching its public, or identifying and targeting political dissident and oppositionist groups that pose a threat to regime stability. Yet even the most draconian restraints are subject to circumvention by the more technically savvy and diligent oppositionists, reducing the overall effectiveness of these technical measures. More importantly, the capability to exert influence over an indigent populace does not hold the same authority as being able to use the same capabilities to influence decision makers in a country next door, no less one thousands of miles away. Whereas nuclear weapons have been used as the platforms from which nuclear states have flexed diplomatic muscle, cyber weapons have not yet reached that revered plateau. To date, cyber weapons have failed to wreak the awe inspiring havoc of their nuclear counterparts and thereby cannot be used as a saber rattling tool of foreign policy. Until such a time, cyber weapons will continue to be more terrifying in theory than practice, and remain a pejorative sound bite used by politicians and military hawks as the harbinger of future threats than an effective policy prescriptive tool for today's governments and militaries.

2. NUCLEAR WEAPONS VS CYBER WEAPONS

Ever since the two atomic bombs were dropped on Hiroshima and Nagasaki, the world has been privy to the devastating effects of nuclear weapons. In 1945, a yardstick had been inadvertently established that would forever become the benchmark for future nuclear weapons development. Superpowers now had a standard by which to convey their might to their adversaries. The implied threat was clear: cross any "red lines" and an assured destruction would take place. The Cold War ushered in a near twenty-year global race for nuclear supremacy. What capability one country had, competitors and allies alike wanted as well. Since 1945, eight countries are believed to have some level of nuclear weaponry; one is believed to have weapons although has not publicly acknowledged it; and in the case of Iran, one may or may not be actively trying to develop them.

Most nuclear states will readily admit that the reason for possessing such a capability is largely to deter the potential hostile actions of their enemies. Nuclear weapons, or the very threat of their acquisition, have succeeded in bringing powerful nations to the diplomatic table. North Korea has consistently used the threat of nuclear weapons development as a bargaining chip to achieve tactical objectives such as receiving food and humanitarian aid. Israel's assumed possession of nuclear weapons has helped enable it maintain its strong position in the Middle East, and Iran's pursuit of this capability can be interpreted as its attempt to gain regional supremacy and create a level playing field with its main adversary, Israel. Nuclear deterrence theory is largely rooted in the fact that a nation state has the capability and credibility to deploy nuclear weapons as does its adversary, thereby creating a military stalemate.

As cyber claims the prize as the 21st century's greatest non-nuclear threat, many experts and influential people such as U.S. Cyber Command's General Keith Alexander and former deputy assistant director for the FBI's Cyber Division Steven Chabinsky believe that offensive actions can be applied to this asymmetric domain to deter hostile adversarial actions in cyberspace. The United States in particular has taken initiatives in getting its military involved in addressing the cyber problem. In May 2011, the White House released its *International Strategy for Cyberspace* outlining how it would approach its use of cyberspace, promoting its core commitments to fundamental freedoms, privacy, and the free flow of information. While not explicit, the strategy states that the U.S. "reserves the right to use all necessary means – diplomatic, informational, military, and economic – as appropriate and consistent with applicable international law."[3] What can be inferred from this is U.S. intent to employ whatever tools at its disposal to defend itself, its allies, its partners, and its interests. In July 2011, the U.S. Department of Defense released its *Strategy for Operating in Cyberspace,* which clearly articulated

the role of the U.S. military to ensure that it has the necessary capabilities to operate effectively in the cyberspace, citing it as a military domain much like air, land, and maritime,[4] with U.S. Cyber Command leading the efforts to "conduct full-spectrum military cyberspace operations to ensure U.S. and allied freedom of action in cyberspace, while denying the same to their adversaries."[5] Whether it wants to admit it or not, like in the nuclear domain, the United States has taken the first step toward militarizing cyberspace.

However, as several articles and publications attest, there is no cyber equivalent to nuclear deterrence, based largely on four factors: 1.) Nation states typically do not assume responsibility for hostile actions taken in cyber space; 2.) There has been no awe inspiring, game changing, show of what a cyber attack can do; 3.) Attribution in cyberspace is extremely difficult and can't be as precise as identifying a nation state that has launched a nuclear weapon, and 4.) Unlike nuclear weapons development, which can be monitored, there is no similar transparency for nation state production of cyber weapons, nor an international watchdog agency to track such developments. These actions ultimately hinder nation states from applying a similar mutual assured destruction concept to cyberspace. Therefore, as the following case studies will demonstrate, nation states using cyber attacks to support foreign policy objectives will ultimately fail to effectively influence decision makers into favorable courses of action, or deterring political and/or military actions.

3. HYPOTHETICAL CASE STUDIES

These hypothetical case studies show how suspected nation state cyber activity was used in the hopes of obtaining a political objective outside its tactical mission. To date, there has been no definitive nation state attribution or linkage to these case studies. While attribution in cyberspace is nearly impossible, certain elements such as actor motivation, intent, behavioral actions, actor profiling can be identified, evaluated, and assessed to help better understand the significance of these cyber events – not just from an operational level, but from a strategic perspective as well.

A. 2007 ESTONIA DISTRIBUTED DENIAL-OF-SERVICE ATTACK

Note: It is difficult to discuss the 2007 cyber attacks against Estonia without bringing up the 2008 distributed denial-of-service (DDoS) attacks directed at Georgia. While both cyber campaigns were suspected instruments of the Russian government conducted through proxies, they differ in that the 2008 attacks were conducted in tandem with the invasion of Georgia's Ossetia region by Russian armed

forces. I assume the premise that the Russian government had at least an informal role in directing the activity. Therefore, if the Russian government was involved, and the DDoS activity was an instrument of its foreign policy, its objective could be construed as an attempt to influence an Estonian government course of action favorable to Russian interests.

In late April 2007, Estonia relocated a Soviet-era bronze statue and inadvertently opened a virtual Pandora's Box of cyber malfeasance against its national networks. For three weeks, Estonia was the victim of politically-motivated cyber attacks targeting websites of political parties, and distributed denial-of-service (DDoS) attacks against governmental, and commercial organizations to include schools, Internet service providers, media channels, and private websites.[6][7] Given that Estonia has long been considered a modern nation and among the most wired countries in the world, this was a serious concern. A flummoxed Estonia searched for courses of action, enlisting international support to mitigate the cyber attacks. The world was witness to what it had long heard about but up until this point had never seen – cyber attacks shut down a country's information infrastructure.

1) It Was The Russians…I Think

Bilateral tensions between Russia and Estonia are deep rooted and have been festering for over eighty years. From 1918 when Estonia first gained its independence from Russia until 1991 when it regained its ultimate independence, Estonia has viewed Russia's presence as an illegal occupation.[8] Moscow's attempts to "Russify" Estonian culture has been compounding this friction, relocating hundreds of thousands of ethnic Russians to Estonia starting in 1940 and continuing throughout the Cold War.[9] Suffice to say, foreign relations between these two countries remain a constant work in progress, ebbing and flowing according to the diplomatic and political environments. Estonia's relocation of the Soviet war memorial from the center of Tallinn to a military cemetery in 2007 was the catalyst for physical and digital protests. At the time, Estonia had a substantial Russian-speaking population, almost a third of Estonia's population of 1.3 million.[10] Patriotic hackers and established Russian youth groups who had previously engaged in hostile cyber activity against Chechen websites,[11] immediately mobilized to defend Russian nationalist interests. Clearly Russian government reaction to the Estonian government was bold and resolute, threatening to sever diplomatic relations and calling Estonia's statue removal as "blasphemous and barbarous."[12] The scene was set – a showdown between a powerful nation state and its smaller, younger, and highly individualistic cousin.

So why a cyber attack?

The 2007 cyber attacks against the Estonian information infrastructure can be

interpreted as an instance where a nation state tried to influence the decisions and actions of another country using cyber weapons. The three-week long DDoS achieved several different outcomes including the expression of diplomatic discontent; the flexing of "virtual" muscles; and the capturing of the Estonian government's attention. More importantly, the world observed firsthand the potential consequences of a serious cyber attack.

- **Expressing Diplomatic Discontent**: The fact that the Pro-Kremlin youth group "Nashi" immediately claimed responsibility for the cyber attack was a signal to the Estonians that the Russian government may have had influence on the group allegedly behind the attacks, if only as the puppet master pulling the strings. Nashi had an established track record of working on behalf of Moscow to include spying on other youth groups and conducting DDoS against unfriendly newspapers.[13] Furthermore, a State Duma official acknowledged a relationship with a Nashi "commissar intimating a possible collusion between the Russian government and this group with regards to the attack.[14] So the message was clear: While it couldn't be proven, the Estonians and the rest of the world for that matter saw Russia's hand in this attack.
- **Flexing "Virtual" Muscle**: The DDoS did not target a sector or a specific organization but a *nation's information infrastructure*. This wasn't a mistake or a serendipitous happenstance, rather, a calculated, planned operation that systematically executed an attack that increased in intensity throughout its duration. As one of the most wired countries in the world, a potent DDoS disrupting but not destroying key services seemed to be sending a potent message: "If our youth group hacktivists can do this to you, imagine what the full fury of the Russian government can do?" Up until that point, DDoS attacks had been primarily used by hacktivists and patriotic hackers to express discontent, but none had ever achieved the magnitude of these attacks.
- **Capturing Estonia's Attention**: The DDoS attack can be considered a virtual "slap" to get Estonia back in line. The fact that the DDoS ended as quickly as it had started further supports the fact that it was a measured response designed to make a statement, rather than cause permanent or irrevocable damage. So the intent was not to bring down the system, which could have easily been done. Russia initially postured, threatening to sever diplomatic relations as a result of the statue's relocation, suggesting that if the statue were replaced, diplomatic relations would be reinstituted. When physical protests did nothing to dissuade the Estonian government, DDoS attacks quickly targeted government networks at the onset. Throughout its duration, the DDoS increased in intensity, particularly on days of historic significance, such as the

on May 9 – "Victory Day" in Russia commemorating the capitulation of Nazi Germany to the Soviet Union in 1945.

2) But Did It Work?

Clearly from a tactical standpoint, the DDoS attacks against the Estonian information infrastructure were an unqualified success. For three weeks, Estonia was the target of these attacks. Each time there was a pause in the activity, it would resurface soon after stronger and more potent than earlier iterations. What's more, the attackers constantly tweaked their malicious server requests to evade filters.[15] More lasting damage could have been done but wasn't.

Estonia remained resolute and did not surrender or acquiesce to Russia's demands. Instead, Estonia solicited international support in mitigating the cyber attacks. Estonia's unique situation encouraged NATO to consider the repercussions of cyber attacks when directed against a nation state, incentivizing NATO's creation of a cyber center of excellence to improve NATO's cyber defense posture.

If involved, Russia may have correctly anticipated NATO's reluctance to consider enacting Article 5 and opting to provide defensive network support instead of escalating the situation by rallying NATO members behind Estonia. Although Russia, if they were involved, might have estimated this course of action correctly, it was not a guaranteed outcome. At the time, while there was no precedent to addressing a state-level cyber attack, NATO had a history of intervening on behalf of a weaker actor as evidenced by its military operations in Bosnia and Herzegovina in 1994 and in Yugoslavia in 1999, for example. While it couldn't conclusively be determined that Russia was behind the cyber attacks, circumstantial evidence certainly pointed in its direction: some Russian computers being involved in the DDoS (as well as other countries in the world)[16], its perceived culpability in the instigating the riots in Tallinn, and the fact it made no overtures to stop or halt the attacks coming from its information space could have encouraged NATO to approach Russia. Diplomatic channels would invariably be exercised. Worst case scenario, diplomatic efforts would fail to achieve positive results, cooling relations between Russia and the Alliance. Therefore, when viewed as an instrument of policy, the DDoS attacks could be considered an unqualified failure that ran the risk of worsening formal relations or escalating into an international incident.

B. STUXNET – CYBERWARFARE HAS ARRIVED

Note: As of this writing, it is largely believed that the United States, and perhaps Israel, was involved in the creation and execution of Stuxnet. While unsubstantiated, this assertion gained additional legitimacy when an unidentified senior administration official "leaked" similar information[17]. If the U.S. government was behind

the Stuxnet attack, the successful deployment of the weapon combined with the unofficial "leak" could have served an important U.S. foreign policy objective – to demonstrate to the Iranian government the complexities of U.S. capabilities and its ability to impact Iran's most sensitive programs.

In 2010, Iran publicly disclosed that a cyber weapon had damaged gas centrifuges in the uranium enrichment facility at Natanz. First identified by VirusBlokAda, Stuxnet was described as "highly sophisticated" and a complex application designed for the sole purpose of sabotaging uranium enrichment centrifuges controlled by high-frequency converter drivers used by the uranium enrichment facility at Natanz.[18] The malware successfully impacted a significant number of Iranian centrifuges, targeting a specific type of industrial controller and causing almost 1,000 of them to spin out of control.[19] That malware had been injected into a network not connected to the Internet was nothing new. Individuals are constantly infecting stand-alone machines and networks via the witting or unwitting insertion of infected removal media. The significance of this event was the fact that this was the first documented incident where an actual cyber weapon was deployed whose intention was to deny, degrade, disrupt, and destroy a specific information system target. What's more, the sophistication of the malware, its functionality, the intent behind its deployment, and its clandestine appearance on a non-Internet connected industrial control system network pointed a finger squarely at nation state sponsorship, thus ushering in the first instance of cyberwarfare. Suspicions that the United States and Israel were the possible perpetrators of this act were bolstered but not confirmed in 2012 when an unnamed U.S. official acknowledged U.S. involvement as part of its classified "Olympic Games" program initiated by President George W. Bush and continued by President Barack Obama. The U.S. government has made no official pronouncement on the subject.[20]

1) Why Stuxnet?

It's little surprise that relations between the U.S. and post-Revolution Iran haven't been friendly. Iran's heated rhetoric, extremist religious views, and insistence on its sovereign right to develop nuclear power have caused the U.S. great concern over its intentions to use that capability to develop weapons grade uranium. Indeed, for the past few years, U.S. has closely monitored Iranian uranium development, and has even offered technological and economic incentives through international cooperation as a viable alternative to developing the capability indigenously and without regulatory oversight. Iran has consistently brushed aside these overtures, withstanding an onslaught of increasingly severe economic sanctions to affirm its right to nuclear self development.

If true, President Bush's decision to employ a cyber weapon of this caliber[21] was commendable in the fact that he saw this as a viable non-lethal option as opposed

to approving a conventional military strike. Already embroiled in two military conflicts, the deployment of Stuxnet could have been intended to impede Iran's nuclear development without alerting them that this was the result of clandestine sabotage. For two years after its discovery, the U.S. remained tight-lipped about its role in Stuxnet despite international suspicions to the contrary. So if they were culpable, why would the U.S. publicly "leak" their involvement in Stuxnet in 2012? Some key events that transpired in the spring provide some illumination:

- March 14, 2012: In an interview with CNN, Iranian officials reiterate that nuclear inspectors would not be allowed to return.[22]
- March 5, 2012: Israeli President Benjamin Netanyahu travels to the United States and warns that a diplomatic solution to Iran's nuclear threat is running out.[23]
- March 3, 2012: U.S. President Barack Obama states that all elements of American power remain an option to prevent Iran from becoming a nuclear power.[24]
- February 24, 2012: The International Atomic Energy Association (IAEA) reports that Iran has significantly stepped up its uranium enrichment program and has concerns about potential military uses.[25]

These events show that over the course of 2012, Iran continually demonstrated its intentions to continue enriching uranium despite the European Union and United States economic sanctions and international disapproval levied against it. Therefore if the unknown U.S. official's admission of deploying Stuxnet is true, it can be interpreted as removing any doubt over U.S. involvement in trying to impede Iran's nuclear development. Not only would it have demonstrated the United States' sophisticated capabilities in the development of advanced cyber weaponry; but it also would have shown that it could "touch" Iran's most secret nuclear development facilities any time it wanted.

2) But Did It Work?

