CERIAS Tech Report 2016-8

The Application of natural Language Processing to Open Source Intelligence for Ontology Development in the Advanced Persistent Threat Domain

by Corey T. Holzer
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

THE APPLICATION OF NATURAL LANGUAGE PROCESSING TO OPEN SOURCE INTELLIGENCE FOR ONTOLOGY DEVELOPMENT IN THE ADVANCED PERSISTENT THREAT DOMAIN

by

Corey T. Holzer

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Technology

West Lafayette, Indiana

December 2016

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. J. Eric Dietz, Chair

Department of Computer and Information Technology

Dr. Biajian Yang

Department of Computer and Information Technology

Dr. Dongyan Xu

Department of Computer Science

Dr. John A. Springer

Department of Computer and Information Technology

Approved by:

Dr. Kathryne A. Newton

Head of the Departmental Graduate Program

This work is dedicated to my wife, Rachael, and our children who selflessly support me in all my endeavors; allowing me to pursue my dreams.

ACKNOWLEDGMENTS

There are several people that I would like to acknowledge for their support of this endeavor.

My wife, Rachael, who has shown me unwavering support for the 23 years that we have been together. I could never have accomplished what I have without her. I also want to thank my children Christina, Gwendolyn, Alannah, Joshua and Destini. Their faith in and love for me is a constant source of strength each and every day.

My Committee Chair, Dr. Eric Dietz for his guidance and mentorship throughout my dissertation and as my advisor while attending Purdue. My committee Dr. Yang, Dr. Xu, and Dr. Springer for their insights and guidance throughout the dissertation process. My thanks to Dr. Eugene "Spaf" Spafford, Director Emeritus of the CERIAS Program, because without his urging and guidance I might not have had this opportunity in the first place. To Mrs. Marlene Walls whose help and guidance through the administrative aspects of graduate life at Purdue saved me many hours of work and needlessly running around campus.

I want to acknowledge U.S. Army Cyber Command for selecting me as one of their FY15 Army Cyber Scholars. They saw in me the potential as a Cyber leader for the Army and demonstrated their confidence in that potential by selecting me. What I have accomplished over the last two years demonstrates that their confidence was well founded.

I wish to acknowledge Colonel Timothy Frambes, Lieutenant Colonel Charles D. (Dean) Smith, and Tommie L. Walker, Lieutenant Colonel (USA Retired) whose mentorship and leadership has been invaluable. I continually seek to emulate and incorporate their leadership styles into my own. I also want to thank James Lerums Colonel (USA Retired), Major Patrick Glass, and Captain Chris Baker fellow students and colleagues in military life. You were always there willing to give me the azimuth check as I made my way through my scholarly pursuits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

APT	Advanced Persistent Threat
C2	Command and Control
CND	Computer Network Defense
CNE	Computer Network Exploitation
CoCA	Corpus of Contemporary American English
CSS	Creative Style Sheets
FTP	File Transfer Protocol
GREAT	Kaspersky Lab Global Research and Analysis Team
HTTP	Hypertext Transfer Protocol
IDS	Intrusion Detection System
NCHC	National Center for High Performance and Computing
IMF	Information Warfare Monitor
NER	Named Entity Recognition
NLP	Natural Language Processing
OSINT	Open Source Intelligence
OWL	Ontology Language
OWL2	W3C's Ontology Language 2.0
PoS	Parts of Speech
RAT	Remote Access Toolkit
SaaS	Software as a Service
SCADA	Supervisory Control and Data Acquisition
SCP	Secure Copy Protocol
SFTP	Secure File Transfer Protocol
SIEM	Security Information and Event Management
SSH	Secure Shell
SSL	Secure Socket Layer
TTP	Tactics, Techniques, and Procedures
VPN	Virtual Private Network
W3C	World Wide Web Consortium's Web

ABSTRACT

Author: Holzer, Corey, T. Ph.D.
Institution: Purdue University
Degree Received: December 2016
Title: The Application of Natural Language Processing to Open Source Intelligence for Ontology Development in the APT Domain
Major Professor: J. Eric Dietz.

Over the past decade, the Advanced Persistent Threat (APT) has risen to forefront of cybersecurity threats. APTs are a major contributor to the billions of dollars lost by corporations around the world annually. The threat is significant enough that the *Navy Cyber Power 2020* plan identified them as a "must mitigate" threat in order to ensure the security of its warfighting network.

Reports, white papers, and various other open source materials offer a plethora of information to cybersecurity professionals regarding these APT attacks and the organizations behind them but mining and correlating information out of these various sources needs the support of standardized language and a common understand of terms that comes from an accepted APT ontology.

This paper and its related research applies the science of Natural Language Processing Open Source Intelligence in order to build an open source Ontology in the APT domain with the goal of building a dictionary and taxonomy for this complex domain.

INTRODUCTION

Sophisticated hackers today employ complex operations in order to achieve their goals. They use intelligence gathering techniques to collect information about a potential target. They study these target networks and organizations to find weaknesses that they can use to their advantage. They execute attacks in a precise and careful manner in order to remain hidden from their targets. These attacks can even go dormant for months at a time or be executed over an extended period so as not to trip standard cybersecurity measures employed by their victims.

Conversely, intelligence about potential threats aids cybersecurity professionals in the defense of their networks as well. Like military forces on the battlefield these professionals need knowledge about those attacking their network and the means of attack that these adversaries employ in order for these professionals to better defend their network.

The rise in frequency and complexity of cyberattacks means that cybersecurity professionals are fighting a pitched battle and the intelligence they have is their best means of dealing with the threats that infiltrate their network and remain hidden within it in order to steal data or to do harm. Over the last decade these complex attacks known collectively as Advanced Persistent Threats (APTs) presented some of the greatest challenges to network and national security as the perpetrators executed these attacks against both the private and public sector.

This paper represents the author's research into the development of an APT ontology through the mining of open source intelligence. By employing the science of Natural Language Processing, it is the goal of this research to build a useful tool which will enable cybersecurity professionals to learn about these attacks

This chapter breaks down into seven sections. The first provides a brief background regarding APTs over the last decade. The second addresses the scope of the research. The third section lays out the significance of the research. The fourth provides the Statement of Purpose for the research. The final three sections outline the assumptions, limitations, and delimitations that could potentially affect the work at hand.

Background

Nearly one decade ago the term APT was coined. It described an emerging and complex form of cyberattacks that were engineered to steal data over an extended period of time while evading detection through a variety of techniques including small data transfers, use of non-malicious applications for malicious ends, etc. (Brill, 2010). These attackers were not seeking to do drastic or violent harm to an infected network. Instead they laid in wait eavesdropping on the network waiting for the right opportunity to present itself.

In his seminal work on APTs, Eric Cole likened these attackers and their attack to shoplifters. Cole notes, "At the point of entry the legitimate customer and shoplifter look identical. … If that shoplifter is only in the store for 5 min[sic], the store has less than 5 min to detect and deal with the problem (Cole, 2012, p. 7)." To carry his analogy one step farther, if the shoplifter browses around the store doing little to cause suspicion and waiting for an ideal time (i.e. when the store owner is busy with a lunchtime crowd) to pilfer the item he desires the shop owner or the shop workers may never be any the wiser.

According to a market research report in 2015, the APT protection market generated $3.9B USD in 2015 (PR Newswire, 2015). The figure is based on expenditures on items including Sandboxing, Endpoint Protection, Forensic Analysis, Security Information and Event Management (SIEM), Intrusion Detection Systems (IDS), Intrusion Protection Systems (IPS), Firewalls, etc. (PR Newswire, 2015). The same report by MarketsandMarkets estimated that this expenditure would more than double to $8.7B USD in 2020 (PR Newswire, 2015). The cost of APT attacks will be discussed in greater depth in Chapter 2.

Scope

The research focused on the development of an ontology in the APT domain. This effort will not attempt to create new signatures or update existing signatures nor will it test to see if modified signatures can improve detection. For the reasons of timeframe and skillset, such an effort is beyond what the researcher would be able to complete to satisfy the requirements for his degree.

Significance

In a world where so much of our commerce, infrastructure, and even national defense depends on the communication pathways provided by the internet, failure to detect and mitigate the threat represented by APTs could impact businesses, civilian infrastructure, and potentially national security of all nations. Current Cyber Network Defense measures are unable to reliably and regularly detect APTs that compromise both commercial and government networks. Part of the problem is incomplete understanding of these APTs and to see the various components of a single APT attack as parts of the larger whole.

The goal of this research is to build a comprehensive ontology based on information resources found in the open source domain. This work uses both open source intelligence and open source tools in the fields of natural language processing and ontology development.

Statement of Purpose

Upon its conclusion, this research will produce an ontology, built to OWL2 specifications in Protégé, for use in the APT culled from Open Source Intelligence (OSINT) resources found in the public domain of the World Wide Web.

Assumptions

The design of this study is based upon the following assumptions:

- An APT domain ontology does not already exist in the public domain.
- The pool of Open Source Intelligence (OSINT) is significantly large enough to create a useful ontology pertaining to the APT domain.

Limitations

The design of this research incorporates the following limitations:

- **OSINT –** The pool of information used for this research will be limited to what information about APTs exists in the public domain.
- **Digital OSINT –** The research will only include digitally based OSINT.

- **Skillset –** While I have some programing experience it is not significant enough nor do I possess the expertise in malware forensics needed to properly develop signatures for use by existing cybersecurity systems to detect and/or mitigate an APT attack.

Delimitations

Factors outside of the control of the researcher creates the following delimitations:

- **Time –** The Army has afforded the author a finite amount of time, two (2) years, to complete all the requirements of a graduate degree.
- **Machine Processing -** The identification of ontological terms will be performed by software. Therefore, the amount of data that can be processed will be limited by how much data can be obtained and by the processing power of the systems being employed.
- **Open Source Intelligence –** What information about APTs is available in the public domain is limited both by what cybersecurity companies wish to keep proprietary and what governments identify as classified information.
- **Exclusion of other OSINT –** Limited assets for scanning and limited time for manually processing non-digital OSINT means that the research must exclude non-digital OSINT.