Aside from being a technological marvel at the time of its discovery, it is debatable if the deployment of Stuxnet achieved its true intended results. While U.S. officials might conclude the success of Stuxnet at "delaying" Iran's nuclear progression, it did not significantly impact Iran's plans or its ability to enrich uranium. On the contrary, the discovery of Stuxnet reaffirmed Iran's commitment to its nuclear program. While reports genuinely agreed that Stuxnet had effectively damaged 1,000 centrifuges in the Natanz facility, Iran had quickly recovered from the attack and replaced the effected centrifuges with new equipment, according to the Institute for Science and International Security, a Washington, D.C.-based non-partisan think

tank.[26] Indeed, current evidence clearly indicates that Iran has actually stepped up its nuclear development capabilities. According to the *Washington Post,* the next IAEA report on Iran's nuclear facility is not due until mid-November 2012, but as of the end of October, Iran had added more than 600 centrifuges to its underground facility at Fordow.[27] Therefore, it is clearly evident that the cyber attack – while minutely slowing Iran's uranium enrichment – did nothing to dissuade it from pursuing its nuclear development objectives. Three truths emerged from this situation: 1.) Stuxnet did not cause Iran to alter its plans; 2.) The deployment of a cyber weapon did not influence the Iranian government to cease its production of enriched uranium; and 3.) Stuxnet did not encourage the government to come to an arrangement with the United States and European Union.

As a potential policy tool, the cyber attack achieved two unexpected consequences: it bolstered Iranian commitment to nuclear development as the government rapidly replaced all damaged centrifuges,[28] and it revealed that if the United States was responsible for the attack it would militarize cyberspace to pursue its objectives. Furthermore, revelation of Stuxnet has since compelled Iran to improve its cyber security posture through a series of government-led mandates and regulations. Perhaps of more concern, Stuxnet has allowed Iranians to study a tool that is designed to target and adversely impact an industrial control system. Given the increased concern expressed by U.S. policymakers and military decision makers of hostile actors targeting the United States supervisory control and data acquisition (SCADA) systems, an escalation over this nuclear issue could prompt Iran to use a similar type tool to "try to retaliate by attacking U.S. infrastructure such as power grid, trains, airlines, and refineries."[29]

Furthermore, evidence suggests that stringent sanctions are doing more to influence Iran into providing more transparency to its nuclear development, than Stuxnet or any subsequent malware discovery on Iranian networks. Since taking effect, sanctions have successfully weakened Iran's economy, causing inflation and deflating the value of the rial, Iran's national currency.[30] While this has not caused Iran to give up its plans for nuclear development, it has been instrumental in changing its views over the possibility of sitting down with the United States to discuss alternatives and possibilities. In November 2012, Iran's Ministry of Intelligence published a report on its website highlighting both Israeli and U.S. positions on Iran's nuclear aspirations with a favorable view of the U.S. desire to resolve the matter diplomatically rather than by military force.[31] Simply, multilateral economic sanctions and diplomatic overtures have had more influence to bringing this topic to a peaceful resolution. The cyber attack, on the other hand, did not dissuade Iranians.

C. 2012 DISTRIBUTED DENIAL OF SERVICE ATTACKS AGAINST U.S. FINANCIAL SECTOR

Note: Although Iran has not claimed responsibility for this activity, it is assumed by U.S officials[32] that the government had at least an informal role in directing the distributed denial-of-service (DDoS) attacks against the U.S. financial sector. If the Iranian government was involved in directing or participating in the DDoS attack, and the activity was used as a foreign policy tool, then it can be interpreted as an Iranian effort to communicate to the United States that Iran had a formidable cyber capability, and to try to influence U.S. government courses of action toward the Islamic state.

From September-October 2012 an ongoing strategic DDoS campaign dubbed "Operation Ababil" was levied against several prominent institutions within the U.S. financial sector (Note: as of this writing, Phase 2 of Operation Ababil occurred in December 2012 and Phase 3 occurred in March 2013) . A self-described hacktivist group dubbed the "Cyber Fighters of Izz ad-Din Al Qassam," assumed credit for the attacks, claiming they were perpetrated in response to the anti-Islam film "Innocence of Muslims," which sparked worldwide controversy and physical protests. The targets of this sustained DDOS campaign were Bank of America, Wells Fargo, US Bank, JP Morgan Chase, Sun Trust, PNC Financial Services, Regions Financial, and Capital One. According to press reports, the attacks effectively cut bank customers off from online services for extended periods.[33]

1) Who Are The Cyber Fighters of Izz ad-Din Al Qassam?

Regardless of their public statements to the contrary, the Cyber Fighters of Izz ad-Din Al Qassam demonstrated little in common with traditional hacktivist groups such as Anonymous, LulzSec, or Anti-Sec. Additionally, while there have been instances of Islamic hackers unifying against a common foe (e.g., Israel during Operation Cast Lead), several facts suggest that this group was not composed of hacktivists at all but of more sophisticated individuals perhaps sponsored by or affiliated with a nation state. Several pieces of evidence support this line of thinking:

- Hacktivist groups typically target the subject of their ire; in this case, there were no cyber actions taken against Mark Basseley Youssef, the director of the controversial film, or anything related to him.
- There were no protest style cyber attacks directed against either Google or YouTube, the unwitting distributor of the film via its website. If the group believed Google to be complicit in posting the video, we could reasonably expect that the hacktivist group would have targeted Google. In this case, the Cyber Fighters of Izz ad-Din Al Qassam did not.

- The emergence of the Cyber Fighters of Izz ad-Din Al Qassam is suspect. While it's not uncommon for like minded individuals to quickly ban together under a common cause, there is no history of Islamic hacktivist groups demonstrating this type of capability. This group leveraged infections of high bandwidth servers as opposed to using participatory DDoS tools, which has generally been the case with Middle Eastern hacktivist groups. This suggests that these individuals had some affiliation with a nation state for training, technical support, and/or sponsorship.

The attackers used servers and customized malware, tailoring the campaign to get around defenses specifically designed to stop floods of data.[34] Given the technical savvy required to maintain a sustained DDOS attack against major financial networks, one would think that these individuals would be frequently engaged in Middle East disputes.

2) Mess With Us And We'll Mess With Your... Banks?

If cyber attacks are a possible government tool to support foreign policy objectives, two major questions need to be addressed in determining the utility of this new capability with respect to the 2012 DDoS campaign: (1) Why was the U.S. financial sector targeted with a sustained DDOS and (2) If the Iranian government had at least an informal role in directing the activity, did it achieve what it had set out to do?

U.S. officials and cyber experts have pondered the potential consequences of cyber attacks directed against the U.S. critical infrastructure and the damage it can cause due to the country's heavy reliance on computer networks and technology. In early 2012, President Obama identified cyber security as a national security priority alluding to the possible destruction of critical infrastructure networks as a real threat.[35] Indeed, even U.S. Secretary of Defense Leon Panetta warned of the possible ramifications of a cyber Pearl Harbor dismantling the nation's power grid, transportation system, and financial networks.[36] Suffice it to say, the U.S. Government has made it perfectly clear that it fears the possible consequences of such attacks against its well networked infrastructure.

Regardless of the motivations of the "alleged" hacktivist group behind the DDoS attacks, the targeting of the U.S. financial sector could be considered a retaliatory action for the Stuxnet incidents and U.S. sanctions levied against Iran for its continued nuclear development. For the former, Iran perceived the United States to be behind the cyber attack trying to destroy or at least disrupt the nuclear development process by infecting Iranian centrifuges.[37] For the latter, U.S. sanctions encouraged the Society for Worldwide Interbank Financial Telecommunication to block Iranian banks from using its service to conduct international banking transfers.[38]

Therefore, it can be argued that if Iran had some level of involvement in directing the DDoS activity, it was exercising a retaliatory strike.

The financial sector, as well as a nuclear industrial control system, is considered critical infrastructure, or networks essential for a functioning society. If the Government of Iran was involved in deploying the cyber weapon, they might have hoped to accomplish the following:

- Demonstrate its capability to target a critical infrastructure network using a technologically-based weapon instead of a more conventional one that would cause physical damage and potential loss of life;
- Retaliate with a measured response;
- Signal to the U. S. Government, as well as the region, that Iran has a cyber capability that can be deployed in a calculated manner against targets of its choosing.

While the attack did not cause any substantial damage to the targeted institutions, it did raise concern at the highest levels of the U.S. Government. Then U.S. Senator Joseph Lieberman in particular made public remarks attributing the activity to Tehran.[39] Although Iran never claimed official responsibility, it certainly made its intentions clear to the United States.

3) But Did It Work?

From an operational standpoint, the DDoS was an unqualified success. Banks were successfully targeted with a DDoS that took information sources offline or caused intermittent outages interrupting services. According to Prolexic Technologies, a company specializing in protecting organizations from DDoS attacks, sustained floods hitting 70 Gbps and more than 30 million packets per second were recorded in some of the attacks. When asked about the attacks, Dimitri Alperovitch, founder of CrowdStrike, said, "These banks… are not tiny. They have massive infrastructures…The fact that these attacks were able to shut down is quite remarkable." However, did Iran signal to the United States that it was a force to be reckoned with in cyberspace, perhaps a more subtle political objective of the government? Aside from signaling an Iranian cyber "show of strength," the DDoS attack failed to influence U.S. decision making or demonstrate to the U.S. government that Iran is a notable cyber force, based on the following:

- The United States did not alter its tough stance on Iran diplomatically nor did it repeal stringent economic sanctions levied against Iran.
- The United States withstood the most severe DDoS attack it has ever faced with relative ease without a prolonged hindrance to operations.

- Iran may have demonstrated the best it could do in a cyber attack capacity and the United States did not cower.

Therefore, the DDoS attacks did not prove to be a viable weapon of influence for the Iranian government if they were involved. It made no impact on U.S. plans and intentions toward Iran and its nuclear development, nor did it alter or amend its foreign policy positions. Viewing it from the narrow lens of foreign policy, the only conclusion that can be drawn was that the DDoS was a failure.

4. FUTURE CONSIDERATIONS

In October 2012, U.S. President Barack Obama signed a directive that enabled the military to act more aggressively against cyber attacks directed against the United States.[40] Lauded as a proactive step in countering the 21st Century's biggest threat, the new "policy" is intended as an equalizer to malicious online activity and a tool for military use. One U.S. defense official was quoted as saying, "cyberoperations… are an integral part of the coordinated national security effort that includes diplomatic, economic, and traditional military measures."[41] In short, the United States appears to be legalizing cyber attacks to be leveraged against those nations it perceives as a security threat without fully exploring the policy considerations that should accompany the deployment of cyber weapons as a policy tool.

Further complicating this scenario is how cyber weapons would be deployed against perceived nation state threats of varying cyber capability and/or information technology/Internet reliance. For a country like North Korea that has a near negligible Internet penetration rate, cyber attacks as a policy tool are almost futile. On the other end of the spectrum, take into account adversarial countries that have suspected and more robust cyber programs such as China and Russia. As one of the more wired countries in the world, and one whose officials routinely express concern about the security of its industrial control systems and critical infrastructures, is the United States prepared to potentially receive the same intensity of cyber attacks as it gives out? Finally, a third consideration addresses the identification of friendly and allied nations engaged in activities deemed a threat to national security such as the theft of sensitive and valuable military research and development, diplomatic, economic, and political information? Both France – a NATO member country – and Israel – the U.S.' strongest ally and a mutual defense treaty partner in the tumultuous Middle East region – have been identified as pervasive economic espionage actors against U.S. interests, according to the Central Intelligence Agency.[42][43] Would cyber attacks succeed in dissuading economic espionage, and is the cost-benefit worth the risk of breaking solid alliances.

5. CONCLUSION

Although the use of cyber attacks to support nation state foreign policy interests is still nascent at best, early indications clearly show it to be unsuccessful at influencing decision makers or their courses of action, and therefore is not an effective policy tool. Several factors account for this. First and most notably, despite the advanced cyber weaponry capabilities demonstrated by Stuxnet, Duqu, and Flame, there has yet to be that one "jaw dropping" effect that makes individuals think twice before booting up their laptops with malicious intent. Second, the threat of offensive cyber operations has a relative limited target base. For example, the threat of unleashing a sophisticated cyber weapon may carry more weight with a "wired" country like China or Russia than North Korea or even Iran, and has even less menace to nonstate actors that do not have a fixed infrastructure from which they operate. Finally, the deployment of cyber weaponry runs the risk of quickly and unnecessarily escalating a situation, particularly if cyber actions are misunderstood or misinterpreted by a clever adversary seeking to divert blame onto a third country.

REFERENCES

[1] John Brodkin; November 29, 2007; "Government-sponsored cyberattacks on the rise, McAfee says;"Network World; McAfee; http://www.networkworld.com/news/2007/112907-government-cyberattacks.html

[2] Ellen Nakashima, Greg Miller, and Julie Tate; June 19, 2012; "U.S. Israel Developed Flame Computer Virus to Slow Iranian Nuclear Efforts, Officials Say;" The Washington Post; http://www.washingtonpost.com/world/national-security/us-israel-developed-computer-virus-to-slow-iranian-nuclear-efforts-officials-say/2012/06/19/gJQA6xBPoV_story.html.

[3] The White House, International Strategy for Cyberspace; May 2011; http://www.whitehouse.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

[4] U.S. Department of Defense; July 2011; Department of Defense's Strategy for Operating in Cyberspace; http://www.defense.gov/news/d20110714cyber.pdf

[5] U.S. Department of Defense, Cyber Command Fact Sheet, 21 May 2010 http://www.stratcom.mil/factsheets/Cyber_Command/

[6] Gadi Evron, Winer/Spring 2008, "Battling Botnets and Online Mobs: Estonia's Defense Efforts During the Internet War," Georgetown Journal of International Affairs, http://journal.georgetown.edu/wp-content/uploads/9.1-Evron.pdf

[7] Eneken Tikk, Kadri Kaska, and Liis Vihul, 2010, "International Cyber Incidents: Legal Considerations," Cooperative Cyber Defence Centre of Excellence, http://www.ccdcoe.org/publications/books/legalconsiderations.pdf

[8] William C. Ashmore; 2009; “Impact of Alleged Russian Cyber Attacks;” Baltic Security & Defence Review; Volume 11, 2009;

[9] Stephen Herzog; 2011; “Revisiting the Estonian Cyber Attacks: Digital Threats and Multinational Responses; Journal of Strategic Study; Volume IV, Issue 2; pp.49-60.

[10] Peter Finn, May 9, 2007, “For Estonia’s Ethnic Russians, Ties to Moscow Fading,” Washington Post Online, http://www.washingtonpost.com/wp-dyn/content/article/2007/05/08/AR2007050801935.html

[11] Irina Borogan and Andrewi Soldatov, April 25, 2012, “The Kremlin and the Hackers: Partners in Crime?” Open Democracy, http://www.opendemocracy.net/od-russia/irina-borogan-andrei-soldatov/kremlin-and-hackers-partners-in-crime

[12] Luke Harding; April 27, 2007; “Russia Up in Arms After Estonians Remove Statue of Soviet Soldier;” The Guardian; http://www.guardian.co.uk/world/2007/apr/28/russia.lukeharding/print

[13] Noah Shachtman; March 11, 2009; “Kremlin Kids: We Launched the Estonia Cyber War;” Wired.com; http://www.wired.com/dangerroom/2009/03/pro-kremlin-gro/

[14] Russian Law; March 6, 2009; “Russian Deputy Admits Involvement in Cyber Attacks on Estonia;” Russian Law Online; http://russian-law.livejournal.com/24179.html

[15] Joshua Davis; August 21, 2007; “Hackers Take Down Most Wired Country in Europe;” Wired Magazine; http://www.wired.com/politics/security/magazine/15-09/ff_estonia?currentPage=all

[16] Europe Edition; May 10, 2007; “A Cyber Riot;” The Economist Online; http://www.economist.com/node/9163598

[17] David E. Sanger, June 1, 2012, “Obama Order Sped Up Wave of Cyber Attacks Against Iran,” New York Times, http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html?pagewanted=all&_r=0

[18] Matthew Schwartz, June 1, 2012, “Stuxnet Launched by United States and Israel,” Information Week, http://www.reuters.com/article/2011/12/02/us-cyberattack-iran-idUSTRE7B10AV20111202

[19] Ellen Nakashima, Greg Miller, Julie Tate; June 19, 2012; “U.S. Israel Developed Flame Computer Virus to Slow Iranian Nuclear Efforts, Officials Say;” The Washington Post; http://www.washingtonpost.com/world/national-security/us-israel-developed-computer-virus-to-slow-iranian-nuclear-efforts-officials-say/2012/06/19/gJQA6xBPoV_story.html.

[20] David E. Sanger, June 1, 2012, “Obama Order Sped Up Wave of Cyber Attacks Against Iran,” New York Times, http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html?pagewanted=all&_r=0

[21] David E. Sanger, June 1, 2012, “Obama Order Sped Up Wave of Cyber Attacks Against Iran,” New York Times, http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html?pagewanted=all&_r=0

[22] CNN Wire Staff; March 19, 2012; “Timeline of Iran’s Controversial Nuclear Program;” CNN.com; http://www.cnn.com/2012/03/06/world/meast/iran-timeline/index.html.