Summary

In this chapter, you were provided with an overview of rise of the APT over the past decade as well as the projected growing costs of protecting systems and networks from these attacks over the next 5 years. We reviewed the significance of the research and the deliverables this researcher expects to provide upon completion. Finally, it addressed assumptions, limitations, and delimitations impacting the project at hand. The next chapter consists of a review of relevant literature pertaining to the current field of research.

REVIEW OF RELEVANT LITERATURE

With the overview of the research complete, we will now move into a review of the relevant existing literature. This chapter covers eight areas of literature relevant to the current research effort. The first three sections provide context for the Advanced Persistent Threat by defining APTs, examining the APT attack in detail, and through an exploration of the decade long history of APT attacks and the costs associated with them. The fourth section defines the present challenges of detecting APT attacks as well as exploring the countermeasures used by malicious individuals in order to evade detection.

The last four sections examine concepts, disciplines, and related research in the APT domain which is relevant to this research. The first addresses the concept of Open Source Intelligence. The second explores the science of Natural Language Processing. The last defines ontologies. This chapter concludes with an exploration of research in the field of Advanced Persistent Threats, with particular focus on the development of APT ontologies.

Defining Advanced Persistent Threats

The United States Air Force first coined the phrase Advanced Persistent Threat in 2006 (Arsene, 2015; Ask et al., 2013; Bejtlich, n.d.). They created the term in order to facilitate discussions about the characteristics, activities involved, and the classification of these types of attacks (Ask et al., 2013).

It is proper to begin this research endeavor by defining how Advanced Persistent Threat will be used for our purposes. The threat is *Advanced* in that adversaries can employ tactics that cover the full spectrum of cyber-attacks (Ask et al., 2013; Bejtlich, n.d.; Bodeau, Graubart, Heinbockel, & Laderman, 2014; Cole, 2012). It is *Persistent* in that he is not an opportunistic hacker searching for easily infiltrated systems (Bejtlich, n.d.). Instead, the persistent attack is one that will operates over an extended period of time with a pre-determined target and desired end state for the cyber-attack (Bejtlich, n.d.; Bodeau et al., 2014; Radzikowski, 2016).

APT as an Attack

The APT attack involves multiple methods, tools, and techniques used in a sophisticated complex manner in order to compromise the target and achieve what is usually a long-term objective (Chandran, P, & Poornachandran, 2015; ISACA, n.d.-a). The attacks are referred to as complex because they involve multiple forms of attack in order to compromise a target. A single APT can include a combination of social engineering of human targets, a phishing campaign as a call to action by a victim, and the use of malware to gain access and elevated permissions on target systems (Ask et al., 2013; ISACA, n.d.-b, n.d.-c).

The actual phases of an APT attack will be discussed in a later section. The discussion will follow the framework outlined by Lockheed Martin in its Cyber Kill Chain (Ask et al., 2013). It is important to note that as the APT attack phases should not be considered as discrete events where one phase ends and the next begins (Chandran et al., 2015). Keep in mind that there can be overlap as the attack propagates across the target network. It is also important to understand that while we use the term persistent it is not intended to mean that the attack is constantly active or that connections between the APTs' C2 server and compromised hosts is constant (Ask et al., 2013; Hutchins, Cloppert, & Amin, 2011; Raj, Chezhian, & Mrithulashri, 2014).

The sophistication and complexity of APT attacks make it hard for organizations to recognize one particular element as being only one piece of a larger plan (Armerding, 2012; Ask et al., 2013). In this game of cat and mouse the attacker has the advantage of having unlimited time, resources, the victim organization's prioritization of its business processes, and less fear of prosecution when the attack takes place across international borders (Auty, 2015).

APT as an Organization

In addition to describing the attack, the term APT is used to describe the organizations that execute these attacks. In this context the discussion focuses on organizations that are well funded, well organized, and patient (Cole, 2012). They can infiltrate a network and remain hidden while monitoring it for a specific target or data to

exfiltrate (Bodeau et al., 2014). Their goal is stealthy execution instead of the kind of attack that draws attention to the person or persons committing the crime.

While it is understood that those responsible for APT attacks are well organized and that they possess significant funds, cybersecurity professionals must not confuse this with meaning that they are sponsored by state actors (Ask et al., 2013). Cybersecurity professionals attribute some APTs to state sponsored actors but state sponsorship in not a required component of the definition (Bejtlich, 2013; Mandiant, 2013). With APT defined let us explore the costs, both in financial and national security terms, associated with the threat.

Phases of the APT Attack

With the definitions complete, let us delve into the several phases that an APT attack employs in order to obtain the attacker's desired end state. One of the more widely accepted description of the APT Attack comes from Lockheed Martin. The "Lockheed Martin Cyber Kill Chain" breaks the attack into seven phases. We will use their model as a means of discussing Lockheed Martin's definition of each phase. The seven phases are as follows:

- Reconnaissance
- Weaponization
- Delivery
- Exploitation
- Installation
- Command and Control
- Actions on Objective

Additionally, the descriptions will include elements that other cybersecurity professionals and professional organizations include when discussing the APT attack.

Reconnaissance

Reconnaissance is the selection and identification of the desired target. In this stage the APT is footprinting the target organization and collecting information including but not limited to names, positions, email addresses, physical locations, operating

systems, etc. (Hutchins et al., 2011). Through this information gathering process the APT determines who has ownership of the desired information that they seek to steal (2013). The APT will determine which employee to compromise as well as a potential means for doing so.

Weaponization

In the Weaponization phase, the APT puts together the code that they will use to compromise a target system (Hutchins et al., 2011). This will often involve the use of existing and proven code but, if needed, APTs will adapt or modify code in order to address a specific configuration or defensive challenge (Ask et al., 2013; Cole, 2012; Hutchins et al., 2011). When using code designed for the specific target, the code has no anti-virus signature which the target company might use to detect it (Websense, 2011).

Delivery

In the Delivery phase, the APT transmits the weapon to the targeted system (Hutchins et al., 2011). Lockheed Martin identifies the most common delivery methods as email attachments, websites and removable media. In addition to those three, Ask, et.al. (Ask et al., 2013) identified social media as another means for launching at attack against an individual within the target organization. For the attack to move beyond this phase, the targeted individual most click on the link, attachment, or application for the attack to move into the next phase (Auty, 2015).

Exploitation

Exploitation involves compromising the host machine on the network. It is where the weaponized tool is triggered (Hutchins et al., 2011). The exploitation can be of a flaw in the operating system or an individual application on the host (Ask et al., 2013; Hutchins et al., 2011).

Installation

The next phase of the attack is the Installation phase. Installation refers to the installation of a Remote Administration Tool (RAT) or backdoor that the APT can use to

gain control of the target's computer (Ask et al., 2013; Hutchins et al., 2011). Once the victim triggers the malicious code (e.g. by clicking the malicious link, opening the infected file, or visiting the compromised site, etc.) the code reaches back to its Command and Control (C2) server and provides the attacker with useful information about the target network's environment that could be useful in executing the later stages of the APT attack (Ask et al., 2013). Once installed the RAT can also lay dormant until the C2 server connects to it (Ask et al., 2013; Sikorski & Honig, 2012).

Command and Control

The Command and Control phase begins once the infected host beacons the C2 server (Hutchins et al., 2011). Attackers need to maintain access to the victim's network means that each communication with a compromised system (Auty, 2015). During this phase the APT will seek to obtain elevated privileges on the system and will install additional software to facilitate the attack (i.e., encryption) on compromised system and network (Ask et al., 2013). While the initial installation is achieved by software designed to exploit a zero-day vulnerability, the additional software is likely to be commonly known software that may even be approved to operate on the network for legitimate activities (e.g., SSH, SecureFTP, etc.) (Ask et al., 2013).

Actions on Objective

The final stage in Lockheed Martin's APT Kill Chain is the Actions on Objective phase. During this phase the APT is actively going after the data that they originally identified as their target (Hutchins et al., 2011). The APT uses previously installed software to determine the network layout including, but not limited to, mapping the hosts of networked drives, database servers, domain controllers, PKI, etc. (Ask et al., 2013). The goal here is to footprint the network and to establish a network account and elevate the privileges for that account (Ask et al., 2013). During this phase, the APT will also seek to compromise more hosts in order to strengthen its foothold in the target network. The extraction of the target data may also be accomplished using custom encryption and/or tunneling within other protocols to hide the data from security professionals (Websense, 2011).

Conventionally, malware will remove itself once its task is complete or it is discovered and removed by antivirus software (Ask et al., 2013). The APT, however, is designed to stay invisible. It maintains persistence by reaching back to the C2 server for updates to the malicious code (Ask et al., 2013). Changing code enables the APT attack to avoid detection. Mandiant's APT Attack model includes cleanup as part of this phase (Mandiant, 2013; Saarinen, 2013). However, it is more likely that the APT will leave some software in place in order to facilitate quicker access if the adversaries wish to exfiltrate more information in the future. The security firm Mandiant has data demonstrating that a group identified as APT1 has left software in place to re-access a target network months, and even years, later (Bejtlich, 2013; Mandiant, 2013; Raj et al., 2014).

History of APTs

With an understanding of the APT attack established, we will next look at some examples of cyber-attacks that were qualified as APTs. By no means is this intended to be an all-inclusive list. It is intended to demonstrate the variety of attack elements and the variety of targets.

Shady Rat

With earliest evidence indicating that this APT collected data in mid-2006, it is possible that it was stealing data even earlier than the logs provide (Alperovitch, 2011). Evidence collected by McAfee indicates that, unlike other APTs discussed here, this APT was used against a wide range of individuals and organizations in multiple industries. Initial installation took place via a spear-phishing email. The attachment triggered the download and installation of malware that, in turn, created a backdoor communication channel with its C2 server. In four of 71 instances where Shady RAT gained control of a target system, it remained persistent for 24 or more months (Alperovitch, 2011).

Night Dragon

This attack targeted the Global Energy Business Community (McAfee, 2011). Commencing in 2009, Night Dragon employed social engineering and spear-phishing

attacks to exploit vulnerabilities and compromise Microsoft Active Directory machines. The initial targets were extranet web servers and then internal desktops and servers to gain elevated permissions within the hosts and target network. Finally, the APT harvested sensitive proprietary and project-financing related information sending that information back to C2 servers on the Internet (McAfee, 2011).

Poison Ivy Attack on RSA

The APT identified two specific independent groups of RSA employees and crafted a spear-phishing campaign tailored to the target employees' job functions (RSA FraudAction Research Labs, 2011). The email contained a spreadsheet which executed code that leveraged an Adobe Flash vulnerability in order to inject a Poison Ivy RAT which, in turn, established a hard to connect reverse connection to the APT's C2 servers. The data stolen included RSA's SecureID two-factor authentication products (Ashford, 2011). Executive Chairman Art Coviello issued an open letter to customers in which he acknowledged that the stolen information "could potentially be used to reduce the effectiveness of a current two-factor authentication implementation as part of a broader attack (Ashford, 2011; Coviello, n.d.)."

Icefog

This APT attack has been used numerous times starting in 2011 with most of the attacks targeting organizations in Japan and South Korea (Kaspersky Lab Global Research and Analysis Team, 2013). Kaspersky Lab's Global Research & Analysis Team (GREAT) researched the attacks and determined that their targets were supply chains of "government institutions, military contractors, maritime and ship-building groups, telecom operators, satellite operators, industrial and high technology companies and mass media (Kaspersky Lab Global Research and Analysis Team, 2013)."

The APT achieved their initial insertion through spear-phishing campaigns and attachments which exploited known vulnerabilities in the host's operating system (Kaspersky Lab Global Research and Analysis Team, 2013). GREAT identified at least 6 variations in the manner in which Icefog exfiltrates data (Kaspersky Lab Global Research

and Analysis Team, 2013). They are as follows (with designations established by GREAT) (Kaspersky Lab Global Research and Analysis Team, 2013):

- **The "old" 2011 Icefog —** sends stolen data by e-mail
- **Type "1" "normal" Icefog —** interacts with C2 servers
- **Type "2" Icefog —** interacts with a script-based proxy server that redirects attacker commands
- **Type "3" Icefog —** observed to use a certain kind of C2 via a different means of communication
- **Type "4" Icefog —** another C2 variation with a different means of communication
- **Icefog-NG —** communicates by direct TCP connection to port 5600

Stuxnet

In 2010 this worm was weaponized with the specific goal of impacting systems that run Supervisory Control And Data Acquisition (SCADA) engineered by Siemens (Kushner, 2013). The worm was initially uploaded via USB to a Windows workstation and started spreading across the target network without impacting any systems that did not run the SCADA software. Once it entered a machine where the SCADA was present it would connect with its C2 server and receive software updates (Kushner, 2013). The worm then compromised the system and began gathering information in order to get the elevated permissions needed to take control of centrifuges making them fail. The software would also provide false information back to the user giving the appearance that everything was functioning normally (Kushner, 2013).

GhostNet/Shadows in the Cloud

In a collaborative effort, Information Warfare Monitor (IMF) and Shadowserver Foundation (2010) documented the complex network of systems used to conduct cybercrime and cyber espionage operations against unsuspecting targets in the government, academic, and corporate sectors. The research built upon a research effort conducted by an Information Warfare Monitor partner, SecDev, titled *GhostNet* which they performed in 2009 (Bradbury, 2010; Information Warfare Monitor & Shadowserver

Foundation, 2010). SecDev initiated the *GhostNet* investigation at the behest of the Office of the Holiness the Dali Lama which was a victim of the APT attack (Bradbury, 2010).

The IMF and Shadowserver (2010) determined the following:

- C2 infrastructure leveraged cloud-based social media services in order to compromise unsuspecting targets.
- The complex network of compromised systems employed C2 servers, malware, botnets, and drop sites in order to compromise targets and to exfiltrate data undetected.
- Stolen data included classified and sensitive documents as identified by classified markings on multiple documents.
- Hackers exfiltrated data from 44 systems across nine countries.

The investigation determined that the entire system required four different types of hacking tasks in order to successfully complete the task of stealing data (Information Warfare Monitor & Shadowserver Foundation, 2010):

- **Malware Authors** to develop the malware that compromised target systems.
- **Website Crackers** to maintain the malicious websites
- **Envelope Stealers** who are individuals who steal username and password combinations on compromised networks.
- **Virtual Asset Stealers and Virtual Asset Sellers** who possess an understanding of what data has value in the underground or criminal economy.

New York Times Attack

In 2013, the New York Times announced that their network was compromised through the installation of malware which led to the extraction of the network's user database (Perlroth, 2013). In its 2014 "M-Trends: Beyond the Breach" Report, Mandiant stated that the suspected APT group took specific steps to change their cyber operations following the disclosure in order to "better hide its activities (Mandiant, 2014, p. 20)."

Trojan.APT.Seinup

Discovered in 2013, this Trojan targeted Google Docs. The APT attack leveraged this legitimate cloud based Software as a Service (SaaS) in order to leverage the legitimate Secure Socket Layer (SSL) of Google Docs to protect their malicious communications (Naval, Laxmi, Rajarajan, Gaur, & Conti, 2014).

With this brief sample the reader can see that detecting APTs is a challenge for cybersecurity professionals. Next, we will examine some of the technical reasons that make it so difficult.

The Cost of the APT Threat

On a regular and increasingly frequent basis companies, government organizations, and industries are reporting breeches of their network and the extraction of thousands if not millions of data records. Despite the security measures these organizations put in place to ensure the security of the data customers provide to them. For example, in 2011 RSA spent $66 million USD to undo the damage caused by an APT attack (TrendLabs, n.d.). In a 2015 study by the Ponemon Institute, the researcher estimated that it can cost up to $161 each record lost (Ponemon Institute, 2015). When one considers that some APT attacks can compromise millions of users' records the cost could potentially bankrupt businesses.

The threat is not limited to consumer market. The US government's Office of Personnel Management reported a data breech in 2014 which involved 25,000 or more personnel records of government employees (Bisson, 2015). Breeches like this could present a risk to national security. The Stuxnet attack in 2010 had the potential of causing significant damage to nuclear facilities which could place lives and national infrastructures at risk as well (Damballa, 2010; Kushner, 2013). It is for this reason that the U.S. President issued an Executive Order in 2013 calling for the development of a Cyber Resiliency Framework (United States. The White House, 2013). It also prompted the U.S. Navy to declare APTs as a "must mitigate" threat (Card & Rogers, 2012).

Understanding the Challenge of Detecting APTs

With the previous examples providing a context for understanding how APTs function we can now address what challenges exist in detecting APTs. While APTs can employ a variety of known attack elements (e.g., phishing, malware, etc.) which can be detected by current security measures, attackers are still able to execute their attacks unnoticed. The question for security professionals, therefore, how are these attackers employing these detectable tools in a manner that leaves them undetected.

Challenges Specific to APT Detection

Conventional means of intrusion detection often fail to detect APTs because they are implemented to mitigate risks associated with automated viruses and worms, not the focused manually-operated attack of an APT (Hutchins et al., 2011). In its annual *M-Trends*, Mandiant (Mandiant, 2010), estimated that only 24% of APT malware is detected by security software. This is due to multiple factors. The target organization's decision not to inspect outbound packets (Ask et al., 2013; Auty, 2015; Villeneuve & Bennett, 2012). Data is encrypted (Ask et al., 2013; Villeneuve & Bennett, 2012). The affected machines send data to a trusted source (Villeneuve & Bennett, 2012).

Anti-detection Methods Employed with Malware

As discussed in the previous section, APTs commonly employ malware as a means to gain their initial foothold into a target network and to gain elevated permissions on host machines. Malware comes in many forms including Trojan horses, worms, rootkits, scareware, spyware, and viruses (Egele, Scholte, Kirda, & Kruegel, 2012; Sikorski & Honig, 2012). Regardless of which family the malware belongs to the software is designed to remain undetected on an infected system (Brand, Valli, & Woodward, 2010).

Malware developers must overcome the various forms of detection the forensics analysts use to reveal the presence of malware as described above. Malware developers have varied methods at their disposal to defeat detection. These means fall into several areas.

Anti-emulation

Malware developers employ techniques which detect that their malware is running in a virtual environment and the malware will either stay dormant or use deception code to provide a false signature to forensic experts trying to dissect the malware (Brand et al., 2010).

Anti-online

Companies offer third party malware analysis services online. However, there are limits to how well these services work because the online environment may not match conditions to trigger the malware or it may act differently than it would in a real-world network (Brand et al., 2010).

Anti-Analysis

Anti-analysis refers to changing the code such that it becomes harder to read during the analysis process (Brand et al., 2010; Shosha, James, Hannaway, Liu, & Gladyshev, 2012). These techniques target the way analysis is conducted. Code is deceptively transformed such that the analysis tools cannot establish a signature for the malware (Brand et al., 2010). De-obfuscation methods of analysis fail because that analysis happens on the files whereas the malware's transformation back into identifiable malicious code happens in memory as part of the unpacking process (Egele et al., 2012).

Anti-hardware

Malware developers can use check to determine whether the malware is being analyzed based on signatures of CPU usage and register usage during the debugging session (Brand et al., 2010).

Anti-debugger, Anti-disassembler, and Anti-Tools

In the same way that malware can detect if the operating system is running in a virtual environment, malware developers can design their malware to detect if the code is being debugged, disassembled, or examined by other tools (Brand et al., 2010).

Anti-memory

In case the malware analyst is clever or experienced enough to dump memory as a means of defeating anti-analysis measures, the malware developer can use anti-memory measures in order to frustrate the forensics analyst's effort. For example, the developer can have the packer that unpacks the code into memory to delete code as soon as it is executed (Brand et al., 2010).

Anti-process

Anti-Process techniques are designed to mitigate the attempts by forensic analysts debugging of running processes. The technique changes the entry point from to a different one which foils the debugging effort (Brand et al., 2010).