[23] CNN Wire Staff; March 19, 2012; "Timeline of Iran's Controversial Nuclear Program;" CNN.com; http://www.cnn.com/2012/03/06/world/meast/iran-timeline/index.html.

[24] CNN Wire Staff; March 19, 2012; "Timeline of Iran's Controversial Nuclear Program;" CNN.com; http://www.cnn.com/2012/03/06/world/meast/iran-timeline/index.html.

[25] CNN Wire Staff; March 19, 2012; "Timeline of Iran's Controversial Nuclear Program;" CNN.com; http://www.cnn.com/2012/03/06/world/meast/iran-timeline/index.html.

[26] Joby Warrick; February 16, 2011; "Iran's Nuclear Natanz Facility Recovered Quickly From Stuxnet Cyber Attack;" The Washington Post Online; http://www.washingtonpost.com/wp-dyn/content/article/2011/02/15/AR2011021505395.html

[27] Walter Pincus; November 1, 2012; "All Politics Are Local In Iran Nuclear Dispute;" The Washington Post; http://articles.washingtonpost.com/2012-10-31/world/35501296_1_nuclear-facilities-iaea-report-nuclear-program

[28] Mark Clayton, January 3, 2011, "Stuxnet Attack on Iran Nuclear Program Came About A Year Ago, Report Says," Christian Science Monitor, http://www.csmonitor.com/USA/2011/0103/Stuxnet-attack-on-Iran-nuclear-program-came-about-a-year-ago-report-says

[29] Brian Ross; March 5, 2012; "What Happens to the U.S. If Israel Attacks Iran?:" ABC News online; http://abcnews.go.com/Blotter/israel-attacks-iran-gas-prices-cyberwar-terror-threat/story?id=15848522#.UJqwOIZCjZ4

[30] Glenn Greenwald, October 7, 2012, "Iran Sanctions Now Causing Food Insecurity and Mass Suffering," The Guardian, http://www.guardian.co.uk/commentisfree/2012/oct/07/iran-santions-suffering

[31] Jason Rezaian, November 7, 2012, "Iran Ministry Suggests Openness to Nuclear Talks," The Washington Post, http://articles.washingtonpost.com/2012-11-07/world/35504040_1_nuclear-program-nuclear-talks-nuclear-facilities

[32] Ellen Nakashima, September 21, 2012, "Iran Blamed for Cyber Attacks on US Banks and Companies," The Washington Post, http://articles.washingtonpost.com/2012-09-21/world/35497878_1_web-sites-quds-force-cyberattacks

[33] Ellen Messmer; October 24, 2012; "DDoS Attacks Against Banks Raise Question: Is this Cyber War?" Network World online; http://www.networkworld.com/news/2012/102412-bank-attacks-cyberwar-263664.html

[34] Robert Lemos; October 4, 2012; "Serious Attackers Paired with Online Mob in Bank Attacks;" Dark Reading Online; http://www.darkreading.com/advanced-threats/167901091/security/perimeter-security/240008534/serious-attackers-paired-with-online-mob-in-bank-attacks.html

[35] Manuel Flores; July 25, 2012; "Obama Makes Cybersecurity a Priority;" Independent Voters Network Online; http://ivn.us/2012/07/25/obama-makes-cybersecurity-a-priority/

[36] Elizabeth Bumiller; October 11, 2012; "Panetta Warns of Dire Cyber Attack on U.S.;" The New York Times Online; http://www.nytimes.com/2012/10/12/world/panetta-warns-of-dire-threat-of-cyberattack.html?pagewanted=all

[37] Associated Press, April 16, 2011, "Iran Blames U.S., Israel for Stuxnet Malware," CBS News, http://www.cbsnews.com/2100-202_162-20054574.html

[38] Don Melvin and Jonathan Fahey; March 15, 2012; "SWIFT Cuts Off as Sanctions Vice Tightens; "The Huffington Post; http://www.huffingtonpost.com/2012/03/15/swift-iran-sanctions_n_1347361.html

[39] David Goldman; September 28, 2012; "Major Banks Hit With the Biggest Cyber Attack in History;" CNN.com; http://money.cnn.com/2012/09/27/technology/bank-cyberattacks/index.html

[40] Ellen Nakashima; November 14, 2012; "Obama Signs Secret Directive to Help Thwart Cyberattacks;" Washington Post Online; http://www.washingtonpost.com/world/national-security/obama-signs-secret-cybersecurity-directive-allowing-more-aggressive-military-role/2012/11/14/7bf51512-2cde-11e2-9ac2-1c61452669c3_story.html

[41] Ellen Nakashima; November 14, 2012; "Obama Signs Secret Directive to Help Thwart Cyberattacks;" Washington Post Online; http://www.washingtonpost.com/world/national-security/obama-signs-secret-cybersecurity-directive-allowing-more-aggressive-military-role/2012/11/14/7bf51512-2cde-11e2-9ac2-1c61452669c3_story.html

[42] From Reuters; August 15, 1996; "France, Israel Cited in CIA Espionage Study;" Los Angeles Times Online; http://articles.latimes.com/print/1996-08-15/business/fi-34524_1_economic-espionage

[43] John A. Nolan; October 2000; "A Case Study on French Espionage: Renaissance Software;" http://www.hanford.gov/files.cfm/frenchesp.pdf

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

The Future of Military Virtue: Autonomous Systems and the Moral Deskilling of the Military

Shannon Vallor
Department of Philosophy
Santa Clara University
Santa Clara, California USA
svallor@scu.edu

Abstract: Autonomous systems, including unmanned aerial vehicles (UAVs), anti-munitions systems, armed robots, cyber attack and cyber defense systems, are projected to become the centerpiece of 21st century military and counter-terrorism operations. This trend has challenged legal experts, policymakers and military ethicists to make sense of these developments within existing normative frameworks of international law and just war theory. This paper highlights a different yet equally profound ethical challenge: understanding how this trend may lead to a *moral deskilling* of the military profession, potentially destabilizing traditional norms of military virtue and their power to motivate ethical restraint in the conduct of war. Employing the normative framework of virtue ethics, I argue that professional ideals of military virtue such as courage, integrity, honor and compassion help to distinguish legitimate uses of military force from amoral, criminal or mercenary violence, while also preserving the conception of moral community needed to secure a meaningful peace in war's aftermath. The cultivation of these virtues in a human being, however, presupposes repeated practice and development of skills of moral analysis, deliberation and action, especially in the ethical use of force. As in the historical deskilling of other professions, human practices critical to cultivating these skills can be made redundant by autonomous or semi-autonomous machines, with a resulting devaluation and/or loss of these skills and the virtues they facilitate. This paper explores the circumstances under which automated methods of warfare, including automated weapons and cyber systems, could lead to a dangerous 'moral deskilling' of the military profession. I point out that this deskilling remains a significant risk even with a commitment to 'human on the loop' protocols. I conclude by summarizing the potentially deleterious consequences of such an outcome, and reflecting on possible strategies for its prevention.

Keywords: *automated methods, ethics, military virtue, professionalism, moral deskilling.*

1. INTRODUCTION

Few images highlight the increasingly automated nature of modern warfare better than a photograph of the eerily opaque, windowless nose of the MQ-1 Predator drone, manufactured by General Atomics Aeronautical Systems and a centerpiece of U.S. military and counterterrorism efforts in the Middle East and Africa, where hundreds of targeted drone strikes are now launched annually. Yet drone warfare is merely the leading edge of a broader worldwide trend toward more autonomous methods of warfighting. From South Korea's armed sentry robots guarding the DMZ, to Israel's 'Iron Dome' anti-munitions defense, to miniaturized lethal drones like the U.S. Army's Switchblade, to long-range intercontinental drones like the U.K. Taranis and the U.S. X47-B, militaries around the world are investing in an increasingly automated future. Nor are such investments limited to weapons in the conventional sense. Military and intelligence agencies worldwide are developing increasingly sophisticated and autonomous software algorithms for use in cyberwarfare – conflicts between electronic agents in electronic space that nevertheless have the potential to inflict considerable human losses. Merging with both trends are advancements in algorithms for analysing massive datasets, which can potentially outperform human calculations of threat potential, target value, operational risk, mission cost, casualty estimates and other key strategic variables. Taken together, these developments represent a profound shift in our traditional understanding of the role of human beings in the conduct of war. In this paper I explore one of this shift's most challenging ethical implications, namely, the risk of a significant *moral deskilling* of professional militaries, and a destabilization of associated ideals of military *virtue*.

The broader legal and ethical implications of this shift are immense in scope; they range from the compliance or non-compliance of automated warfighting systems with the laws of war and requirements of just war theory (Asaro 2008), to problems of moral and legal accountability for actions taken by autonomous agents (Sparrow 2007), to the concern that automated methods of warfare are leading to greater 'moral disengagement' among soldiers (Sharkey 2010; Royakkers and van Est 2010). Together, these concerns mandate extensive and widespread critical inquiry and reflection on the automation of war; fortunately, this conversation is now well underway. In addition to scholarly articles by ethicists and legal experts, recent years have seen several high-profile books on related topics (Singer 2009; Arkin 2009; Krishnan 2009; Lin et. al. 2012). Major media outlets from *The New York Times* and *The Wall Street Journal* to online magazines like *Wired* and *Slate* regularly cover emerging developments in automated war technology and their political, legal and ethical ramifications. We are seeing the welcome emergence of a vigorous scholarly and public discourse on the legality and ethics of automated warfare, one likely to

continue to evolve for decades as the possibilities, risks and benefits of automated systems become clearer.

Yet one important subset of these concerns is likely to be less visible to public and political interests than the legalistic and utilitarian dilemmas presently driving the global conversation about the ethics of automated warfare. The subject to which I wish to call attention concerns the future of military virtue in an era of increasingly automated military action. My claim is that unless we take steps to secure that future, we face the possibility of a dangerous *moral deskilling* of the military profession. In what follows, I call the reader's attention to the importance of habitual moral practice and skill acquisition as a precondition for the cultivation of military virtues, which are in turn critical to the standing of militaries as professional bodies with a distinct moral status recognized by combatants and civilians alike. I argue that by depriving soldiers of the opportunity to practice and develop the skills critical to the ethical use of military aggression, increasingly automated methods of warfighting could endanger the moral and professional status of the military vocation. In my conclusion I offer some speculations about how this outcome might be prevented.

2. MORAL PRACTICE AND MILITARY VIRTUE

Before I develop and support my claims, let me briefly explain what 'virtue' in the phrase 'military virtue' entails. The concept of virtue is rooted in classical traditions going as far back as the ancient Greek philosophies of Plato and Aristotle and, in the East, Confucian and Buddhist ethics. It endures today in the writings of contemporary virtue ethicists like Rosalind Hursthouse, Alasdair MacIntyre and Martha Nussbaum, and has found its way into various applied and professional codes of ethics, including business ethics, medical ethics, environmental and engineering ethics (Axtell and Olson 2012). Virtues are habituated states of a person's character that reliably dispose their holders to excel in specific contexts of action, and to live well generally; so moral virtues are states of character that, once acquired, dispose their possessors to perform excellent moral actions of particular sorts, and, more broadly, to excel in moral living. Cardinal examples of moral virtues include wisdom, honesty, courage and moderation; others commonly recognized include loyalty, integrity, respect, honor, patience, compassion and benevolence, though this list is far from exhaustive. How particular virtues are defined and prioritized varies among cultures, historical periods and social roles; yet there is substantial overlap or convergence among diverse virtue traditions, indicating that the qualities seen as most supportive of human flourishing are, while not entirely universal, rooted in widely shared or similar human practices.

Because virtues are habituated rather than inborn, whether or not a person develops

a particular virtue will largely depend on whether they engage repeatedly in the kinds of practices that cultivate it. The virtue of honesty, for example, can only be acquired through repeated practice of truth-telling. Initially, such practice requires guidance by a virtuous model, e.g., someone who is already honest. Over time, repeated practice can lead a person to see for themselves what honesty is, to see it as good in itself and to embody it better and more easily; a person who has cultivated the virtue of honesty is not only consistently inclined to tell the truth, they have learned how to excel at truth-telling in any situation that might arise: who to tell the truth to, when and where, in what way, and to what extent. Moral virtue thus requires more than good will and a steady desire to do the right thing – it requires the cultivation of a kind of *practical wisdom* that directs this right desire intelligently, perceiving and quickly adapting to the unique moral demands of each situation. In his *Nicomachean Ethics*, Aristotle named this practical wisdom *phronesis*; a sort of 'über-virtue' that orchestrates one's individual qualities of moral excellence and integrates them within a complete and flourishing life (1984, 1140a25-30;1145a). The concept of virtue, then, picks out those aspects of persons that enable them to live as moral exemplars for others, qualities of character that we ourselves can strive to cultivate through the same sorts of repeated practice.

The role of virtue in military ethics has long been recognized, and a rich body of existing literature details the way in which virtues like courage, duty, integrity, honor, loyalty and service have historically been inseparable from the ideal of the good soldier (Olsthoorn 2011; Robinson 2007; Reichberg, Syse and Begby 2006; French 2003; Toner 2000). This does not mean that the enterprise of war itself can or should be seen as virtuous. Rather, ideals of virtuous military character, when exercised as normative *expectations* (not just indicators of supererogatory or heroic performance), express a society's unwillingness to wholly exclude its warfighters from the broader responsibilities and benefits of the moral community.[1] As I argue elsewhere (Vallor 2013), ideals of military virtue, when embedded in the practice and professional identity of military bodies, block the cultural displacement of war to an extra-moral realm where its conduct would be indistinguishable from criminal or mercenary violence.

The ideal serves as a kind of contract between warfighters and the larger community, and when in force, it offers considerable benefits to soldiers and civilians alike. In addition to motivating restraint on the part of soldiers in inflicting civilian harms, it can motivate limited restraint between enemy combatants when each recognizes the other as a professional fighting with honor and moral purpose. It can also support

[1] The term 'moral community' here is left deliberately ambiguous in its reference; ideals of military virtue can be seen as ties that bind a soldier to the ethical life of her own nation or culture, or, in more cosmopolitan views, to the ethical life of the global human community of which she is a part.

the psychological integrity of soldiers themselves, by providing a moral context for what are, taken in themselves, brutal and deplorable actions. Finally, it preserves the sense of moral community between warfighters and civilians that allows returning soldiers to be welcomed home, and even valorized. To see the importance of this contingency, one need only be familiar with the starkly different experiences in the United States of veterans of World War II, treated to grand welcoming parades and to this day labelled "The Greatest Generation," and veterans of the Vietnam War, who returned home to a largely indifferent and often hostile society no longer able to contextualize their service as virtuous.

Thus while war itself cannot be virtuous, because it characteristically impedes rather than supports human flourishing, humans who take on the burdens of military service *can* be - insofar as they aspire to fight only in the manner of an excellent human being. Of course, moral virtue is expressed differently according to the demands of particular circumstances; what compassion and courage call for in battle looks very different from what these demand in civic life. In writing on war and its apparent incompatibility with moral norms, Augustine wrote that precisely because war foments evil ("the desire for harming, the cruelty of revenge, the restless and implacable mind, the savageness of revolting, the lust for dominating"), it is all the more essential that soldiers cultivate the virtuous dispositions of compassion and benevolence to accompany them in battle, so that the "mutual bond of piety and justice" that constitutes common morality has not been irrevocably destroyed by the time that material conditions for peace return (Augustine 1994, 221-222).

Military virtue, then, imperfect as its professional cultivation and practice may be, functions to keep warfighters morally continuous with society. It allows us to see ourselves, and the other, as *worthy* of membership in a moral community, even when engaged in conduct that is in itself destructive to moral community. When the professional cultivation of military virtue is not attempted or its aspirations are abandoned, as in Cambodia, Rwanda, and Srebrenica, the aftermath of war is often precisely what we would expect from Augustine's account: a shallow peace poisoned by deep distrust, resentment and fear lasting for generations. Survivors of such a moral calamity do not stop suffering when the bloodshed stops: they are burdened with the crippling social degradation that comes from the death of civic norms of trust, mercy, forgiveness, justice and goodwill. Such norms, once destroyed, are not easily reborn; while military virtue may not be able to shield them from assault, it can keep them on life support. For all of these reasons, then, it is essential to the mitigation of the harms of war that military virtue be preserved, both as a meaningful moral concept and as a practical and attainable commitment to ethical warfighting. In what follows I explain why the increasing automation of warfighting methods may jeopardize this imperative.