Packers and Protectors

Obfuscation and its subset, packing, are techniques used by malware developers to make static analysis more difficult for the forensics experts (Brand et al., 2010; Sikorski & Honig, 2012). Obfuscation is a means of hiding or disguising code (Sikorski & Honig, 2012). Packing uses compression and a wrapping program as a means of disguising the true purpose of program (Sikorski & Honig, 2012). Even more challenging for analysts and malware detection is recursive packaging which obfuscates code in multiple layers of recursive compression (Egele et al., 2012).

Metamorphic or Polymorphic

This type of malware is constructed in such a manner that it can re-engineer or recode itself (Raj et al., 2014; Sikorski & Honig, 2012). This recoding takes place each time it propagates or is distributed through a network. This type of malware hinders the use of signature-based protection tools (Raj et al., 2014).

Defining and Evaluating Open Source Intelligence

With related research understood, we can start to explore the independent elements that will go into the present research. Therefore, the next three sections will

address Open Source Intelligence (OSINT) which will be used to establish our ontology and as our data for text-mining and link analysis, which will be discussed in the following two sections.

Intelligence organizations and law enforcement at all levels of government use data and information found in open sources for decades (Fleisher, 2008; Steele, 2007). Traditionally, OSINT was characterized by searching through publicly available sources of information to include books, journals, magazines, etc. (Burwell, 1999; Fleisher, 2008). Steele formalizes the definition of OSINT as "information that has been deliberately discovered, discriminated, distilled and disseminated to a select audience (Steele, 2007, p. 132)."

Steele (2007) notes that the change to OSINT is the result of three distinct trends (1) the proliferation of the Internet; (2) the consequence of the related "information explosion" of published useful knowledge which is experiencing an exponential growth; and (3) the availability of formerly restricted sources of information resulting from failed states and newer non-state threats. Best (2011) acknowledges that the challenge with OSINT is not the collection of information but the filtering and distillation of the retrieved content into meaningful metadata that can be analyzed.

The Science of Natural Language Processing

In this section, we will explore the science of Natural Language Processing (NLP). It is broken down into three subsections. First, we will define the concept of Natural Language Processing (NLP) and terms used within this field of study that are applicable to the present research. Next, we will provide some examples of applications of NLP in the public domain. The third section provides a brief examination into how NLP analyzes human language. The last section explores some of the NLP tools available in the public domain with particular attention to the tools used within the present research.

Defining NLP and Key Terms

NLP is the science which enables computers to breaking down unstructured human language (Radev, 2015; Sims, 2015). It is a subset of machine learning which

draws upon the fields of Linguistics, Theoretical Computer Science, Mathematics, Statistics, Artificial Intelligence, Psychology and more (Radev, 2015). When we say unstructured human language, we refer to language as it is written or spoken.

While human language has its own structure it is not a structure that is conducive for machine searching or for relating ideas, concepts, or terms to one another (Radev, 2015). Documents that cover a particular area of study are referred to as the corpora (Börner, 2007; International Business Machines, n.d.).

The following list identifies commonly used NLP terms.

- **Information Extraction –** The process of extracting structured information from unstructured natural language texts (Waldron, n.d.).
- **Bag of Words –** A modeling techniques for training the NLP software to identify the sentiment of a given passage of text (Waldron, n.d.).
- **Coreference** – is the ability of software to identify relationships between specific and general terms within a text. Figure 1 is an illustration of determining a relationship between the multiple instances of his with President Xi Jinping mentioned at the beginning of the sentence.

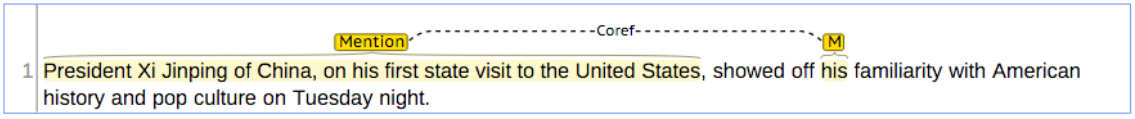

Figure 1: Visual Representation of Coreference (Stanford NLP Group, n.d.-a)

- **Named Entity Recognition (NER) –** the process of identifying or classifying terms within text by a broader classification (Finkel, Grenager, & Manning, 2005). Figure 2 illustrates how NLP software uses NER to classify words in analyzed text. Table 1: NER Classifications lists the 13 NER classifications used by Stanford CoreNLP. It is important to note that the Classification 'O' is a catch-all for any term that cannot be distinguished as one of the other NERs.

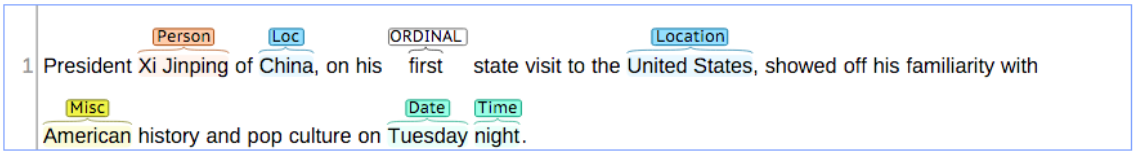

Figure 2: Visual Representation of NER (Stanford NLP Group, n.d.-a)

Table 1: NER Classifications

Named Entity Recognition Classification
DATE
DURATION
LOCATION
MISC
MONEY
NUMBER
O
ORDINAL
ORGANIZATION
PERCENT
PERSON
SET
TIME

- **Sentiment Analysis –** The ability to extract not only terms but also the positive or negative tone of the text (Waldron, n.d.).
- **Corpora or Corpus –** A large collection of texts within a specific domain (Waldron, n.d.)
- **Lemma** – the root word of tokens appearing in the corpora or in a single document (Jones, 1999). Table 2.1 below provides an illustration. The words capture, captured, captures, and capturing appear in the corpus. All of these words are derivatives of the lemma capture.

Table 2: Capture as an example lemma

Lemma	Word	Appearances
capture	capture	195
	captured	93
	captures	60
	capturing	35

Examples of NLP in the Public Domain

A variety of well-known and frequently used tools apply NLP to accomplish specific goals. The most common example is search engines (Radev, 2015). These websites, and their related robots, use NLP as a tool to index websites and facilitate user searches. Other examples include web-based translation tools (i.e. Google Translate), and Natural Language Assistants (e.g. Siri, Cortana, etc.) (Radev, 2015).

IBM's Watson

Another example of an NLP application is International Business Machines' (IBM) Watson Application Program Interface (API). The Watson system, like the other NLP tools described later, processes various forms of the written human word (e.g. Word Documents, Portable Document Format (PDF), HTML web pages, etc.) and pairs of questions and answers are stored within Watson's database which users can then query (International Business Machines, n.d.).

The Watson application has been used to analyze thousands of medical documents and build searchable databases of diseases and their symptoms. **Error! Reference source not found.**, below, provides an example of how Watson processes numerous medical resources to build a structured understanding of cancer and other diseases with similar symptoms, treatments, and side effects.

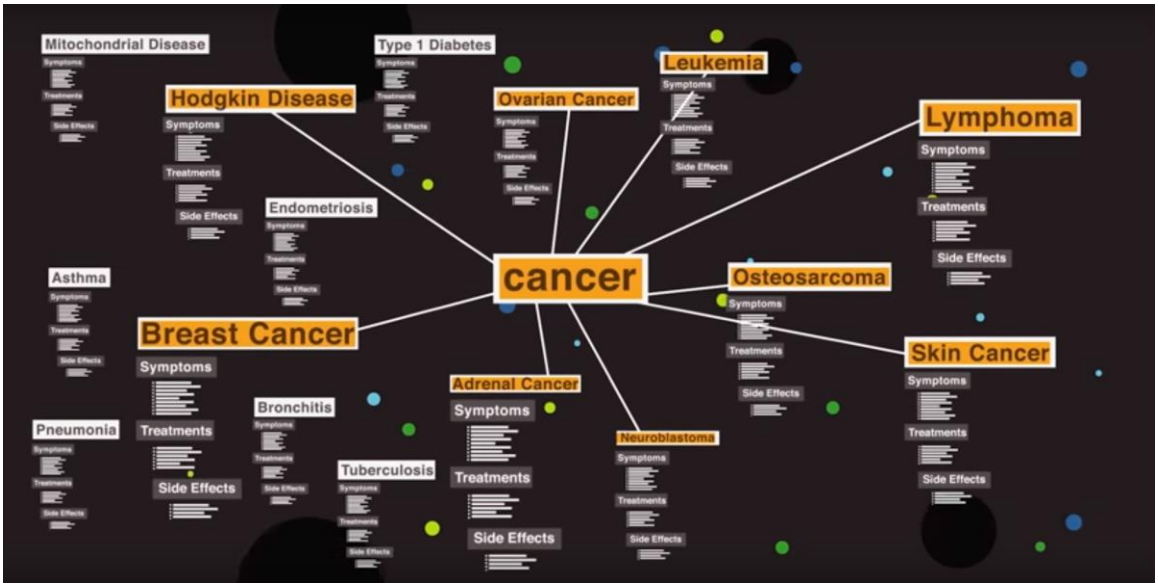

Figure 3: Watson structures Natural Language Resources within a domain (International Business Machines, n.d.).

Stanford's CoreNLP

Stanford's CoreNLP (CoreNLP) is actively developed by Stanford's Natural Language Processing Group (NLP Group). The NLP Group members include faculty, post doctorates, graduates, undergraduates, and programmers (Stanford NLP Group, n.d.-b). CoreNLP is a collection of tools, built in Java, which can extract word lemmas, parts

of speech, proper names of individuals and corporations, markup sentence structure and more (Manning et al., 2014).

BookNLP

BookNLP is an extension of the CoreNLP. It is designed to scale CoreNLP's functionality to larger source documents (Bamman, Underwood, & Smith, 2014). It also leverages MaltParser for dependency parsing (trained on CoreNLP's typed dependencies). The project was last updated over 3 years ago, based on its GitHub repository (https://github.com/dbamman/book-nlp),

Natural Language Toolkit

Natural Language Toolkit (NLTK) is a collection of tools written in Python designed to work with human language data (Bird, Loper, & Klien, 2009). An open source package currently in its third version, NLTK has over 100 predefined corpora and training models and can process 14 languages (Bird et al., 2009).

How NLP Processes Human Language

The two main functional roles of NLP are semantic and syntactic analysis (Collobert et al., 2011). Syntactic focuses on the way language is structured and deriving an understanding of the language through the understanding of the structure. Semantics focuses on the mapping of language to derive meaning from the way that words are used within natural language(Wang, Berant, & Liang, 2015). Collobert et. al. (2011) further divides these two areas into four main roles for NLP as follows:

- **Parts of Speech (PoS) Tagging** which is used to identify words by their syntactic role (e.g. nouns, verbs, adjectives, etc.) (Collobert et al., 2011; Wang et al., 2015);
- **Shallow Parsing or Chunking** which is used to identify syntactic phrases, commonly Noun Phrases and Verb Phrases (Collobert et al., 2011);
- **Named Entity Recognition (NER)**, which identifies parts of the sentence into categories like people and places (Collobert et al., 2011), and;

- **Semantic Role Labelling** analyzes the semantics of a sentence and establishes relationships between the semantic components (Collobert et al., 2011).

Beyond these parsing functions, NLP offers the following functionality when processing text:

- **Deep Analytics** involves advanced data processing techniques to facilitate more precisely targeted and more complex searches (Sims, 2015);
- **Machine Translation** which involves converting human text from one language to another (Sims, 2015);
- **Co-reference Resolution** is a tool that resolve the relationship between terms that refer to the same subject (Sims, 2015), and;
- **Automatic Summarization** functionality in NLP can be used to produced readable summaries of denser texts (Sims, 2015).

In order for NLP to provide usable data regarding a particular domain it is often useful to train the NLP system with a lexicon (Collobert et al., 2011; Hake & Vaishalideshmukh, n.d.; Lee et al., 2011; Wang et al., 2015). With the lexicon, the NLP application has a starting point from which to analyze documents within a domain. It is for this reason that the methodology in the first iteration includes a manual processing of documents by the research team. By default, CoreNLP is trained using the coreference library developed out of the SIGNLL Conference on Computational Natural Language Learning 2011 (CoNNL-11) (Lee et al., 2011).

The Role of Ontology

An ontology is designed to establish a common vocabulary and give practitioners the ability to easily share concepts (Mundie & McIntire, 2013). An ontology helps improve knowledge management within a domain (Dey, Rastogi, & Kumar, 2007). The development of an ontology is further intended to streamline the process of information sharing and to avoid problems of misrepresentation or miscommunication resulting from parties using incompatible terminology (Huang, Acampora, Loia, Lee, & Kao, 2011). Previous research in the area of malware analysis and APT detection will be discussed in the section on Prior Related Research.

Ontology Basics

Dey, Rastogi, and Kumar (2007) notes that the role of ontologies is to help improve knowledge management by formalizing the representation of domain-specific knowledge. Hitzler, et al. (2012) adds that these terms, often referred to as vocabulary, are used to precisely describe the domain. These terms can be identified by the frequency of their appearance in natural language texts (Dey et al., 2007; P Velardi, Fabriani, & Missikoff, 2001). The importance of terms and their relationship to other terms can be determined by the co-occurrence of terms within natural language texts (P Velardi et al., 2001). Karoui, Aufaure, and Bennacer (2007) take co-occurrence one step farther by weighting the relationship based on the proximity of terms to one another within the document.

Ontology Development

Ontologies are developed both vertically and horizontally (Navigli & Velardi, 2004; P Velardi et al., 2001; Paola Velardi, Faralli, & Navigli, 2013). From top-to-bottom, terms move from broader terms to more specific ones (Paola Velardi et al., 2013). For example, if the term at the top of the chain is communication, the next level could contain two terms secure and unsecure. Under Unencrypted Communications there would be more specific individuals including Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), and email. Individuals under Secure Communications could include Secure File Transfer Protocol (SFTP), Secure Shell (SSH), Secure Copy Protocol (SCP), Virtual Private Network (VPN), etc. In other ontologies the subordinate terms may also be a part of the parent term (e.g. the lobby of the hotel) (P Velardi et al., 2001). Individuals are specific instances of the parent term (i.e. Apache or Nginx are instances of a web server) (P Velardi et al., 2001). Figure 4: Example of the Vertical and Horizontal Aspects of an OntologyFigure 4, below, is a visual representation of the same.

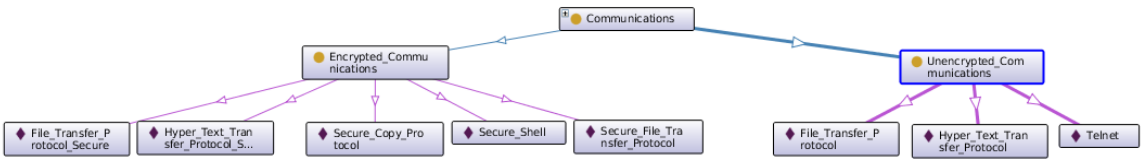

Figure 4: Example of the Vertical and Horizontal Aspects of an Ontology

In addition to classes and individuals, ontologies can specify properties of classes and properties of individuals (Horridge, Knublauch, Rector, Stevens, & Wroe, 2004). Properties define attributes of a class or an individual. In the example, above, a property might be the type of encryption used (e.g. MD5 or SHA2) to encrypt the communication or it could be the ports used to communicate between hosts.

Prior Related Research

This section will look at some of the research that focus on the use of ontologies for malware analysis and the development of fuzzy logic and cognitive agents for use in detecting APT attacks. As has been previously illustrated, malware is a significant tool in the execution of an APT attack. Therefore, research done in this area is impactful on the work in the present effort.

Mundie and McIntire

In Mundie and McIntire's (Mundie & McIntire, 2013) research they sought to develop an ontology for Malware Analysis. Their work was motivated by four challenges in the business of malware analysis: (1) security teams and their customers were wasting time negotiating requirements because they did not "speak the same language;" (2) human resources departments couldn't hire the right malware analysts because they could not properly explain job requirements; (3) certification programs did not have a standardized way to assess the abilities of malware analysts; and (4) information sharing within the malware analysis community is impeded by a lack of shared foundation (Mundie & McIntire, 2013).

Their work employed six increasingly complex levels of knowledge representation (Mundie & McIntire, 2013):

- **Controlled Vocabulary –** collection of preferred terms.
- **Taxonomy -** hierarchically related terms in a controlled vocabulary.
- **Static Ontology –** an ontology that describes static aspects of the world.
- **Dynamic Ontology –** an ontology hat describes changing aspects of the world.

- **Intentional Ontology –** a subjective ontology based on the motivation of agents.
- **Meta-model –** An ontology template that can generate ontologies by filling-in the parameters.

Their work yielded a vocabulary of approximately 270 malware analysis terms and a taxonomy outlined in World Wide Web Consortium's Web (W3C) Ontology Language (OWL). Mundie and McIntire built their initial ontology using the email archive of a malware analysis team. They also included various recognized textbooks in the malware analysis field and some Internet resources (Mundie & McIntire, 2013).

Huang, Loia, and Kao

Huang, Loia, Lee, and Kao (Huang et al., 2011) research sought to apply fuzzy logic and ontologies for their application of inferring knowledge about malware, and designing an intelligent decision making system whose behavioral rules can be used to detect viruses and other malicious programs. As with Mundie and McIntire's research they employed OWL to build their malware behavior ontology (Huang et al., 2011).

To test the effectiveness of their decision-making system they evaluated its performance against 30 "attendance records" from the National Center for High Performance and Computing (NCHC) malware repository. Their reported results and conclusion indicate that the employment of fuzzy logic and ontology was feasible and usable for a malware behavioral analysis system (Huang et al., 2011).

Meckl, Tecuci, Boicu, and Marcu

Meckl, Tecuci, Boicu, and Marcu (2015) attempted to improve cyber defense against APTs using an operational semantic approach. The motivation for their work was to reduce false positives thus increasing the efficiency and reducing the costs associated with automated APT analysis. To achieve this, they are proposing developing collaborative cognitive agents that can apply updates based on new intelligence. In theory, the work of this present study could potentially inform this agent (Meckl et al., 2015).

Lundquist, Zhang, and Ouksel

In Lundquist, Zhang, and Ouksel (Lundquist, Zhang, & Ouksel, 2015), the research focused on applying an ontology to the analysis of network traffic in order to determine the nature of the traffic as a threat or as innocuous. The extent of the threat (or non-threat) is also determined as part of this scan. If the sentiment scan is inconclusive the traffic data goes through further processing that involves alternate ontologies and/or expanded data from a longer observation window (Lundquist et al., 2015).

These and potentially other research efforts have some overlap with our current effort, but the researcher contends that this does not invalidate our efforts. Instead, the researcher holds that the research could augment these other research efforts and others like them.

Summary

In this chapter, we defined of the term Advanced Persistent Threat both in terms of the attack and the actor. With that groundwork laid, we explored the phases of the APT attack in the context of the widely-accepted Lockheed Martin "Cyber Kill Chain." We looked at several illustrative historic examples of known APT attacks and the current challenges cybersecurity professionals face when trying to detect APT attacks. We also examined the costs associated with data breeches resulting from APT attacks.

We examined the concepts of OSINT, ontology development and natural language processing as an intelligence building tool. Finally, we examined prior research efforts in the fields of APT and malware ontologies and the application of those ontologies.

FRAMEWORK AND METHODOLOGY

With definitions, history, and a review of prior APT research complete, it is time to layout the approach that the author intends to apply in the present endeavor. This chapter will examine the framework, researcher's biases, the proposed methodology including the two phases the research will take, a review of the data to be collected, the tools that will be used to analyze the raw data, and finally the researcher's credibility.

Framework

As demonstrated in the literature review, previous research efforts sought to develop a meaningful ontology as a part of malware analysis and incorporating a semantic approach to improve defense against APT attacks. However, none of these prior efforts sought to apply link analysis to the existing APT knowledgebase in an effort to refine or improve the threat intelligence that exists for each individual APT.