3. AUTOMATED WARFARE, VIRTUE AND THE MORAL DESKILLING OF MILITARY PRACTICE

Having offered reasons to take the concept of military virtue seriously, I turn to the primary burden of my argument: to show how increasingly automated methods of warfighting challenge the future of military virtue by potentially contributing to a *moral deskilling* of the military profession.

First, let us consider the link between virtues and skills. Aristotle was clear that virtues and skills share many common features – both are acquired by habit and practice, both must be guided by intelligence, and both must be adapted to the demands of given situations. But he also reminds us that virtue is *more* than just skill or know-how; it is a state in which that know-how is reliably put into action when called for, and is done with the appropriate moral concern for what is good: "The agent also must be in a certain condition when he does [virtuous acts]; in the first place he must have knowledge, secondly he must choose the acts, and choose them for their own sakes and thirdly his actions must proceed from a firm and unchangeable character" (1984, 1105a30-35). Someone could have moral skills in the sense of practical moral knowledge but fail to be virtuous because they are unreliable in acting upon this knowledge, or because they act well only for non-moral reasons. Still, moral skills are a *necessary* if not a sufficient condition for moral virtue. Without the requisite cultivation of moral knowledge and skill, even a person who sincerely wishes to do well consistently and for its own sake will be unsuccessful. It follows that if the advancing automation of military conflict were to bring about a significant 'moral deskilling' of the profession, the future of military virtue would be gravely endangered.

What would a 'moral deskilling' of the military profession amount to, and how might the advancing automation of warfighting systems contribute to it? *Deskilling* is a familiar concept in the analysis of the social impact of technology; for example, we might think of the way in which the skills of machinists and other classes of mechanical labor were devalued by widespread factory adoption of automated machine tools (Braverman 1974). Or consider the worry that the professional work of highly skilled nurses is increasingly given over to a combination of less skilled aides and advanced medical monitoring and medication delivery technologies (Rinard 1996). However, the concept of deskilling has declined in academic usage in the last few decades, in large part because unlike the earlier automation of factory work, the information revolution has thus far seemed to deliver as much *up*skilling as deskilling –workers in many industries have been freed by computers to shift their duties from mindless tasks like filing, copying and collating to more

challenging and knowledge-laden responsibilities. Yet some new applications in information technology may warrant renewed concerns about deskilling, including moral deskilling (Manders-Huits 2006). Whether any automated technology produces deskilling, then, is an empirical question that depends upon the particular context of use. Let us look more closely at the critical meaning of the concept, and how it may apply to the context of automated warfare.

The concept of deskilling has at least two critical implications. The first and most commonly discussed implication is that the deskilling of a given profession may decrease the socioeconomic value, autonomy and power enjoyed by workers, potentially causing them significant psychological and economic harm. A second critical implication, the one I wish to highlight, is that at least for some professions, we may have reason to regret the loss of the professional context for cultivating the given skills because we think the skills themselves are intrinsically valuable. For example, many have mourned the declining skills of artistic handicraft lost to mass manufacture of ready-made objects (Roberts 2010), resulting in renewed interest in 'handmade,' 'custom' or 'artisanal' products. In this context, it is not only the economic welfare of the artisan that we value, and not only the quality of the end product, but also the connection between an artifact and a human whose artistic excellence and knowledge was responsible for its production. We think that it is *good* that humans are skilled at making beautiful and useful objects for their own living, and that even if machines could produce all such goods for us, it would be sad and regrettable if humans were no longer capable of doing the same.

I suggest that the intrinsic value of artisanal skills is not only paralleled but dwarfed by the intrinsic value of moral skills. The concept of *moral* deskilling is only rarely employed used in the sociological literature on technology, in part because sociologists tend to shy away from normative judgments of ethics, and in part because concerns about moral deskilling are sometimes associated with reactionary 'moral panics' in reaction to technological change – for example, worries in the 1920's that the telephone would result in crippling social isolation and the unravelling of people's capacities for moral interaction. However, the concept remains meaningful, and I suggest that it may have profound significance with respect to the professional impact of military automation. While deskilling has been recognized with respect to the threatened obsolescence of abilities such as those cultivated by military snipers (Townsend and Charles 2008), the more worrisome possibility of a *moral* deskilling of the military profession has yet to be widely acknowledged.

Consider the parallel drawn earlier with artisanal skills. Just as the widespread loss of such skills by humans would not be fully expunged by machines that produce comparable products, a widespread loss of moral skills in the context of military

conduct would not be rendered insignificant by the emergence of machines that produce equivalent, or arguably even better outcomes. This fact has unfortunately been lost in the otherwise rich debate about the legal and ethical implications of automated warfare. A world in which humans involved in warfighting are no longer skilled in the *moral* conduct of war is a world in which the concept of 'military virtue' has no meaning. As I have argued elsewhere (2013), where this concept has lost its meaning, the recognition of soldiers as professionals devoted to the selfless service of the moral community is no longer possible.

A. AUTOMATED WEAPONS SYSTEMS AND MORAL DESKILLING

Methods of automated warfare may be divided between those involving cyber-conflict, and those involving (directly or indirectly) the application of military force. Let's first consider the latter, starting with the ongoing debate about autonomous and semi-autonomous weapons. Most of the literature on this subject has focused on the inability of weaponized robots and drone aircraft to act with the moral knowledge, restraint, compassion, discrimination, proportionality and accountability demanded by modern laws of war. A recent Human Rights Watch report on the topic states that their primary concern is the price civilians will pay for these inevitable ethical shortcomings of autonomous weapons systems (Human Rights Watch 2012, 1). But not only is this open to challenge from those with more optimistic projections for artificial moral intelligence (Arkin 2009), it remains silent on the human cost of no longer asking soldiers to cultivate and reliably exercise the same moral capacities. If the optimistic predictions of roboticists are anywhere near correct, we may be moving toward a future where humans start wars, oversee them, and suffer from them, but are no longer *fighting* them, in the concrete sense of making informed and morally reflective choices about who or what gets targeted, or when, in which circumstances, or with what degree of force. My claim is that there is a price to pay here *even if* civilians do not suffer more direct harm as a result.

Consider that the skill set for supervising, approving or vetoing the decisions of semi-autonomous robots seeking to apply lethal force will be much narrower than the skill set required for humans to make those moral decisions themselves. One reason involves the time constraints under which human supervisors of autonomous or semi-autonomous weapons will operate. Many scholars believe that the much-touted principle of humans staying 'on the loop,' with veto power over system targeting or firing actions, will soon be rendered largely meaningless when the human operator is given only a fraction of a second to make the veto decision, as is the case with several systems already in operation (Human Rights Watch 2012). One of the key tactical advantages of autonomous weapons systems is that they can make

and execute decisions far faster than humans can. These narrowing time horizons will likely preclude human operators from conducting substantive investigation of, or careful reflection upon, the morally salient features on the ground warranting the robotic application of force. This has already been acknowledged as a fundamental technical obstacle to humans remaining on the loop of engagements between unmanned combat fighters; the delay time injected by satellite communications is simply incompatible with the timescale of air combat (Sharkey 2012).

Add to this the likelihood that human operators of semi-autonomous systems will be tasked with supervising dozens or hundreds of drone or robotic agents at one time, as described by the Swarms.org website for the U.S. military's SWARMS initiative (Scalable sWarms of Autonomous Robots and Mobile Sensors), and the potential for moral deskilling becomes even more evident. We might be tempted to envision human supervisors of autonomous or semi-autonomous weapons systems as elite military judges chosen for their Solomonic wisdom and discretion in the ethical use of lethal force; but in reality they may have even less room for discretion and fewer degrees of decision freedom than air traffic controllers. What sort of moral skill set can we reasonably expect such practices to cultivate? And if moral skills in the use of military force are not cultivated at the level where force is applied, or even at the level where its application is being directly supervised, where *will* it be cultivated, and through what practices?

That advances in automation will result in revolutionary shifts in the skills needed for modern warfighting is news to hardly anyone. In envisioning a future where thanks to advancing automation, "systems and equipment can deploy forward with little if any human presence unless required for acceptance," the U.S. Air Force's published "Flight Plan 2009-2047" for unmanned aerial systems acknowledges that "a key challenge to realizing the vision will be to develop and maintain the right skill sets of systems and operational software developers, mission directors and USAF leaders…leaders will also require different skills to employ air power that is largely non-human" (USAF 2009, 51). But nowhere in this plan is it acknowledged that military leaders traditionally are expected to exercise *moral* as well as technical and strategic skill in the use of weaponized systems; it is worth asking *where* those moral skills will be cultivated in a future of automated warfighting where "relatively few mission directors will be needed" and the skills needed to "prepare, launch and perform" combat operations have been shifted from the field of action to "technology development offices" (*Ibid*.).

One might object by pointing out that decisions to use military force are very rarely conducted under conditions conducive to deep moral reflection. Human soldiers already have to make snap judgments in the field under highly demanding constraints, and even those decisions that can be reviewed by commanders are

rarely evaluated under ideal conditions. Yet it remains the case that a remarkable amount of moral knowledge and skill is presupposed by the human ability to keep a military operation involving lethal force from descending into utter moral chaos. Not every commanding officer has this kind of knowledge and skill, or even adequate exposure to the practices needed for its cultivation; but if *no one* in the chain of command has it, the chances of moral catastrophe are greatly increased. Neither sound rules of engagement nor advance commitments to 'human values' will prevent disaster if there is no one who can *apply* them in morally expert ways.

Indeed, moral virtue entails precisely the kind of expertise that allows us to quickly perceive the right course of action even in unpredictable or rapidly changing circumstances, without laborious calculations or clumsy recourse to formal principles. Yet according to many roboticists, a chief advantage of future autonomous weapons systems is that once programmed, they will not need human experts to tell them how to avoid morally catastrophic uses of force, begging the question of whether human soldiers will still be expected to cultivate that expertise for themselves. Remember that repeated moral practice is essential to the cultivation of moral virtue. How might our moral development suffer from transferring the most critical of those practices to machines, whose response times and cognitive architecture will be sufficiently unlike our own to prevent them from serving as models of virtue for us?

Furthermore, if the advance of autonomous weaponry were to lead to a significant 'moral deskilling' of the military profession, how would that impact the cultivation of military virtue, which as I claimed earlier, performs a critical function in mitigating the tendency of wars to produce lasting civic devastation? For the sake of argument, let us assume with roboticists like Ron Arkin (2009) that the most optimistic predictions regarding the emergence of artificial moral intelligence will be realized, and that in the not so distant future, human soldiers are no longer regularly called upon to judge when lethal military force is warranted and when it is not, or how it should be applied. We have handed over these judgments to robotic systems without any of our defective dispositions to anger, vengeance, bias, fear and laziness, and with computational abilities that keep their margins of error well below the best-trained of human soldiers. Without opportunities to exercise the skilled moral judgments that the expert application of lethal force requires, what level of 'moral deskilling' of professional soldiers may result, and with what consequences for the cultivation of military courage, honor or compassion?

One might interrupt to remind me that the United States has issued a new policy directive ensuring that, as the title of *Wired* magazine's coverage ably summed up, 'A Human Will Always Decide When a Robot Kills You' (Ackerman 2012). This new commitment to reject fully autonomous targeting and lethal engagement should not preclude us from seriously entertaining our thought experiment for

three reasons. First, this directive expires in ten years (Carter 2012), well before most scholars expect reliable technology for fully autonomous lethal robots to be available. Second, it binds only U.S. armed forces, and in no way precludes the development of such systems for other markets. Third, it does not change the above-noted fact that even semi-autonomous weaponry is rapidly shrinking the window for decisions on target selection and engagement below the timescale of human decision-making. Add to this the recognized 'automation bias' that leads humans to trust computer judgments over their own (Asaro 2009, 22), and the United States' promise to preserve human control over lethal means of warfighting may seem less meaningful. Well, then, so much the worse for underperforming, unpredictable and irrational human soldiers, and so much the better for programmable, precise and obedient killer robots, say roboticists like Arkin (2010). They may be right, and from a consequentialist point of view, there is no question that reductions in civilian casualties as a result of increased precision and reduced error rates, if realized, will have to factor into any moral assessment of the use of autonomous lethal weapons.

But the consequentialist equation cannot be the whole story. We must also consider the value of the moral skills that make military virtue and professionalism possible and what their loss might mean for professional soldiers and civilians alike. Before offering some concluding thoughts on how their loss might be prevented, let us move beyond automated weapons systems of the traditional sort, and extend our inquiry to automated methods of cyber-conflict.

B. *BEYOND AUTONOMOUS WEAPONS: ALGORITHMIC AUTONOMY, CYBER-CONFLICT AND MORAL DESKILLING*

The role of automated systems in warfighting is not limited to drones, robots or autonomous defense munitions. The software algorithms that enable autonomous or semi-autonomous operation of such systems can also be used to automate military or intelligence decision processes that may or may not involve the deployment of autonomous weapons. Consider, for example, a recent paper on an algorithm developed at West Point's Network Science Center and funded by the U.S. Army Research Office for potential use in ranking the most valuable targets in a terrorist network (Shakarian et. al. 2012). The paper's authors suggest that the performance of the algorithm, which tends to select mid-level lieutenants in a terrorist network as more valuable targets than high-level commanders, may be superior to independent human assessments of target value. Granting targeting authority to such an algorithm could lead to an operation involving an automated drone strike on the target, but it could also motivate a Special Forces assault or attempt at capture. Yet even without the use of automated weapons, such an operation would embody the trend toward automated warfighting. The moral implications of letting a computer program decide which individual humans deserve to be military targets are starkly

apparent, but again, setting aside obvious worries about the *justice* of automating such a decision, consider its additional implications for human cultivation of moral skills, knowledge and virtue. Imagine that we come to rely upon algorithms of this kind for military and intelligence targeting, but also for determining, for example, whether killing or capturing a particular target is more ethical and prudent, and the best operational design, occasion or ordnance for doing so. What kind of moral character would be required for military officers and other personnel to successfully support such an operation? Would any moral skills or qualities of note be required? If so, what would they be? In what actions would they be cultivated, or exercised? Or would soldiers and mission leaders be called upon strictly as technical specialists, tasked and trusted with nothing more than ensuring informational integrity in the communication of algorithmic decisions down the chain of command?

Such questions clearly extend into the realm of military operations involving no direct deployment of force whatsoever, such as cyber-warfare, or more broadly, cyber-conflict. Consider an algorithm that is programmed to defend government networks from intrusion, and to launch a counter-attack upon any electronic system or network it identifies as the host of the intruding informational agent. Set aside for now the technical questions about how to effectively design such an algorithm, such that it does not frequently mistake benign interactions for a cyberattack, or misidentify the agent responsible for an attack. To run parallel with our thought experiment about autonomous robots and other weapons systems, let us assume, just for the sake of argument, that we will soon develop sufficiently advanced artificial intelligence such that we can trust such algorithms to select, as justice requires, a proportional and discriminating response to any given cyberattack. What skills and virtues would be required of the human operators and supervisors of a cyberdefense system driven by such algorithms?

Let us say that we adopt the policy that such systems must maintain a human 'on the loop,' who in each case is tasked with approving or rejecting the system's request to launch a counterattack. Even setting aside the 'automation bias' mentioned above, which can already predispose us to defer to computer decisions (Asaro 2009), how would such a supervisor ever become qualified to make that judgment, in a professional setting where the decision process under review is no longer regularly exercised by humans in the first place? An expert supervisor of another's decision, in order to be worthy of the authority to override it, must have acquired expertise in making decisions of the very same or a similar kind. The requisite skills and wisdom that constitute such expertise could only be acquired, according to most theories of expertise, by having repeatedly and habitually practiced the actions in question, with an opportunity to learn from mistakes and successes, and to receive corrective feedback from others who already have the expertise one seeks to acquire. Where will the human supervisors of automated cyber-conflict acquire such practice, and the expertise in the proportionate and discriminating use of cyber-power that it

alone can engender? And what are the implications for human beings engaged in cyber-conflict, and for those impacted by cyber-conflicts, if they do not?

4. CONCLUSIONS

Of course, the reduction of human decision-making to mechanistic, formulaic or quasi-algorithmic processes can happen by means other than technological automation. We can easily conceive of military environments in which soldiers and officers are encouraged to eschew complex moral reasoning in favor of legalistic templates, decision-trees and other formal mechanisms of reducing the cognitive burdens (and freedoms) of human judgment.[2] Thus any 'moral deskilling' of the military profession need not be *essentially* linked to advances in the technological automation of war – it may have other causes as well. That said, the considerations above make it clear that advances in technological automation may greatly exacerbate any existing defects in the ability of today's military bodies to cultivate moral skills and virtues among their members and within their leadership ranks. What can be done to prevent such an eventuality? One option, of course, is a wholesale reversal of the shift toward automated methods of warfighting. While theoretically possible and perhaps even ideal, the expedience of such methods makes this reply of questionable utility. Are there other options?