The author acknowledges that some corporate entities may have explored employing link analysis to expand their threat intelligence about APTs but, that work being of a proprietary nature, it has not been published in the public domain as open research.

Researcher Bias

While there is always a risk of bias when dealing with qualitative or hybrid research efforts, steps within the methodology outlined below are designed to minimize the possibility that it will affect the outcome of this research effort. My only vested interest in this research is to provide the most thorough effort that I can within the time parameters available to me.

Methodology

This section lays out the methodology used in the research related to this paper. It describes the workflow illustrated in , below.

Figure 5: Methodology Workflow

The first step in the research process involved the collection of OSINT pertaining to APT incidents over the past decade. During the literature review research, the search for resources uncovered two repositories of documents pertaining to APT attacks. These two repositories contain over 700 documents include reports by cybersecurity professionals and organizations covering individual incidents and studies of trends about cyberattacks over time. Organized by year, these primary sources span from 2008 through 2016. The size of these repositories provides a significant foundation for building this study's ontology.

As a part of the merging of the two repositories, this step included the removal of duplicates documents. Other documents were removed for other reasons as explained in Chapter 4.

As discussed in the 'Science of NLP' section of the Literature Review, in order for NLP software to produce the most useful analysis regarding a domain, the software needs to be trained regarding the lexicon of the domain. Therefore, the researcher selected 5% of the documents randomly from the final knowledgebase and read through the selected documents and identified terms and phrases that are meaningful within the APT domain.

The next two steps on the methodology were performed in parallel as neither required the other as a preceding task. The two steps were the NLP processing of the remaining documents in the corpus and the building of an initial ontology using the terms identified during the manual analysis.

The researcher included this step for multiple reasons. First, to understand the breadth and scope of an APT domain ontology. Second, as a familiarization with the ontology building software used in the research.

After completing the two parallel tasks, the next step was the extraction of domain relevant terms from the processed documents. Using knowledge gleaned from the manual processing of documents, I determined to use a statistical approach to initially identify ontological terms and then to use contextual analysis to identify additional terms. From there I used a tool I am familiar with to build the relational structure of the ontology and then connected Protégé to the database in order to develop and the final ontology (presented in OWL2 in Appendix C).

Data Collection

The data collected in this research comes in two forms. The first is the statistical analysis of the tokens outputted by NLP. The second is the identified terms in the APT ontology and the lexicon used in the NLP step of the methodology.

As mentioned in the previous section, statistical analysis of the NLP output (in the form of tokens) enabled the identification of terms for use in the ontology and would facilitate the identification of additional terms through their contextual use.

In order to facilitate the statistical analysis of the NLP output, I first turned to Excel but the number of tokens in the entire corpus made statistical analysis in excel difficult. Excel spreadsheets are limited to 1,048,576 rows or records and, as mentioned previously, the corpus of 425 documents consists of 2,423,738 tokens. Therefore, it would take three separate tabs of a spreadsheet just to contain the raw data. This would make analysis more complex than it needed to be.

As a result, I turned to MySQL for its ability to handle the 2.4+ million tokens (records) in the corpus and use its querying capability to summarize the data the output to build the domain ontology. I used MySQL to summarize the data in the corpus and then

analyzed those results using Excel and SAS. By summarizing, I am referring to counting the number of appearances of original words, lemmas, and NER. I then used that.

Before selecting terms for the ontology, it was important to gain both a contextual understanding of terms as they are used within the texts of the corpus as well as statistical analysis of word frequency within it. This process is explained in more detail in Chapter 4.

Analysis

Analysis of the available open source intelligence employed a recognized industry standard and the processing of documents were performed with multiple software packages. The next section examines the standard and software for developing the ontology. It will be followed by a review of the software package chosen by the researcher to perform the link analysis of the available OSINT.

As discussed previously, an accepted industry standard for ontology development is OWL, which is currently in its second edition (Huang et al., 2011; W3C OWL Working Group, 2012). OWL2 provides the structure to show the vertical relationship between entities and their properties (e.g. communication to secure communications) as well as the horizontal relationships between like properties (e.g. SSH to SFTP) (Hitzler et al., 2012).

While OWL2 provides the structure, Protégé provides the user interface to work with ontologies. Developed as a tool to develop Semantic Web Applications, Protégé provides access to ontologies, as well as tools to visualize, edit and use ontologies (Horridge et al., 2004; Knublauch, Fergerson, Noy, & Musen, 2004). Protégé is designed to simplify the use of the OWL2's notoriously difficult structure (Knublauch et al., 2004). It's incorporation in several other ontology development research projects weighed heavily in the researcher's decision to use it for the current research effort (Abulaish & Dey, 2007; Huang et al., 2011; Mikroyannidis & Theodoulidis, 2007; Mundie & McIntire, 2013). This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health (Musen & The Protégé Team, 2015).

Credibility

The credibility of any study is paramount to the researcher and to the product developed. The credibility for the present research project will be built on the foundation of industry standards for the work being performed as well as widely used and recognized tools to perform the work. The use of such standards and tools ensure that the research is repeatable by any researcher with access to the tools and the raw data used in this endeavor.

The researcher has 23 years of experience in the arena of networking and cybersecurity. Ten years of the 23 have been with the United States Army. As an Officer his focus has been on the installation, operation, maintenance, and security of both the tactical and operational networks within the Army up to and include theater level operational network spanning seven Army installations in the United States Army Pacific footprint which stretches from Japan up to Alaska and down to Hawaii.

Summary

This chapter examined the framework and methodology the researcher intends to employ in the current research. It also addressed the technical tools the research will use that are widely accepted within the fields of natural language processing, and ontology development.

RESEARCH AND ANALYSIS

This chapter of the paper is an examination of the methodology employed during the research and the analysis of the same. This chapter is broken down into ### sections. The first section focuses on building the corpus. The second outlines the manual process for building the training lexicon. The third examines the processing of the documents through the NLP software and addresses issues with migrating that data into MySQL for statistical analysis. The fourth section presents an analysis of the NLP output as well as the steps taken to filter through the output in order to find ontological terms more efficiently. The fifth section addresses the selection process used to identify ontology appropriate terms and the building of the ontology structure itself.

Building the Corpus

As previously discussed, the first step of the research relied on two APT documentation repositories found on GitHub. After downloading both repositories they contained a combined 742 files. Several files had issues (e.g. unprintable characters in the filename, filenames greater than 256 characters, etc.) that were correctable without impacting the integrity of the file's content. However, many files were excluded from the process for various reasons. These reasons are as follows:

- Two hundred and ninety-two (292) files were duplicates as determined by the tool *fslint*.
- Ten (10) files were excluded because they were written in foreign languages. While there are online tools that can translate documents; given the limited number of documents, my lack of knowledge of these languages, and time it was more efficient to eliminate them.
- Eight (8) files were eliminated because they were encrypted, or had errors that prevented text extraction by the text extraction tool.
- Six (6) files generated errors for invalid PDF formats while trying to extract text. A manual review of these files revealed that the extracted text was unusable.

- Despite all software support recommendations, the parsing of one (1) continually failed.

With 314 files either duplicates or containing issues that made them unusable for this study, the ontology development relied upon 425 documents pertaining to various APT attacks over the last eight years.

During this process, each document received a document number, primarily as a means to track the processing of documents. Secondarily to facilitate the random selection of the documents that I used to establish the initial lexicon.

Building the Initial Training Lexicon

As described in Chapter 2, NLP software needs to be "training" when it starts working with documents in a new domain in order to better identify parts of speech, named entities, etc. In order to establish the training lexicon, I selected 22 documents randomly using the document numbers assigned during the "Establishing the Corpus" phase, described above, and a random number generator.

The 22 documents selected, represented just over 5% of the total corpus (5.02%) and the random number generator ensured that the documents covered a diverse number of incidents. While reading through the original documents provided the initial lexicon, formatting requirements for the lexicon necessitated that I also process the documents through the NLP software in order to simplify the building of the training file. However, there were issues with the NLP software's output that could not be avoided and which will be discussed in the next subsection.

Processing Documents through BookNLP

After extracting the plain text out of the documents in the corpus the natural language processing of the documents began. As discussed in Chapter 3, I employed Stanford CoreNLP through the extension BookNLP to process the text in all the documents.

Numerous documents still contained what the software flagged as 'Untokenizable Characters.' These were presented as unprintable characters in the ASCII character set.

These messages appeared to be warnings as the software successfully completed its execution.

Errors with NLP Output

During my early experimentation with CoreNLP and BookNLP I ran into errors processing some of the PDF documents including the types of issues mentioned above. In order to avoid eliminating more documents for technical issues, I extracted the text of these files using the Python script *pdf2txt.py*. Even after performing this step there were some issues which are outlined below along with steps taken to correct these errors.

Unhandled Exception

While MySQL imported the tab delimited output from BookNLP, some of the documents generated the error in Figure 6 and the process simply failed.

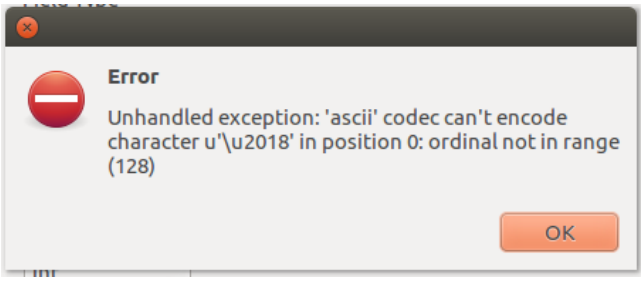

Figure 6: “Unhandled Exception” Error

Examining the file revealed a non-ASCII character in the “Original Word.” The character in question is shown in line 442 in Figure 7. Given that this token is not one that would be used in the ontology the easiest solution was to replace the character with a number.

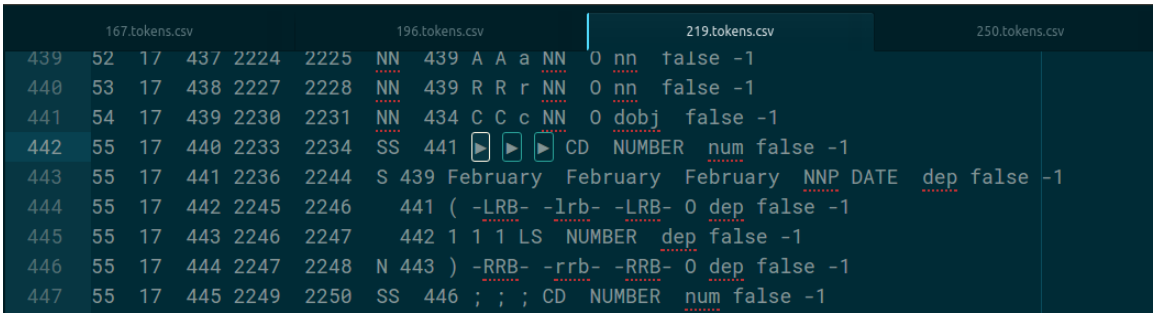

Figure 7: Sample character that createsd import issues

Import File Error – Non-ASCII Character

When there were non-ASCII characters (as shown in Figure 7) in the first row of the of the tab delimited file the error messages in Figure 8 were displayed and no records were imported. Clicking through this error resulted in the error in Figure 9. As with the "Unhandled Exception" error above, replacing these with an ASCII punctuation character.

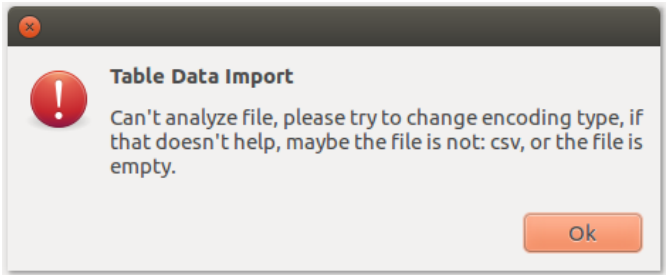

Figure 8: "Import Error" Issue

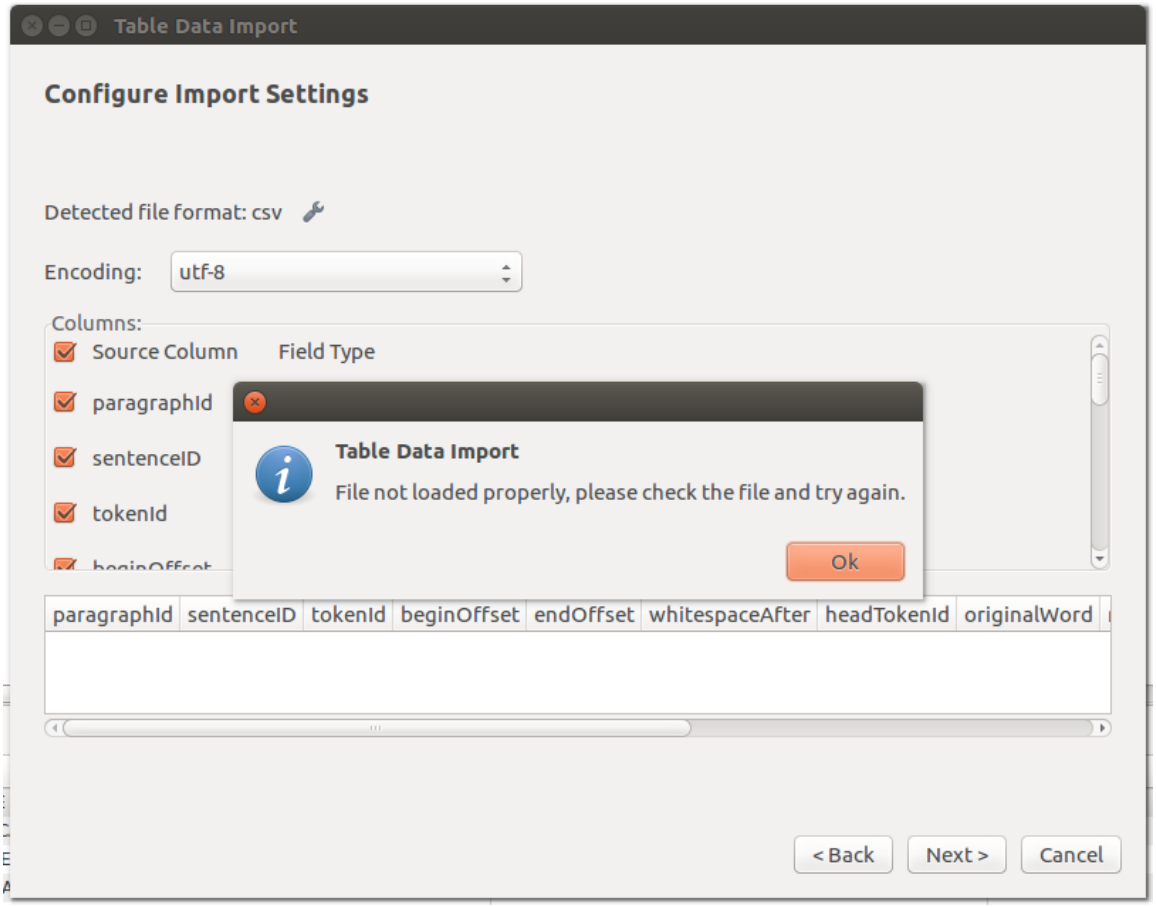

Figure 9: "File Not Loaded Properly" Message

Import File Error – Quotation Marks

This error was more challenging to uncover because no errors were generated. The file would process correctly but later I would discover a discrepancy in the number of records in the document table and the largest token ID as generated by BookNLP. Through trial and error, I realized that the errors were being caused by different quotation marks within the original text as shown in Figure 10. Escaping the quotations with a backslash (\) corrected this problem.

Figure 10: Quotation Marks Example

Escaping Characters

As with the problem described in the previous subsection regarding quotation marks. There were many other characters that generated errors during the MySQL import because they are special characters that MySQL treats differently including the backslash (\). In order to be processed properly these special characters need to be escaped with a backslash just as in the example above. With the NLP processing complete and all the data imported into the MySQL database, it is time to proceed with the statistical analysis of the NLP output itself.

Examining the Documents in the Corpus

A person reading a book processes what he reads very differently from an NLP program that breaks the document down into tokens. While this can be understood conceptually, when the processing was complete and all the data was imported into the database it was daunting to see that there were more than 2.4 million tokens in the document. It would be impossible to process through all those records and create the

ontology in the time afforded to me. I determined that in order to build the ontology it was necessary to reduce the number of records in order to get to the most meaningful and useful terms.

The following five subsections provide statistical analysis of the database and walk the reader through the steps taken to remove extraneous words and get to those words that are useful for ontology development in this domain. The first section looks at the concept of the token. The second examines all the tokens produced by BookNLP. The third explains the removal of common word tokens. The fourth explores the use of lemma. The last examines the NER as another means of interpreting the data.

Understanding the Tokens

Reading through the Processing the 22 documents produced 128,120 total tokens with only a small percentage of these tokens being significant for building the APT Ontology. there was a significant amount of clutter or noise within these documents that had to be sifted through to find the needed keywords. Figure 11, below, is a sample of the tokenized output generated by BookNLP for one of the selected documents. Notice that line 10 is the possessive apostrophe s (‘s) that is actually part of the word in the previous token (Storm). Figure 11 contains a partial list of this and similar tokens in those 22 documents.

```
1  paragraphId sentenceID  tokenId beginOffset endOffset whitespaceAfter headTokenId originalWord  normalizedWord  lemma pos ner deprel  inQuotation
2  0 0 0 0 10  NN  3 2015/10/31  2015/10/31  2015/10/31  CD  DATE  num false -1
3  1 0 1 12  14  S 3 An  An  a DT  O det false -1
4  1 0 2 15  23  S 3 In-Depth  In-Depth  in-depth  JJ  O amod  false -1
5  1 0 3 24  28  S 20  Look  Look  look  NN  O nsubj false -1
6  1 0 4 29  31  S 3 at  at  at  IN  O prep  false -1
7  1 0 5 32  35  S 12  How How how WRB O advmod  false -1
8  1 0 6 36  40  S 7 Pawn  Pawn  Pawn  NNP O nn  false -1
9  1 0 7 41  46    10  Storm Storm storm NN  O poss  false -1
10 1 0 8 46  48  S 7 ’s  ’s  ’s  POS O possessive  false -1
11 1 0 9 49  53  S 10  Java  Java  Java  NNP MISC  nn  false -1
12 1 0 10  54  62  S 12  Zero-Day  Zero-Day  Zero-Day  NNP O nsubjpass false -1
```

Figure 11: Sample of Document Tokens

Another issue with how the NLP software parsed the documents can be seen in lines 8 and 9 of Figure 11. BookNLP broke up a two-word phrase which is actually the name of the APT discussed in the document. Therefore, in order to build an accurate lexicon required human intervention to modify the normal function of the software.

The issues described in this subsection are not an indictment of the software, it is functioning as the designers intended. In fact, my inexperience with the software probably contributed to some of the challenges described.

Examining the Corpus Tokens

With an understanding of what tokens are established, we can move forward and look at the corpus. The corpus consists of various types of documents including blog posts, corporate white papers, and academic explorations of various APTs. The 425 documents contain 2,423,738 tokens in total.