Perhaps military institutions will compensate for the loss of moral skills in combat personnel by instead cultivating them in the software engineers responsible for programming automated systems to act ethically. But this does not answer the question of how, or through what new professional practices, software engineers could gain the needed moral expertise. Professional education would not be sufficient - the study of ethics textbooks, articles on just war theory or legal briefs on international laws of war do not by themselves enable skillful moral *action* or virtue– only repeated practice of the activities those books describe can produce the requisite capacities. Wargames or virtual-reality simulators might aim to engender in programmers and supervisors of automated systems the required habits and talents of moral discernment; but it is highly questionable whether simulations would carry the situational richness and moral gravity that produces genuine virtue.

Perhaps the best option is to restrict the deployment of automated methods of warfare to just those contexts in which human judgments are consistently and gravely inadequate *and* lead to morally intolerable error – while preserving robust opportunities elsewhere for, and expectations of, human soldiers to cultivate and exercise moral virtue in the conduct of war. This policy could be adopted alongside other uses of automated systems that create positive opportunities for moral

[2] Thanks to Don Howard for pointing this out in personal correspondence.

*up*skilling of military professionals. For example, rather than developing artificial moral intelligence that supplants human decision-making in the use of lethal force, artificial intelligence systems might instead be usefully deployed to provide soldiers with enhanced information about morally salient features of the battlefield, or to offer improved feedback concerning the alignment of soldiers' habits and decision patterns with norms of military honor, courage and restraint. However, such ethically constructive policies are unlikely to be pursued until and unless military leaders, educators, officers *and* the designers of automated systems jointly acknowledge the importance of preserving in military practice a developmental path for moral skills and virtues.

Acknowledgments

Thanks to Don Howard and John Sullins for helpful discussion of several of the key issues discussed in this paper.

REFERENCES

Ackerman, Spencer. 2012. "Pentagon: A Human Will Always Decide When a Robot Kills You." *Wired*. Last modified November 26, 2012. http://www.wired.com/dangerroom/2012/11/human-robot-kill/.

Aristotle. 1984. *The Complete Works of Aristotle: Revised Oxford Translation*. Edited by J. Barnes. Princeton: Princeton University Press.

Augustine. 1994. *Political Writings.* Edited by Ernest Fortin and Douglas Kries. Indianapolis, IN: Hackett Publishing.

Arkin, Ronald. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.

Arkin, Ronald. 2010. The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics* 9:4, 332-41.

Asaro, Peter M. 2008. "How Just Could a Robot War Be?" In *Current Issues in Computing and Philosophy*. Edited by Philip Brey, Adam Briggle and Katinka Waelbers, Amsterdam: IOS Press. 50-64.

Asaro, Peter M. 2009. "Modeling the Moral User." *IEEE Technology and Society Magazine* 28:1, 20-4.

Axtell, Guy and Olson, Philip. 2012. "Recent Work in Applied Virtue Ethics." *American Philosophical Quarterly* 49:3, 183-203.

Braverman, Harry. 1974. *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century.* New York: Monthly Review Press.

Carter, Ashton B. 2012. "Department of Defense Directive: Autonomy in Weapons Systems." Last modified November 21, 2012. *www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.*

French, Shannon E. 2003. *The Code of the Warrior: Exploring Warrior Values Past and*

Present. New York: Rowman and Littlefield.

Human Rights Watch. 2012. "Losing Humanity: The Case Against Killer Robots." Last modified November 19, 2012. http://www.hrw.org/reports/2012/11/19/losing-humanity-0.

Krishnan, Armin. 2009. *Killer Robots: Legality and Ethicality of Autonomous Weapons.* Surrey, UK: Ashgate Publishing Limited.

Lin, Patrick, Abney, Keith and Bekey, George A. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge, MA: MIT Press.

Manders-Huits, Noemi. 2006. "Moral Responsibility and IT for Human Enhancement." *Proceedings of the 2006 ACM Symposium on Applied Computing*, 267-71.

Olsthoorn, Peter. 2011. *Military Ethics and Virtues: An Interdisciplinary Approach for the 21st Century.* New York: Routledge.

Roberts, John. 2010. "Art After Deskilling." *Historical Materialism* 18: 77-96.

Reichberg, Gregory M., Henrik Syse and Endre Begby. 2006. *The Ethics of War: Classic and Contemporary Readings.* Oxford: Blackwell Publishing.

Rinard, Ruth G. 1996. "Technology, Deskilling and Nurses: The Impact of the Technologically Changing Environment." *Advances in Nursing Science* 18:4, 60-9.

Robinson, Paul. 2007. "Magnanimity and Integrity as Military Virtues." *Journal of Military Ethics* 6:4, 259-69.

Royakkers, Lambèr & van Est, Rinie. 2010. "The Cubicle Warrior: The Marionette of Digitalized Warfare." *Ethics and Information Technology* 12:3: 289-96.

Shakarian, Paulo, Devon Callahan, Jeffery Nielsen and Anthony N. Johnson. 2012. "Shaping Operations to Attack Robust Terror Networks." *Human Journal* 1: 15-25.

Sharkey, Noel. 2010. "Saying 'No!' to Lethal Autonomous Targeting." *Journal of Military Ethics* 9:4, 369-83.

Sharkey, Noel. 2012. "Automating Warfare: Lessons Learned from the Drones." *Journal of Law, Information and Science* 21:2.

Singer, Peter W. 2009. *Wired for War: The Robotics Revolution and 21st Century Conflict*. New York: Penguin Group.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy* 24:1, 62-77.

Toner, James H. 2000. *Morals Under the Gun: The Cardinal Virtues, Military Ethics and American Society.* Lexington, KY: University Press of Kentucky.

Townsend, Keith and Charles, Michael B. 2008. "*Jarhead* and Deskilling in the Military: Potential Implications for the Australian Labour Market." *Australian Bulletin of Labour* 34:1, 64-78.

United States Air Force. 2009. *Unmanned Aerial Systems Flight Plan 2009-2047.* Last modified May 18, 2009: Washington, D.C. http://www.govexec.com/pdfs/072309kp1.pdf.

Vallor, Shannon. 2013 (forthcoming). "Armed Robots and Military Virtue." In *Ethics of Information Warfare*. Edited by Luciano Floridi and Mariarosaria Taddeo, New York: Springer.

2013 5th International Conference on Cyber Conflict
K. Podins, J. Stinissen, M. Maybaum (Eds.)
2013 © NATO CCD COE Publications, Tallinn

An Ethical Analysis of the Case for Robotic Weapons Arms Control

John P. Sullins
Philosophy Department
Sonoma State University
California, U.S.
707-664-2277
John.sullins@sonoma.edu

Abstract: While the use of telerobotic and semi-autonomous weapons systems has been enthusiastically embraced by politicians and militaries around the world, their deployment has not gone without criticism. Strong critics such as Asaro (2008), Sharkey (2008, 2009, 2010, 2011, and 2012) and Sparrow (2007, 2009a, 2009b, 2011) argue that these technologies have multiple moral failings and their deployment on principle must be severely limited or perhaps even eliminated. These authors and researchers along with a growing list of others have founded the *International Committee for Robot Arms Control* as a means for advancing their arguments and advocating for future talks and treaties that might limit the use of these weapons. Others such as Arkin (2010), Brooks (2012), Lin, Abney and Bekey (2008, 2012), Strawser (2010), have argued that there are some compelling reasons to believe that, at least in some cases, deployment of telerobotic and semi-autonomous weapons systems can contribute to marginal improvements to the state of ethical and just outcomes in armed combat. This presentation will trace the main arguments posed by both sides of the issue. Additionally this paper will suggest certain considerations motivated by the philosophy of technology that might be worthy of addition to future robotic arms control treaties. This position argues that these technologies through the process of reverse adaptation can change our notions of just war theory to the point that caution in their use is recommended until further analysis of these effects can be accomplished. A realistic stance towards robotic weapons arms control will be argued for without losing sight of the positive role these technologies can play in resolving armed conflict in the most just and ethical manner possible.

Keywords: *Robotic Arms Control, Autonomous Weapons Systems (AWS), Just War Theory, Robot Ethics, Machine Ethics*

1. INTRODUCTION

The use of robotic weapons systems is accelerating around the globe. While the date of the first introduction of telerobotic weapons to the battlefield is debatable, it is clear that they have grown out of the use of guided missiles and radio controlled bombs in the last century, to the "smart" missiles and unmanned weapons systems of today. These systems are constantly evolving and a vast array of telerobotic and semi-autonomous weapons systems are being deployed in all potential theaters of conflict; land, sea and air (Singer, 2009). The epochal change in conflict resolution represented by the rise of more and more capable and autonomous weapons systems has not come without criticism. This paper will describe why the rise of autonomous weapons is so seemingly unavoidable and will look at the strong arguments in favor of severely limiting the research in, and deployment of, robotic weapons systems. An argument in favor of the cautious use of these weapons systems will also be developed.

2. THE FATAL ATTRACTION OF DRONES

If you are a politician in a liberal democracy, then the technology of unmanned weapons is the answer to your dreams. While armed aggression between liberal democracies is rare, they are involved in many military conflicts driven by clashes with nondemocratic countries or interventions in unstable regions of the world. Given the norms and values espoused in liberal democracies there is a political will to spread the virtues of democracy and check the aggressions of non-democratic powers. But other values and norms such as the distribution of political power to the voting population, severely hampers the governments of these countries who try to act on their military aspirations. There are massive political costs to be paid when it comes to deploying large numbers of troops in foreign lands. Sauer and Schörnig (2012), note that the governments of democracies want to engage in military activities but the citizens of democracies demand that these adventures be low cost with no casualties from their own military and low casualties inflicted on the enemy and local population.

> [T]he need to reduce costs, the short-term satisfaction of particular 'risk-transfer rules' for avoiding casualties, and the upkeep of a specific set of normative values – constitute the special appeal of unmanned systems to democracies" (Sauer and Schörnig, 2012, p. 365).

Unmanned weapons systems would seem to allow all constituents within a liberal democracy to achieve their goals. The weapons are not expensive compared to the massive expense of building, deploying and maintaining manned systems. The

missions are secret and so small they hardly warrant mention in the daily news back home. There is almost no risk to military personnel in their use and the number of enemy casualties per strike is relatively small. Also, politicians such as President Barack Obama have made claims that these weapons are far less indiscriminate and more tightly controlled in their killing than alternative modes of attack (Landler, 2012). We should note that this claim is backed up by what appears to be questionable methods used in the official calculation of civilian casualties since every adult male in the blast of the weapon is often considered a combatant by default, a claim that is often disputable (Qazi and Jillani, 2012). So the more precise targeting available on modern drones does not necessarily correspond to less civilian casualties (ibid; Zubair Shah, 2012). These weapons also come with a moral veneer that comes from the assumption that a more ethical conduct of war is possible using these machines. Sauer and Schörnig go on to conclude that the political goals along with the normative drives of liberal democracies necessitates that unmanned systems will continue to be armed with more powerful weaponry and that they will be given more demanding missions which will require greater autonomy from the machine (Sauer and Schörnig, 2012, p. 370). In addition to this they can also help with complex political relationships such as those between Pakistan and the United States. Drone strikes have arguably benefited the Pakistani government by allowing them a tool to attack their own political enemies while simultaneously being able to criticize the United States for those killings (Zubair Shah, 2012). In some cases the residents in tribal areas of Pakistan are sometimes grudgingly in favor of the strikes:

> Many favor the drone strikes over the alternatives, such as military operations or less selective bombardments by Pakistani bombers and helicopter gunships. Better a few houses get vaporized than an entire village turned into refugees (Ibid, p. 5).

This argument shows that we can expect the research into autonomous weapons systems to increase and for these systems to proliferate into every aspect of military activities across the globe. Recent history has given ample empirical evidence to back this theory up. Even though President Obama was elected largely on an anti-war vote, it has been reported that there has been an 8% increase in the use of drones during his first term and in 2011 drones were used in, "…253 strikes – one every four days…. [And] Between 2,347 and 2,956 people are reported to have died in the attacks – most of them militants" (Woods, 2011). This trend has only increased since that time. Other countries such as Israel and Italy are known to operate reconnaissance drones but recently the German government has announced plans to invest in both armed and unarmed drones (Gathmann et al, 2013; Kim, 2013; Medick, 2013). As the enthusiasm for this technology grows, a mounting opposition movement has also emerged that claims that this technology is not a

cheap, easy, casualty free means of propagating just war. Instead they claim that these technologies contribute to unjust military adventures and a indefensibly immoral push button warfare that claims the lives of thousands of innocents caught in the cross fire. In the interest of furthering the ethical use of these technologies, it is important that we give these counter arguments our full attention in this paper.

We should note that some technologies can cause what philosophers of technology call *reverse adaptation*; "…the adjustment of human ends to match the character of the available means" (Winner, 1978, p. 229). This is where the social and political milieu of a society changes to accommodate the inflexibility of a technology, rather than waiting to deploy the technology when it is more suited to the human environment. A prime example would be the way that human societies changed due to the adoption of mass production necessitating all manner of social upheaval that proved to be the fault lines of human conflict over the last two centuries. There are many other examples of this in recent history, think of the social disruption caused by the introduction of the automobile or cellular telephone, etc. It is obvious that autonomous weapons are again confronting our societies with the problems of reverse adaptation. These weapons are completing a trend in technological warfare begun in the nineteenth century that is making traditional notions of ethics in warfare largely untenable. These notions were always on shaky ground to begin with, but the tradition of just war that reaches back to the middle ages has become almost mute in its ability to influence decisions made on the modern battlefield. If this was not worrisome enough, as aging drone technologies are replaced with better equipment, the surplus will find use in civilian law enforcement duties, this will complete the circle and technologies that liberal democracies gladly used to control their enemies will now be used in ways that challenge and potentially curtail cherished civil liberties at home. Because of this, it is vital that we engage in discussion of these technologies at every opportunity. We will take up this issue again and apply it to the problem of creating a realistic robotic arms control at the end of this paper.

3. THE CALL FOR ROBOT ARMS CONTROL

Altmann (2009), Asaro (2008), Sharkey (2008, 2009, 2010, 2011, and 2012) and Sparrow (2007, 2009a, 2009b, 2011) have all challenged the perceived political expediency of robotic weapons systems as described above. They disagree with the claim that these weapons present a more limited and more just way of deploying military force and argue that their proliferation must be regulated. The most interesting and thoughtful counter arguments to the raise of the drones come from

the International Committee for Robot Arms Control (ICRAC).[1] This committee formed as an NGO in 2009 and its members represent concerned roboticists, legal scholars, philosophers, and other academics that seek to foster discussions in favor of robotic arms control. The five positions that they are in favor of supporting are listed on their website as follows.[2] Robotic weapons have the potential to lower the threshold of armed conflict. There should be a prohibition of the development, deployment and use of armed autonomous unmanned systems. No machines should be allowed to make the decision to kill people. There should be limitations on the range and types of weapons carried by "man in the loop" [telerobotic] unmanned systems and on their deployment in postures threatening to other states. There should be a complete ban on arming unmanned systems with nuclear weapons. And there should be a prohibition of the development, deployment and use of robot space weapons.

In addition to these propositions there are a number of other statements that are not agreed upon by all members of ICRAC but that there is broad agreement on such as: Limits on the endurance of unmanned weapons systems; size restrictions and or negotiated numbers of systems allowed by class and size; restrictions on operational range and size of payload; uninhabited weapons systems do not have the right to violate the sovereignty of states by incursions into their airspace or other territory.

We can distil these various claims into three main categories of proposed limitations. First, there should be limits on the authority given to decisions made solely by the machine. Second, bounds must be placed on the technological capabilities of these machines. And third, there must be restrictions placed on the deployment of autonomous weapons systems. Each one of these can be looked at from a technical, legal and/or ethical standpoint. In this paper we will deal only with the ethical justifications for and against these propositions. Let us now look at each one of these assertions in turn.

A. LIMITS TO AUTONOMOUS DECISION MAKING

The question of autonomous machines deciding when to use lethal force is the most ethically challenging of the three categories of proposed limitations and as such we need to pay more attention to it than the other two categories. Asaro (2008) noted that robotic weapons systems are developing along a spectrum from non-autonomous, through semi-autonomous, to fully autonomous. As we move up the scale to full autonomy there will be a critical step taken when we allow machines to

[1] http://icrac.net

[2] http://icrac.net/statements/

select and engage military targets on their own with little or no input from human operators (Asaro, 2008). His primary argument being that doing so will cloud our ability to ascribe guilt or responsibility to anyone in the event that something goes wrong. The machine might be capable of finding and attacking targets but it is unlikely to be capable of taking a moral stand to justify its own decisions (ibid, p. 2). So, in building this decision making into the machine, we are uploading our moral responsibility to the machine as well and abdicating our duties as moral agents. Furthermore, we can see that if robots are not capable of making the moral decision to kill a human being, then this situation must be avoided. In recent public talks he has begun to wonder if we ought to claim the human right not to be killed by autonomous weapon systems.[3]

Sharkey (2010), as a robotics researcher himself, argues mainly from the view that machines are never going to be able to reliably make the right choices in difficult situations on the battlefield. Robots can barely tell the difference between a human and a trash can, which begs the question of how they are going to be able to tell the difference between an innocent civilian caught on a battlefield and an irregular soldier who is posing as a civilian. This is a challenging task for a human being and well beyond the capabilities of a machine. This limitation would make an autonomous fighting machine somewhat indiscriminant and therefore unjustifiable from a moral standpoint. In an interview, Sharkey has suggested that those funding research into autonomous weapons systems have an almost mythic faith in the ability of artificial intelligence to solve these kinds of problems in a prompt manner and that this belief is far from the reality of what these systems are capable of, "[t]he main problem is that these systems do not have the discriminative power to do that," he says, "and I don't know if they ever will" (Simonite, 2008). Again, we are mistakenly placing our trust in a machine that is actually incapable of making good moral decisions, a position that is morally suspect indeed.

Sparrow (2007), argues that it would be immoral to give machines the responsibility of choosing their own targets even if we can somehow transcend the technical problems of complex decision making and target discrimination. He asks us to consider what we would do in the case of a machine that decided on its own to commit some sort of action that if a human had done it would constitute a war crime. In that case he argues we would find that there is no good answer when we try to decide where to affix the moral blame for the atrocity (ibid, p. 67). Asaro believes this is due to the fact that in the end, there is no way to punish a machine as they have neither life nor liberty to lose nor would it be reasonable to assume

[3] Asaro, P. (Forthcoming). "On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-making," Special Issue on New Technologies and Warfare, International Review of the Red Cross.

that the responsibility for the act rested in the machine itself, or its commanding officers or even in its builders and programmers (ibid). *Jus in bello* requires that there be an ability to assign responsibility for war crimes and that the perpetrators be punished. "If it turns out that no one can properly be held responsible in the scenario described above, then AWS [autonomous weapons systems] will violate this important condition of *jus in bello*" (ibid, p. 68) Consequently, Asaro concludes that the use of this kind of weapon would be immoral and hence must be avoided.

The above arguments call into question not only the morality of having a machine decide to kill an individual human being but even their use of force to simply injure an opponent or follow opponents across political boarders as this would no doubt incite retaliation and could even lead to an escalating situation where decisions by a machine might lead to open warfare between humans. This leads ICRAC to suggest that these decisions should never be left to a machine alone.

While it is quite reasonable to seek limits on the use of autonomous weapons in situations where they could inadvertently escalate a conflict, the argument that autonomous weapons need to be banned due to the fact that they are incapable of affixing moral blame to is much harder to follow. Even if it were problematic to ascribe moral agency to the machine for metaethical reasons, there would still be legal recourse and the commanding officers that deployed the weapon as well as its builders and programmers could be held liable. Of course if these people were also shielded from culpability through various legal means, then Sparrow would be correct in making a strong claim that the lack of a responsible agent renders the use of these weapons immoral. It is not clear that that is happening yet so this argument should be tabled until there is evidence suggesting that military commanders are claiming autonomous weapons have a rogue status that absolves anyone but the machine itself of moral responsibility.

It is difficult to find arguments in favor of giving the choice to use lethal force solely to machines. Yet it has been argued that if machines truly did have the ability to accurately distinguish between targets, then we might expect a great deal of precision in these judgments and in that case it would be moral to allow them some autonomy on the battlefield (Lin, Abney, and Bekey, 2008). Given that machines would not experience the same combat related stresses that make human judgment prone to error on the battlefield, there might be good reason to take this claim seriously. Higher reasoning powers in humans are often the first casualty when violence erupts causing some to make regrettable decisions. A machine, for instance, would not have to instantly respond to enemy fire since it does not have a right to self-preservation. It could instead wait and fire only when it was sure of its target. Ostensibly, it would be able to deliver return fire accurately with less chance of harming innocent civilians which might marginally improve the ethical outcomes

of violent encounters. If the members of ICRAC are wrong in their assessment of the ability of these machines to discriminate between targets, then that somewhat weakens their case.

Arkin (2007, 2010), argues a more subtle point. He might agree that machines should only fire their weapons under human supervision but he would like to see that machines have the ability to autonomously decide not to fire their weapons even when ordered to do so by humans. He would rather design a system that independently reviewed the constraints on its operation imposed by the rules of operation, laws of war, just war theory, etc., that it was operating under. This system, called an "ethical governor," would continuously assess the situation and if the machine decided that the operation was beyond set parameters then it would disengage its ability to fire. In this way the machine's artificial ethical governor would also be able to control human decisions that might be immoral or illegal but that emotion or the heat of the battle had made the human actors unable to accurately process (ibid). In an interview Arkin said that, "[o]ne of the fundamental abilities I want to give [these systems] is to refuse an order and explain why" (Simonite, 2008). Again, since the machine has no right to self-preservation, it can legitimately disarm itself if needed. Additionally he argues that, the machine can gauge the proportionality of the fire it delivers to suit the situation it is in. Where a human might need to fire to ensure that the enemy is killed and no longer a threat to his or her person, the machine can take a calculated risk of destruction and instead only would apprehend an enemy rather than always delivering lethal force. An additional strength of this design would be that it would put a safety layer on the possibility that the human operators of the machine might be making an immoral decision to fire based on emotion, stress, or improper understanding of the situation on the ground. The robot would be able to disobey the order to fire and explain exactly why it did without any fear of dishonor of court martial that a human soldier might succumb to in a similar situation (Arkin 2007, 2010; Sullins 2010a). The system Arkin proposes would be far less likely to cause the false positive errors of excessive or indiscriminant use of force that other critics worry about, but it does leave open the possibility of a false negative, where a legitimate target may get away due to situations that cause the ethical governor to engage. What if this enemy then went on to commit his or her own war crimes? Surely this would be an immoral outcome. And is most likely why we have yet to see this system deployed.

We can see that the stance on banning autonomous targeting decisions is indeed one that requires more discussion and it is appropriate to place it on the table for potential restrictions in any robotic arms control deliberations.

B. *LIMITS TO THE TECHNOLOGICAL CAPABILITIES OF ROBOTIC WEAPONS*

There is a vast array of unmanned systems in use or in development in the world today. Everything from tiny surveillance drones that look like hummingbirds or spiders, to autonomous fighting vehicles and ordnance removal systems, to massive aircraft or sea vessels loaded with lethal missiles. Major world powers such as the U.S. and China are vigorously pursuing the deployment of these systems (US Department of Defense, 2007, 2011, 2012; Von Kospoth, 2009). As this arms race continues unabated, the question remains as to whether or not we could have a more just world without these systems. Altmann (2009) has argued for very strong restrictions on the proliferation of unmanned military vehicles if not a complete ban on them. Citing the success of arms control in keeping the Cold War cold, he argues that robotic arms control must be used today as a means of preventing these weapons from growing out of the ability for human control and he suggests that:

> Whereas military UAVs for surveillance already are deployed by dozens of countries, providing them with weapons has only begun recently. If unchecked by preventive arms control, this process will spread to many more countries. Later, similar developments are possible in uninhabited vehicles on land, on and under water and – to a more limited extent – in outer space. Seen from a narrow standpoint of national military strength, these developments will provide better possibilities to fight wars and to prevail in them. However, if one looks at the international system with its interactions, the judgment will be different, in particular concerning armed robots/uninhabited systems. Destabilization and proliferation could make war more probable, including between great/nuclear powers. Criminals and terrorists could get more potent tools for attacks, too (ibid).

Proliferation and escalation are the main arguments that Altmann brings to bear on this problem. If we allow the technology to continue its development without any checks to its growth, then he argues that this could lead to destabilizing events and weaponry finding its way into unsavory hands. Presumably terrorists and other ne'er-do-wells would also like cheap, reliable weapons that can cause harm to others with no risk to themselves.

ICRAC seems to agree, at least in principle, with this assessment and specifically asks for a ban on nuclear armed autonomous weapons.

There is unlikely to be anyone that would argue for uncontrolled growth in autonomous weapons, though as Singer (2009), notes in his book *Wired For War,* there was little reason for the U.S. to self-impose limits on this technology since they were the first to make extensive use of it, but the first mover advantage has

slipped. Now the interests of the U.S. would be best served by being a party to robotic arms limitation negotiations.

Another reason to place limits on robotic weapons is that there is a potential that these weapons might be successfully engaged through cyberwarfare and hijacked. An enemy could then turn the weapons systems against their owners or use them for terrorist activities. For this reason, it may be prudent to keep them unarmed and small to limit the damage they are capable of. Currently there is no known successful hijacking of a military system given that these systems utilize strong encryption on the commands and communications between the drone and its operators. There have been reports of a successfully hacked drone using the proposed civilian communications protocols now under development by the Federal Aviation Authority for the use of drones by civilian operators (Homeland1 Staff, 2012). It was found that the GPS systems could be manipulated by a third party causing the craft to veer off course and potentially crash at the bidding of the researcher from the University of Austin posing in this instance as a terrorist hacker (ibid). It is in the self-interest of parties that might be the targets of autonomous weapons systems to seek means to defeat or control these weapons through cyberwarfare, so we should expect an arms race in this sub-discipline of cyberwarfare. Paradoxically, one way to help defeat these attacks would be to build systems that do not interact that much with their operators and can do their mission stealthily and autonomously and thus avoid the notice of enemy cyberwarriors before the mission is complete. Thus it is more likely that we will see both increased encryption of military systems and more autonomous decision making by the system itself. This may also happen in the civilian sphere of operations but that is a separate topic.

C. *RESTRICTING THE DEPLOYMENT OF ROBOTIC WEAPONS*

These restrictions are concerned with where, and for what purpose, robotic weapons are deployed. ICRAC proposes a complete prohibition of deploying robotic weapons in space. Presumably this is meant to cover both autonomous and semi-autonomous machines. In addition they propose a ban on the deployment of all autonomous unmanned systems regardless of the theater of operation. Yet they do seem to tolerate some minimal use of teleoperated systems as long as they are not space based.

The ethical arguments opposing the deployment of robotic weapons in space tends to appeal to extending existing bans of the deployment of weapons of mass destruction (WMDs) in orbit or on the moon. When it comes to robotic weapons armed with WMDs, then this is a strong argument. There are grey areas however in

that many satellites have both a civilian and military use, GPS is a prime example. If we imagine a semi-autonomous satellite that provides both civilian and military functionality, then should such a machine be banned completely? Furthermore, in any future war, controlling the very high ground of space would be a vital military objective and it seems ambitious to believe that any country with advanced capabilities in space would consent to sign on to such a ban. Also, as countries and corporations begin to mine our nearby planetary and asteroid neighbors, it is very likely they will wish to protect their investments with armed planetary rovers alongside the mining bots. It is hard to see this as an immoral impulse. Of course using these machines to wantonly attack the mining operations of others is a different matter. But we have international law to appeal to in that eventuality. A better solution would be to attempt to limit the size and capabilities, or the numbers deployed, of autonomous military satellites and/or planetary rovers.

The moral support for a ban on the deployment of any autonomous robotic weapons depends entirely on whether it is decided that there is a human right not to be the target of a robotic weapon as described in the section above on limiting autonomous decision making. We were unable to come to a full conclusion on that concept. The precautionary principle would suggest that until we do, a ban is justified. But if a supra human robotic moral agent or a good moral reasoning augmentation system of the sort that Arkin proposes with his ethical governor is indeed developed, then it would actually be immoral not to deploy robotic weapons so constructed.

Even when it comes to telerobotic weapons systems, certain limits on deployment should be considered. There is wide agreement that these systems along with other high tech advances have already been used in ways that can challenge interpretations of 2(4) of the UN Charter governing the resort to force as well as the International Humanitarian Law, and the rules of armed conflict (Altmann, 2009; Arquilla, 2010, 2012; Asaro, 2011; Brooks, 2012; Caroll, 2012; IHLRI, 2013; Khan, 2002; Lin, 2010; Lin, Abney and Beekey, 2008; Marchant et al., 2011; Oudes and Zwijnenburg , 2011; Rohde, 2012; Sauer and Schörnig, 2012; Sharkey, 2008, 2009, 2010, 2011, 2012; Singer, 2009; Sparrow, 2007, 2009a, 2009b, 2011; Sullins, 2010a, 2011; Wallach and Allen, 2009).

It is quite difficult to find arguments in favor of no special controls on unmanned weapons, but Strawser (2010), has argued that there is actually a moral duty to use uninhabited aerial vehicles.

> "…any new technology that better protects the just warfighter is at least a prima facie ethical improvement and is morally required to be used unless there are strong countervailing reasons to give up that protection. I have argued that if using UAVs (as a particular kind of remote weapon) does not incur a significant loss of capability particularly the operators' ability

> to engage in warfare in accordance with *jus in bello* principles then there is an ethical obligation to use them and, indeed, transition entire military inventories to UAVs anywhere it is possible to do so" (Strawser, 2010).

It is vitally important to note that Strawser makes this claim pertains only to unmanned aerial vehicles and under the condition that these systems be used only if *jus ad bellum* (lawful state of war) has been achieved, a situation he doubts has actually obtained in all of the recent uses of these weapons.

4. SUGGESTIONS FOR THE DESIGN OF FUTURE ROBOTIC ARMS TREATIES

As we have seen in this paper so far there are some developed positions on robotic arms control. One is held by countries that are prime movers in the early development of telerobotic, semi-autonomous and autonomous weapons systems who seek to let these technologies develop largely unregulated. This position is best illustrated through the promises made by leaders such as Barack Obama who assure us that these weapons are under "tight control" but controls that are not made public (Landler, 2012). Another strong position on the other end of the spectrum is that held by the members of ICARC as described in this paper which seeks to place strong publicly stated limits on the semi-autonomous versions of this technology and an outright ban on autonomous weapons systems.[4] As mentioned above, ICARC is a coalition of academic philosophers, roboticists, and other scholars. It is important to also recognize that there has been important research into the legal definitions and justifications of cyberwarfare. The Tallinn Manual represents a three year attempt to apply international law to cyberwarfare and it outlines the legal territory that justifies certain uses of these weapons as long as there use comports to international laws and treaties (Schmitt, 2013). The Tallinn Manual does not have the force of a treaty but it is a powerful tool in the construction of future treaties for the control of cyberwarfare. The document does serve as a detailed look at the legal justifications for military operations in cyberspace as implemented by the leaders of NATO nations such as the US executive branch (Ibid). Unfortunately for our purposes here, The Tallinn Manual is specifically designed for cyberwarfare (Ibid). As such, it only strictly applies to robotic weapons systems when they are either the targets of cyber-attacks or in their potential roll as attack vectors for launching acts of cyberwar. There are many points of overlap between cyberweapons and robotic weapons but it is not a one to one match. For instance, cyberweapons as of this point in time are incapable of directly launching kinetic attacks, whereas

[4] See: http://icrac.net/statements/

this is commonly done with robotic weapons systems. Thus one of the primary conundrums with cyberwarfare is what activities in cyberspace actually count as acts of war and would it be just to launch a kinetic attack in reaction to a successful hack of a computer system? Since robots function primarily (but not entirely) in physical space, this particular moral question does not arise with the same force as it does with cyber-attacks. Also, it is difficult to delineate the 'zone of conflict' in cyberspace, whereas a robot inhabits physical coordinates that can be precisely determined to be either inside or outside a 'zone of conflict.' In fact that is one of the current debates in robotic warfare that seems perfectly resolvable but due to reverse adaptation, has become muddled. Drones habitually fire weapons in areas that are far from known zones of conflict and thus make life very hard for civilians who can't know for sure if they are in the line of fire (Glaser, 2013). This seems to be a very immoral situation even if it is not found to be a strictly illegal situation.

This brings us to my main point in this section. One of the earlier commenters to this paper asked: Why then cannot the "old" ethics principle be similarly applied to the cyber conflict? That is a good question, why don't the standard ethical principles developed over the last three millennia along with the laws that they have inspired just settle the issue before it arises? There are two factors that challenge this very good common sense notion which we have discussed above but will go over again here.

The existing laws of armed combat are far less capable of controlling the misuse of robotic as well as cyber weapons than we might wish given the phenomenon of technological reverse adaptation and to the special problems that law encounters around the quickly information technologies it is trying to control. Certain new information technologies can rapidly alter the social milieu and given that these technologies change faster than the laws attempting to regulate them can the result is a permanent policy vacuum (Lessig, 1999). Networked web applications, cloud computing, mobile phones, autonomous robots, are all information technologies that display this behavior. Since these technologies are all used in cyberwarfare and since warfare by its very nature is about gaining advantage over an opponent, we can expect this policy vacuum to be almost insurmountable. In addition to this, technological reverse adaptation as described early in this paper causes political policy to eventually adapt to the technological change and this process can mask the policy vacuum as the changes in social norms brought about by these transformative technologies quickly become the new normal, standard of behavior or policy. As an example look at how quickly the social norms around what is public or private information has shifted over just a generation.[5]

[5] For a good example see (Zick, 2012).

From this we can derive two cautions for developing robotic arms control. One is to acknowledge that no set of existing laws will be sufficient and that new laws and policies will have to attempt to keep pace with developments. The second is to recognize that since the technologies push beyond the borders of legal guidance, the very design and testing process of new technologies must be guided by moral commitments in order for these values to be expressed in the resulting technologies.

This means that concepts of *jus ad bellum* and *jus in bello* must enter into the design process and not only be imposed after the technology is released on the world. This is exactly the opposite of what common sense or current industrial practice might dictate in the design of robotic weapons, but it is a necessary addition to the design process in order to assure that these weapons comport to the values in just warfare we hold as a society. What I am saying is that these weapons will only display these values if their designers and programmers personally hold these values, they cannot be effectively imposed on the technology only from outside regulation. Let us now turn to the ethical values that are germane to the project at hand.

A. ETHICAL NORMS NEEDED TO GUIDE THE DESIGN OF ROBOTIC WEAPONRY AND POLICIES FOR THEIR USE

Above we covered the three major categories proposed for the regulation of robotic weapons systems; Autonomous decision making, technological capabilities, and deployment. In each case we looked at arguments both pro and con for limits on each. Here I will succinctly layout my suggestions, not as an attempt to write law but as an attempt to craft ethical justifications that could guide the design of robotic weapons systems or the design of laws that attempt to regulate them.

Autonomous decision making by robotic weapons systems—we are ethically justified in placing stringent controls onto automatic target acquisition and engagement by robotic weapons systems. Documents such as the Tallinn Manual provide detailed descriptions of legal targets in cyberwarfare (Schmitt, 2013). But it would be a mistake to assume that we are capable of designing systems that can properly discriminate these targets to the level necessary to comply with international law (Sharkey, 2010). There is also the unresolved but morally compelling argument that it might be a human right not to be killed by an autonomous system (Asaro, 2011, 2008). Although we need to protect robotic weapons systems from hijacking by increasing autonomy, we should not allow military expedience to reverse adapt our moral intuitions here.

Technological capabilities of robotic weapons systems—it is impossible to predict the exact path of growth in the future technological capabilities of robotic weapons systems so this means we are ethically justified in demanding precaution

over unbridled advance. We should assume that every advance will lead to reverse adaptation unless we consciously attend to ensuring that it does not. And demand that these weapons progress in more accuracy and less damage to people and property. Ethical values must guide this design. For instance, greater target acquisition capabilities and target discrimination algorithms can lead to both ethical and unethical designs and we have to consciously choose the ethical design. For instance these capabilities could advance a nations ability to commit extra judicial killings or they could be used to create a system that could target just the weapon, a Lone Ranger bot if you will. Funny as it sounds now, it just might be possible in the future and that machine would be a good ethical choice to add to our tools for conflict resolution.

Deployment of robotic weapons systems—assuming the above values are in effect, it is more ethical to use unmanned weapons systems than manned weapons systems. Unmanned systems can take chances that manned systems cannot and can therefor risk destruction to ensure that they have a legal and just reason for the use of lethal force. They can also risk destruction and use less lethal or even non-lethal weapons that would be a foolish risk if deployed by a human. But these deployments must not be such that they extend the zone of conflict beyond reason. For instance, weaponized drone satellites would require that entire hemispheres be considered zones of conflict with any civilian at any time potentially putting her or him at risk of being collateral damage with no chance of refuge. This would be an unjust world so robotic weapons must be deployed in regulated zones of conflict and every effort made to warn innocent noncombatants of their potential risk.

If all three of these sets of ethical norms are respected in the design of robotic weapons and or the design of treaties limiting their use, then we will better succeed in fostering jus *ad bellum* and *jus in bello.*

5. CONCLUSION -- A VALUES BASED CONTROL OF ROBOTIC WEAPONS

The first sections of this paper have shown that telerobotic drones in particular and semi-autonomous weapons systems in general have become a permanent weapon in the arsenals of the world. Simply put, modern political and military values are strong motivators for the continued development and deployment of robotic weapons. This means they are not going to go away. But given the weight of the ethical discomfort that has resulted from the recent use of telerobotic weapons systems, and the threat of technological reverse adaptation, it is reasonable to argue for placing limits on the design and use of robotic weapons systems. But only with the caveat that we take seriously the claim that these weapons could also be significant tools for complying

with *jus in bello* and that it would be immoral to limit them if that were the case. An accurate answer to that last question requires much more research. As we have seen in this paper there are many arguments both pro and con on this issue, but we also have the potential of settling this case with information gathered from an analysis of the last two decades of the use of telerobotic and semi-autonomous drones on the battlefield and in covert actions. This kind of research will have to wait for all of these reports to become declassified but over time they will and we will be able to say with much more certainty whether or not this technology has contributed to a more just and moral world.

It is vital to continue research such as that done by Arkin and other roboticists who seek to explore how much ethics and morality we can put into the design of our fighting machines. In fact, as was argued in the last section we are ethically required to do so. Since all technologies are expressions of the values of their makers, if we care about ethics and morality, it will show in the machines we build. In that way I humbly disagree with some of the members of ICRAC such as Sharkey when he argues that roboticists should avoid working on drones (see Sharkey, 2010). I agree that there are a large number of roboticists and engineers I would wish were not working on drones, but someone like Sharkey is precisely the kind of person that values centered design required to be working on these technologies as he has a very well developed sense of moral value and is also skeptical of military jingoism—and that is the kind of dissenting voice needed on design teams to create innovative and ethical machines.

Robotic arms control treaties must now be negotiated but we should not expect that a complete ban on these weapons is a realistic goal, except in the case of robots armed with WMDs and use of these weapons outside of the norms described in the last section. But we must also remember that the trend toward informational and cyberwarfare; of which robotic weapons is just a part, has already begun to challenge traditional notions of *jus ad bellum* and *jus in bello* through the effects of technological reverse adaptation to the point where even those cherished norms need to be redefined and renegotiated.

New information technologies have challenged traditional ethical norms over and over in the last fifty years and the pace of those challenges is accelerating. Theorists have argued that these challenges require a strong rethinking of our traditional moral norms and that we cannot rest on our laurels when it comes to moral theory (Bynum, 2000; Floridi and Sanders, 2003; Moor, 1985; Sullins, 2010b; Tavani, Herman, 2004). What worked in the past is not guaranteed to work in the future which requires a nimble regulatory structure that is proactive during the design stage of robotic weapons systems.

In this paper a realistic stance towards robotic weapons arms control was argued for but not at the cost of losing sight of the potentially positive role robotic weapon systems might play in resolving armed conflict in the most just and ethical manner possible. This is achieved by adhering to ethical norms of limiting certain aspects of autonomous decision making in regards to targeting humans, limits to the technological capabilities of robotic weapons systems, and limits to their deployment or use. And there limits must be consciously addressed during the design of the machines themselves in order to limit the effects of technological reverse adaptation.

REFERENCES

Altmann, J. (2009). Preventive Arms Control for Uninhabited Military Vehicles, in *Ethics and Robotics*, R. Capurro and M. Nagenborg (eds.) AKA Verlag Heidelberg. Accessed Jan 25 2013 at: http://e3.physik.tu-dortmund.de/P&D/Pubs/0909_Ethics_and_Robotics_Altmann.pdf

Arkin Ronald C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics* 9(4): 332–341.

Arkin, Ronald C. (November, 2007): Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, Technical Report GIT-GVU-07-11, Mobile Robot Laboratory College of Computing, Georgia Institute of Technology. (http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf)

Arquilla, John. (2012). Cyberwar is Already Upon Us: But can it be controlled? *Foreign Affairs*, March/April, Accessed Jan 25, 2013 at: http://www.foreignpolicy.com/articles/2012/02/27/cyberwar_is_already_upon_us

Arquilla, John. (2010). The New Rules of War, Foreign Policy, March/April, Accessed Jan 25, 2013 at: http://www.foreignpolicy.com/articles/2010/02/22/the_new_rules_of_war

Asaro, P. (2011). Military robots and just war theory. In: Dabringer, G. (ed.) *Ethical and Legal Aspects of Unmanned Systems*. Vienna: Institut für Religion und Frieden, 103–119.

Asaro, Peter M. (2008). How Just Could a Robot War Be? In P. Brey, A. Briggle, & K. Waelbers (eds.), Current Issues in Computing and Philosophy (pp. 50-64). Amsterdam: Ios Press.

Brooks, Rory. (2012). What's Not Wrong With Drones? Forign Policy. Accessed Jan 25, 2013 at: http://www.foreignpolicy.com/articles/2012/09/05/whats_not_wrong_with_drones

Bynum, Terrell W. (2000). Ethics and the Information Revolution. *Ethics in the Age of Information Technology,* Linkoping University, Sweeden, pp. 32-55.

Caroll, R. 2012: Drone Warfare: A New Generation of Deadly Unmanned Weapons. The Guardian. Accessed Jan 25, 2013 at: http://www.guardian.co.uk/world/2012/aug/02/drone-warfare-unmanned-weapons

Dabringer G (ed.) (2011). *Ethical and Legal Aspects of Unmanned Systems*, Vienna: Institut für Religion und Frieden.

Floridi, Luciano and Sanders, J.W. (2003). The Foundationalist Debate in Computer Ethics. *Readings in CyberEthics (2nd ed),* Jones and Bartlett Publishers, Inc. Canada.

Gathmann, F., Gebauer, M., Medick, V., and Weiland, S. (2013). Deutschlands Drohnenpläne: Merkel rüstet auf, Spiegel Online, January 25. Accessed February 6, 2013 at: http://www.spiegel.de/politik/deutschland/kampfdrohnen-plaene-der-regierung-stossen-auf-heftigen-widerstand-a-879701.html

Glaser, John (2013). Terrorized by Drones, Afghan Civilians Increasingly Flee Homes. *Anti War.com March 28. Accessed March 29 at: http://news.antiwar.com/2013/03/28/terrorized-by-drones-afghan-civilians-increasingly-flee-homes/*

Homeland 1 Staff, (2012). Researchers: Drones vulnerable to terrorist hijacking. *Homeland1,* July 2. Accessed on March 29, 2013 at: http://www.homeland1.com/Security-Technology/articles/1309966-Researchers-Drones-vulnerable-to-terrorist-hijacking/

International Humanitarian Law Research Initiative (IHLRI), Harvard University. Accessed on February 2, 2013 at: http://ihl.ihlresearch.org/index.cfm?fuseaction=page.viewpage&pageid=2083

Kahn Paul W. (2002). The paradox of riskless warfare. *Philosophy & Public Policy Quarterly* 22(3): 2–8.

Kim, Lucia (2013). Germany and Drones. *International Herald Tribune,* February 5. Accessed Feb 5, 2013 at: http://latitude.blogs.nytimes.com/2013/02/05/germany-and-drones/?nl=opinion&emc=edit_ty_20130205

Landler, Mark (2012). Civilian Deaths Due to Drones Are Not Many, Obama Says, *The New York Times,* January 30. Accessed March 28, 2013 at: http://www.nytimes.com/2012/01/31/world/middleeast/civilian-deaths-due-to-drones-are-few-obama-says.html?_r=0

Lessig, Larry (1999). The Code is the Law. *Industry Standard, April 19-26, 1999.*

Lin, Patrick (2010). Ethical blowback from emerging technologies. *Journal of Military Ethics* 9(4): 313–331.

Lin Patrick, Abney, Keith,, and Bekey George. (2008). Autonomous Military Robotics: Risk, Ethics, and Design. San Luis Obispo, CA: California Polytechnic State University.

Lin P, Bekey G and Abney K (eds) (2012). Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA: MIT Press.

Marchant, G.E. et al. (2011). International governance of autonomous military robots. The Columbia Science and Technology Law Review 12: 272–315.

Medick, V. (2013). 'Credible Deterrence': Germany Plans to Deploy Armed Drones, *Spiegel Online*, January 25, 2013. Accessed February 6, 2013 at: http://www.spiegel.de/international/germany/germany-plans-to-deploy-armed-drones-in-combat-abroad-a-879633.html

Moor, James H. (1985). What is Computer Ethics? *Metaphilosophy,* 16 (4), pp. 266-275.

Qazi, Shehzad H. and Jillani, Shoaib (2012). Four Myths about Drone Strikes. *The Diplomat,* June 9. Accessed March 28, 2013 at: http://thediplomat.com/2012/06/09/four-myths-about-drone-strikes/

Oudes C and Zwijnenburg W. (2011). Does Unmanned Make Unacceptable? Exploring the

Debate on Using Drones and Robots in Warfare. Utrecht: IKV Pax Christi.

Rohde, David. (2012). The Obama Doctrine: How the President's War is Backfiring. *Foreign Policy,* March/April. Accessed on Feb 1, 2013 at: http://www.foreignpolicy.com/articles/2012/02/27/the_obama_doctrine

Sauer, Frank and Schörnig, Niklas. (2012). Killer Drones – The Silver Bullet of Democratic Warfare? Security Dialogue 43:4, 363-380.

Schmitt, Michael N. (2012). International Law in Cyberspace: The Koh Speech and Tallinn Manual Juxtaposed, 54 *Harv. Int'l L.J.* Online 13. Accessed March 29, 2013 at: http://www.harvardilj.org/2012/12/online-articles-online_54_schmitt/

Schmitt, Michael N. (Gen. Ed.) (2013). Tallinn Manual on the International Law Applicable to Cyber Warfare. Cambridge University Press. Accessed on March 3, 2013 at: http://www.ccdcoe.org/249.html

Sharkey, N. (2008). Grounds for Discrimination: Autonomous Robot Weapons. RUSI Defence Systems, 11 (2), 86-89.

Sharkey Noel. (2009). Death strikes from the sky. IEEE Technology and Society Magazine 28(1): 16–19.

Sharkey Noel. (2010). Saying 'no!' to lethal autonomous targeting. Journal of Military Ethics 9(4): 369–383.

Sharkey Noel. (2011). Moral and legal aspects of military robots. In: Dabringer G (ed.) Ethical and Legal Aspects of Unmanned Systems. Vienna: Institut für Religion und Frieden, 43–51.

Sharkey, Noel. (2012). Killing Made Easy: From Joysticks to Politics. In Lin, P., Abney, K., and Bekey, G.A., Robot Ethics: The Ethical and Social Implications of Robotics (pp. 111-128). Cambridge, MA: MIT Press.

Simonite, Tom. (2008). ‚Robot Arms Race' Underway, Expert Warns. *New Scientist,* February 27. Accessed on Feb 1 2013 at: http://www.newscientist.com/article/dn13382-robot-arms-race-underway-expert-warns.html

Singer, Peter, W. (2009). *Wired For War*, New York: Penguin Press.

Sparrow, Robert, W. (2007). Killer Robots. Journal of Applied Philosophy, Vol 24, No. 1.

Sparrow Robert, W. (2009a). Predators or plowshares? Arms control of robotic weapons. IEEE Technology and Society Magazine 28(1): 25–29.

Sparrow, Robert, W. (2009b). Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Applications. Science and Engineering Ethics 15: pp. 169-187.

Sparrow Robert W. (2011). The ethical challenges of military robots. In: Dabringer G (ed.) Ethical and Legal Aspects of Unmanned Systems. Vienna: Institut für Religion und Frieden, 87–102.

Strawser BJ. (2010). Moral predators: The duty to employ uninhabited aerial vehicles. Journal of Military Ethics 9(4): 342–368.

Sullins, John. P. (2010a). RoboWarfare: Can Robots be More Ethical Than Humans on the

Battlefield? Ethics and Information technology 12(3): pp. 263-275.

Sullins, John P. (2010b). Rights and Computer Ethics. *The Cambridge Handbook of Information and Computer Ethics, Floridi, L. (ed), pp. 116-133,* Cambridge University Press, UK.

Sullins, John P. (2011). Aspects of telerobotic systems. In: Dabringer G (ed.) Ethical and Legal Aspects of Unmanned Systems. Vienna: Institut für Religion und Frieden, 157–167.

Tavani, Herman (2004). Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology. *Wiley,* New York USA.

US Department of Defense (2007). Unmanned Systems Roadmap 2007–2032. Washington, DC: US Department of Defense.

US Department of Defense (2011). Aircraft Procurement Plan: Fiscal Years (FY) 2012–2041. Washington, DC: US Department of Defense.

US Department of Defense, November 12, (2012). Autonomy in Weapons Systems. Directive Number 3000.09. Accessed Jan 25 2013 at: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf

Von Kospoth, N. (2009). China's leap in unmanned aircraft development. Available at: http://www.defpro.com/daily/details/424/ (accessed 20 October 2011).

Wallach, W. & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong (New York: Oxford University Press).

Winner, Langdon (1978). Autonomous Technology: Technics-out-of-control As a Theme in Political Thought. *MIT Press*.

Woods C (2011). Drone war exposed. Accessed Jan 25 2013 at: http://www.thebureauinvestigates.com/2011/08/10/most-complete-picture-yet-of-cia-drone-strikes/

Zick, Colin J. (2012). Survey Reveals Generation Gap in Employee Attitudes Toward Confidential Information. *Security, Privacy and the Law,* Blog published by Foley Hoag LLP. Accessed March 29 2013 at: http://www.securityprivacyandthelaw.com/2012/06/survey-reveals-generation-gap-in-employee-attitudes-toward-confidential-information/

Zubair Shah, Pir. (2012). My Drone War. *Forign Policy,* March/April. Accessed Feb 1, 2013 at: http://www.foreignpolicy.com/articles/2012/02/27/my_drone_war?page=0,4

Biographies

EDITORS

Karlis Podins, civilian working for the Latvian Ministry of Defence, has been a scientist in NATO CCD COE for more than four years. Before joining public sector he has worked in academia and private sector with research and security related tasks.Karlis has a distinguished master's degree in Computer Science from University of Latvia. His current interests, among others, include digital forensics.

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 Senior Analyst at the NATO CCD COE. Jan Stinissen holds a Master in Law from the University of Utrecht, The Netherlands.

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.

AUTHORS

Lieutenant Colonel **Scott D. Applegate** is a United States Army Information Systems Management Officer with more than 21 years of leadership, management, communications and security experience. LTC Applegate holds two Masters Degrees, one in Military Studies and one in Information Technology and Assurance, and is currently pursuing a PhD in Information Technology with a focus on Cyber Conflict at George Mason University in Fairfax, Virginia. His current research interests include cyber conflict, cyber militias, security metrics, cyber security policy, information assurance and cyber law. LTC Applegate currently resides in Northern Virginia with his wife Sara and their two children.

Dr **Bill Boothby** retired as Deputy Director of Legal Services (RAF) in 2011 having obtained his Doktor Iuris from the Europa Universitaet Viadrina, Frankfurt (Oder) in Germany in 2009. He published his first book, Weapons and the Law of Armed Conflict, with OUP the same year and his second book, The Law of Targeting, appeared with the same publisher in 2012. He is currently working on a third volume looking at the future of conflict and related legal issues. He was a member of the Group of Experts that prepared the HPCR Manual on the Law of Air and Missile Warfare and was also a member of the Group of Experts, and of the Drafting Committee, of the Tallinn Manual. He teaches at Royal Holloway College, University of London, at the Australian National University, Canberra and at the University of Durham. He writes and presents regularly on a variety of international law issues.

Jeffrey Caton is President of Kepler Strategies LLC, a veteran-owned small business specializing in national security, cyberspace theory, and aerospace technology. His recent work includes space and cyberspace presentations to the Kazakhstan National Defense University supporting the Partnership for Peace Consortium. He is also an Intermittent Professor of Program Management with Defense Acquisition University. Prior to this, Mr. Caton served five years on the U.S. Army War College faculty including Associate Professor of Cyberspace Operations and Defense Transformation Chair. He served 28 years in the U.S. Air Force working in engineering, space operations, joint operations, and foreign military sales.

Gregory Conti is an Associate Professor and Director of West Point's Cyber Research Center. He holds a BS from West Point, an MS from Johns Hopkins University, and a PhD from the Georgia Institute of Technology, all in computer science. Dr Conti is the author of Security Data Visualization (No Starch Press) and Googling Security (Addison-Wesley) as well as over 60 articles and papers covering cyber warfare, online privacy, usable security, and security data visualization.

He has spoken at numerous security conferences, including Black Hat, Defcon, VizSec, HOPE, Interz0ne, ShmooCon, and RSA. His work can be found at www.gregconti.com

Michael J. Covington, Ph.D. is the product manager for Cisco's Security Intelligence Operations; he is focused on delivering technologies that monitor emerging threats and deliver actionable intelligence to next-generation networks and platforms.

As a security researcher with eight patents pending and as the author of numerous papers that have been published in leading academic conferences and journals, Dr Covington's research has explored formal access control modeling, cutting edge authentication techniques, and security approaches for pervasive computing environments. He is interested in bringing more intelligent systems to market that can assist with security-relevant decision-making, policy enforcement, and investigation efforts.

Dr Covington received his Ph.D. and MSCS degrees from the Georgia Institute of Technology's College of Computing in Atlanta, Georgia. He also holds a B.S. degree from Mount Saint Mary's College in Emmitsburg, Maryland.

Mr **Luc Dandurand** joined the NATO Communications and Information Agency in January 2009 where he performs R&D work in Cyber Defence and supports projects such as the NCIRC FOC. Prior to that, as a Signals Officer in the Canadian Forces, he was an analyst in the Directorate of Scientific and Technical Intelligence, he led the CF's Network Vulnerability Analysis Team, and he founded the CF Joint Red Team, responsible for assessing the security of CF networks by conducting controlled cyber-attacks. He then joined the Communication Security Establishment of Canada to lead a team that prototyped novel solutions in Cyber Defence.

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.

Alessandro Guarino is an experienced information security professional and independent researcher. He is CEO of StudioAG, a consulting firm based in Italy whose services were used by the industry and the public sector. He holds a degreee in

Industrial Engineering and is completing work on a thesis on Information Security Economics for a degree in Economics and Business. He is an ISO active expert in JTC 1/SC 27 (IT Security techniques commitee) and contributed in particular to the development of cybersecurity and digital investigation standards. He represents Italy in the CEN-CENELEC-ETSI Cybersecurity Coordination Group.

Kim Hartmann studied Computer Science and Mathematics at the Royal Institute of Technology, Stockholm, Sweden and at Otto von Guericke University Magdeburg, Germany. Kim Hartmann specialised in computer security and mathematical modelling, worked on protocol security analysis, mathematical computer security modelling, computer security risk assessment and risk analysis of critical network infrastructures. Her research interests are secure network design principles, risk analysis and assessment of networks, network components and protocols. Since 2011, Kim Hartmann has been employed at the Institute of Electronics, Signal Processing and Communication at Otto von Guericke University Magdeburg, Germany.

Dr **Jorge Lopez Hernández-Ardieta** holds a B.Sc. and M.Sc. in Computer Engineering from the University Autonoma of Madrid, and a Ph.D. in Computer Science from the University Carlos III of Madrid (UC3M). Dr Lopez is the Head of the Cybersecurity Research Group at Indra (Spain), where he leads research and innovation in cybersecurity and cyberdefence. In addition, he is Part Time Professor and Affiliate Researcher in the Computer Security (COSEC) Lab at UC3M. He participates in standardisation efforts and high-level consultancy activities, being a member of EDA IAP4, NIAG, ISO/IEC JTC1 SC27 IT Security techniques (Chairman of Spanish WG3 Security evaluation, testing and specification), CEN/TC 224, IEEE and IETF.

Janine Hiller is a Professor of Business Law at Virginia Tech in Blacksburg, Virginia, USA. Hiller has published legal and interdisciplinary research in the area of electronic privacy and security. She has organized a conference addressing the public-private partnership for privacy and security, contributed to ABA publications on privacy and cybercrime, and received a National Science Foundation grant to study legal and technical means to protect children's online privacy. She served in the Fulbright-Lund Distinguished Chair of Public International Law in Sweden in 2010, and is a member of the Virginia Tech Hume Center for National Security and Technology.

Emilio Iasiello is the Chief Threat Analyst at iSIGHT Partners, a global cyber intelligence firm, supporting federal and commercial entities to manage cyber risks, understand their threat environment, and help prioritize their investments against those threats impacting their business or mission. He has worked in cyber threat

analysis since 2002 both as a government contractor and a government civilian with the Department of State and the Department of Defense, respectively. Emilio has written papers on the development of a new cyber threat analytic methodology, and on the IT Supply Chain.

Barry Irwin has a PhD in Computer Science, and holds the CISSP certification. He is currently an Associate Professor and has headed the Security and Networks Research Group (SNRG) since its founding in 2003 in the Department of Computer Science at Rhodes University, South Africa. His research interests include network traffic analysis, data visualisation, web based malware, botnets and anti-phishing tools.

Dr **Gabriel Jakobson** is Chief Scientist of Altusys Corp., a consulting firm specializing in situation management technologies for defense and cyber security applications. Dr Jakobson is vice-chair of the Tactical Communications and Operations Technical Committee of IEEE ComSoc, chair of the IEEE ComSoc Sub-Committee on Situation Management, member of IEEE Technical Committee on Security and Privacy in Complex Information Systems, and Distinguished IEEE Lecturer. Dr Jakobson holds Honorary Degree of Doctor Honorius Causa from Tallinn University of Technology, Estonia, and is General Chair of the Conference on Cognitive Methods of Situation Awareness and Decision Support (CogSIMA 2011-2013. He received PhD degree in Computer Science from the Institute of Cybernetics, Estonia.

Kaarel Kalm is a graduate student at the Department of Security and Crime Science, University College London were he is specializing in countering organised crime and terrorism. His research interests are asymmetric conflicts and non-state agent networks. He interested in how conflicts with non-military and non-state agents (organised crime, hacker networks, religious groups, covert state-sponsored networks) influence state deterrence structures and policies. Prior to that, he held different civil service posts in Estonian Government Agencies. He also has a BA in Political Science and MA in International Relations from the University of Tartu, Estonia.

Igor Kotenko is a professor of computer science and Head of Research Laboratory of Computer Security Problems of the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. He graduated with honors from St.Petersburg Academy of Space Engineering and St.Petersburg Signal Academy, obtained the Ph.D. degree in 1990 and the National degree of Doctor of Engineering Science in 1999. He is the author of more than 150 refereed publications, including 12 study books and monographs. Igor Kotenko has a high experience in the research on computer network security and participated in several projects on developing

new security technologies. For example, he was a project leader in the research projects from the US Air Force research department, via its EOARD (European Office of Aerospace Research and Development) branch, EU FP7 and FP6 Projects, HP, Intel, F-Secure, etc. The research results of Igor Kotenko were tested and implemented in more than fifty Russian research and development projects. The research performed under these contracts was concerned with innovative methods for network intrusion detection, simulation of network attacks, vulnerability assessment, security protocols design, verification and validation of security policy, etc. Igor V. Kotenko is a laureate of the St. Petersburg Government award for outstanding scientific achievements in the field of science and technology in 2012, a laureate of the program "Outstanding Scientists. Doctors of Sciences of the Russian Academy of Sciences" in 2007-2008, and a winner of many grants of the Public Science Support Foundation, the Russian Foundation of Basic Research, the Program of fundamental research of the Department for Nanotechnologies and Informational Technologies of the Russian Academy of Sciences and several State contracts, a winner of the best works in the field of artificial intelligence in 2004-2006. The main results of his research from 2002 to 2011 have been included many times in the list of major scientific achievements of the Russian Academy of Sciences. He has chaired several conferences and workshops, and serves as editor on multiple editorial boards.

Owen McCusker has been researching and developing the use of Network Behavioral Analysis (NBA) in cyber defense. In 2006 DHS S&T funded a NBA-based fusion prototype. In 2009, this work has transformed into establishing a Cyber Behavior Analytics capability. Mr McCusker has been invited to a number or workshops and symposiums including the 2009 National Cyber Leap Year NITRD, 2010 NATO R&T Cyber Defense Workshop in Estonia, and in 2011 Global Cyber-physical Supply Chain Summit in Wales with MIT. Mr McCusker holds a Masters in Computer Science from Rensselaer Polytechnic Institute and is an SME for MIT's Geospatial Data Center.

Daniel Plohmann studied Computer Science at the University of Bonn. Since 2010, he is a PhD student as well as a security researcher of the Cyber Defense Research Group at Fraunhofer FKIE in Wachtberg, Germany. His main research field is reverse engineering with a focus on malware analysis and botnet mitigation. For more information on his work and his publications, please visit: http://pnx.tf.

Jody Prescott's research and writing focus on three major evolving national security topics: gender, alternative energy, and cyber. His recent work includes:

NATO Gender Mainstreaming and the Feminist Critique of the Law of Armed Conflict, Georgetown Journal on Gender and the Law (Winter 2013)

Ridgelines and the National Security Implications of Commercial Wind Energy Development in Vermont, Vermont Journal of Environmental Law (Fall 2012)

Direct Participation in Cyber Hostilities: Terms of Reference for Like-Minded States?, Proceedings, 4th International Conference on Cyber Conflict, NATO Cooperative Cyber Defence Center of Excellence (Summer 2012)

LTC **David Raymond** is an Armor officer in the United States Army. He is assigned as an Assistant Professor in the Department of Electrical Engineering and Computer Science at the United States Military Academy, West Point. LTC Raymond has a Master's degree in Computer Science from Duke University and a Ph.D. in Computer Engineering from Virginia Polytechnic and State University. He teaches senior-level computer networking and cyber security elective courses at West Point and conducts research on information assurance, network security, and online privacy.

Prof. **Alexander V. Smirnov** is head of Computer Aided Integrated Systems Laboratory at St.Peterburg Institute for Informatics and Automation of the Russian Academy of Sciences - SPIIRAS (1994), Deputy-Director for Research (1996). He received his Ph.D (1984) and Dr.habil. (1994). He has been involved in projects sponsored by Ford, Nokia, US DoD, European Research Programs (Information Society Technologies, Esprit, Eureka/Factory, etc.), and Russian agencies in the areas of distributed intelligent systems, ontology management, intelligent decision support systems, etc. He is a member of IEEE SMC TC on Self-Organized Distributed and Pervasive Systems. He published more than 300 research papers.

John P. Sullins is an associate professor of philosophy at Sonoma State University in California where he has taught since 2004. He received his PhD in 2002 from the Philosophy, Computers, and Cognitive Science program at Binghamton University in New York. His current research and publications involve the study of computer ethics, malware ethics, and the analysis of the ethical impacts of military and personal robotics technologies. He is the 2011 recipient of the Herbert Simon Excellence in Research award from the International Association of Computers and Philosophy.

Shannon Vallor, Ph.D. is Associate Professor of Philosophy at Santa Clara University in California, where she teaches the philosophy of science and technology as well as

engineering ethics. Her primary research project concerns the impact of emerging technologies on the cultivation of moral and intellectual virtues, and her research has been published in journals such as Ethics and Information Technology, Philosophy of Technology and Techne, as well as in Springer's upcoming edited collection, Ethics of Information Warfare. She is currently working on a book, 21st Century Virtue: An Ethical Framework For Living Well With Emerging Technologies. Professor Vallor is a member of the Executive Board of the international Society for Philosophy and Technology, a Scholar of the Markkula Center for Applied Ethics and a steering member of SCU's Center for Science, Technology and Society.

CCDCOE

NATO Cooperative Cyber Defence
Centre of Excellence
Tallinn, Estonia

Contact & Feedback
publications@ccdcoe.org
CFP1326N-PRT

ISBN 978-9949-9211-4-0