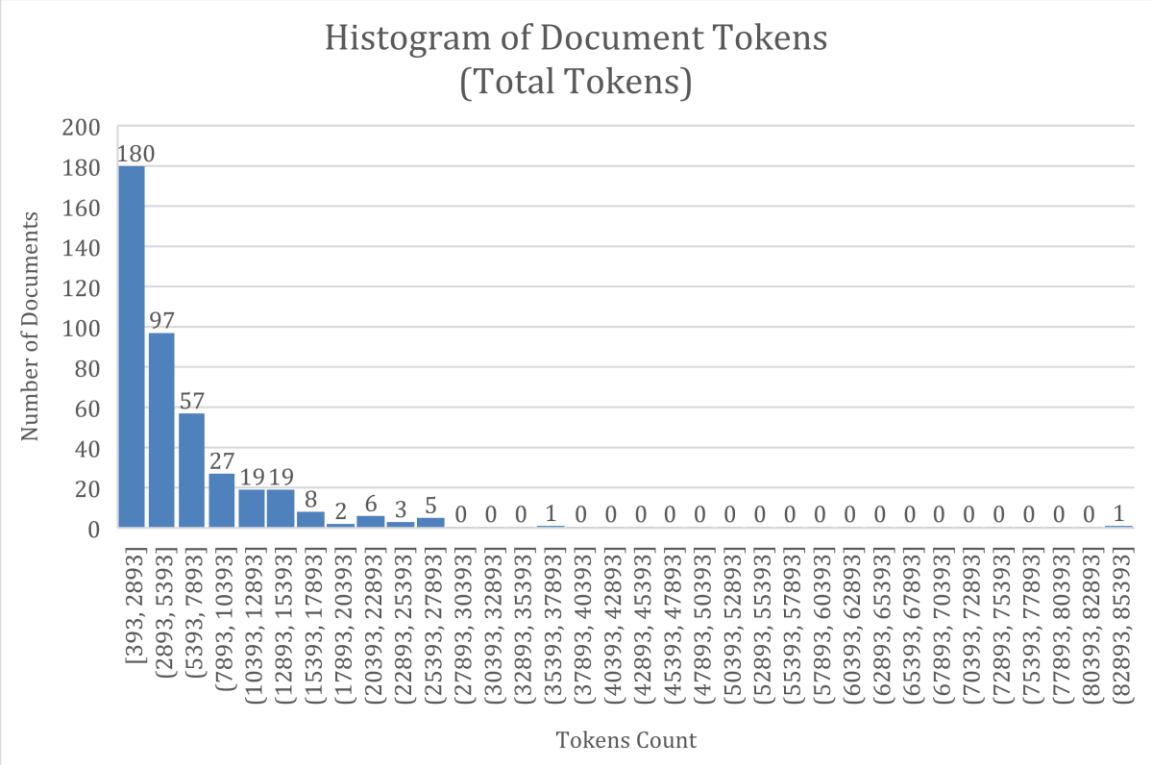

Figure 12: Distribution of Tokens (by Document)

The mean tokens in a given document is 5702.91294. The largest document contains 84,473 tokens while the smallest consists of 393 tokens. Table 3, below, highlights the quantile data for all tokens in the Corpus Documents. The histogram (Figure 12), illustrates the breakdown of the documents by the number of tokens extracted during the NLP processing of the corpus. It shows 277 documents (65.18% of the total) contain less than 5393 tokens. Seventy-five percent (75%) of the documents contain less than 7,144 tokens. These two statistics coincide with the fact that most documents were website reports and summary white papers. The two outlying documents explain the discrepancy between the mean

Table 3: Quantiles for All Tokens in the Corpus

Level	Quantile
100% Max	84,473
99%	26,705
95%	16,562
90%	13,116
75% Q3	7,144
50% Median	3,611
25% Q1	1,924
10%	1,203
5%	986
1%	509
0% Min	393

These numbers reflect all tokens within the documents. They include punctuations, symbols, and numbers as well as actual words. While these tokens provide important information and context they are not likely to contain ontologically meaningful terms. Removing these tokens from the documents we are left with the words that may be included in the ontology.

Table 4: Quantiles for All Words in the Corpus

Level	Quantile
100% Max	44,824
99%	22,008
95%	12,938
90%	9,824
75% Q3	5,521
50% Median	2,838
25% Q1	1,561
10%	990
5%	818
1%	408
0% Min	198

The mean number of words in the corpus documents is 4441.64706 while the median is 2,838 as shown in Table 4. Ninety-five percent (95%) of the documents are between 0 and 13,827 words in length. With most documents being less than 10 pages in

size, these sizes make perfect sense. However, this is still a large number of terms that need to filtered down in order for me to more readily identify ontological terms.

Removing Common Words from the Data

Examining the corpus tokens, it quickly became obvious that there were frequently appearing words that would not provide anything useful for ontology development. With some additional investigation and research, I realized that the words appearing in the corpus with the greatest frequency were contained within the list of the most common words in the Corpus of Contemporary American English (CoCA) (Davies, 2016). Table 5, below, shows the 10 most common words in the CoCA with their appearance rate in all 425 corpus documents.

Table 5: 10 Most Common words in Contemporary American English

Common Word	Appearances	Appearances per Corpus Document	% of Corpus Tokens
the	107,486	252.9082	5.6940
be	50,884	119.7271	2.6956
and	38,463	90.5012	2.0376
of	42,756	100.6024	2.2650
a	41,026	96.5318	2.1733
in	28,241	66.4494	1.4961
to	41,211	96.9671	2.1831
have	10,678	25.1247	0.5657
it	9,145	21.5176	0.4845
I	1,583	3.7247	0.0839

Filtering out the 5,000 most common words in the contemporary American English language removed 1,248,000 of the words in the corpus or 66.1122% of the word tokens within the corpus. This process left the corpus with 639,700 original words.

Using Lemma

In an effort to further reduce clutter within the summarization, I sought more potential ways to combine rows in the summary table. The lemma provided a tailored solution for this problem. Table 6, below is a sample of the summary table for all lemma

beginning with the word ‘alert.’ The search produced 12 rows of data based with 9 unique lemmas.

Table 6: Sample Lemma

Lemma	Variation	Part of Speech	Document Count by Lemma	Lemma Appears	Document Count by Variation	Variation Appears
alert	alert	JJ	107	265	68	172
	alerted	VBD			19	25
	alerting	VBG			12	12
	alerts	VBZ			37	56
alert-14-281-01p	ALERT-14-281-01P	NN	1	1	1	1
alert-avoiding	Alert-avoiding	JJ	1	1	1	1
Alert/Scareware	Alert/Scareware	NNP	1	1	1	2
alert5	alert5	NN	1	1	1	1
alerted16	alerted16	CD	1	1	1	1
Alerter	Alerter	NNP	1	1	1	3
Alerts	Alerts	NNP	5	5	5	5
alertwall.exe	alertwall.exe	NN	1	1	1	1

From this table, we can see that the lemma alert has 4 variations for the text actually found in the documents (alert, alerted, altering, and alerts). The lemma alert appears in 107 total documents and appears a total of 265 times.

There are two important items to note at this juncture. One is a point about what might appear as a mathematical discrepancy and the other is a result of the NLP software.

Regardless of the issues that will be explained momentarily the final compilation of tokens resulted in 98,945 lemmas with 97,886 variations. Only 2,772 Lemma or 2.80156% had more than one variation.

Discrepancy between Variation and Lemma Counts

Looking at Table 6, once again, one will note that there is a difference between the number of documents by Lemma and by word variation for the words with the lemma alert (107 vs. 136). After some evaluation, the reason for the difference is the overlapping appearances of variations within the same document. For example, the variations alerts

and altered may appear in the same document and will be counted once for each variation but only once for the lemma.

Multiple Lemma for One Variation

The second issue was far subtler to detect. While the first lemma in Table 6 includes the variation *alerts* there is a second lemma (Alerts in the second to last row) that has the same variation, *Alerts*. Taking into account just the variation both instances of the word alerts should fall under the lemma alert and appear in 38 documents a total of 61 times. However, upon further investigation, the difference comes down to the part of speech. The first instance is a verb while the second is part of a noun phrase. Therefore, CoreNLP and, by extension, BookNLP recognizes the lemma for the noun as *alerts* instead of *alert.*

In total, there were only 1,057 distinct variations identified by multiple lemma. Only 1 variation had more than 2 lemmas associated with it. Table 7, below, provides a sample of the typical situations where this occurred.

Table 7: Sample of One Variation with Two or More Lemmas

Variation	Related Lemmas	Various Lemmas	Total Appearances
SYS	3	sy, sy., sys	102
02i	2	02i, 02us	14
0xed	2	0xe, 0xed	3
10ms	2	10m, 10ms	1
2000s	2	2000, 2000s	10
3DES	2	3de, 3des	44
4-bytes	2	4-byte, 4-bytes	11

Examining the NER

As discussed in Chapter 2, NLP employs a classification system when identifying Named Entity Recognition (NER) for each token. The tool identifies each token by one of 13 classifications. The first column in Table 8 shows the classification type. The second and fourth columns are the counts for each classification within the corpus before and

after all punctuation, numbers only, and common word tokens were removed, respectively. The third and fifth columns represent the percentages of corpus tokens, again before and after the removal of less useful tokens. As a reminder, the Classification 'O' is a catch-all for all tokens that the Classifier cannot identify as one of the other 12 classifications. Figure 13, Provides a stacked graph illustrating the same calculations.

Table 8: NER Classifications

NER Classification	Base Tokens		Adjusted Tokens	
	Count	Percent	Count	Percent
O	2,117,681	87.37%	1,698,627	89.98%
Number	136,643	5.64%	57,833	3.06%
Date	41,099	1.70%	19,228	1.02%
Organization	40,617	1.68%	39,861	2.11%
Misc	20,985	0.87%	20,478	1.08%
Location	16,864	0.70%	16,859	0.89%
Percent	13,998	0.58%	4,754	0.25%
Person	11,236	0.46%	11,229	0.59%
Time	8,598	0.35%	7,403	0.39%
Duration	6,194	0.26%	5,603	0.30%
Money	6,012	0.25%	2,069	0.11%
Ordinal	3,267	0.13%	3,267	0.17%
Set	544	0.02%	489	0.03%
	2,423,738	100.00%	1,887,700	100.00%

That there are still tokens which are classified by numbers, dates, time, and other numerically related classifications is the result of CoreNLP's identifying words that represent numbers (e.g. one, ten, fifth, etc.). The NER classifications for locations and organizations, and people proved useful in the current research as it helped to extract target locations, organizations and individuals for some of the APT attacks. However, the NER classification system was not perfect for the current research. During the processing of documents there were two issues which reduced this method's usefulness for the current research.

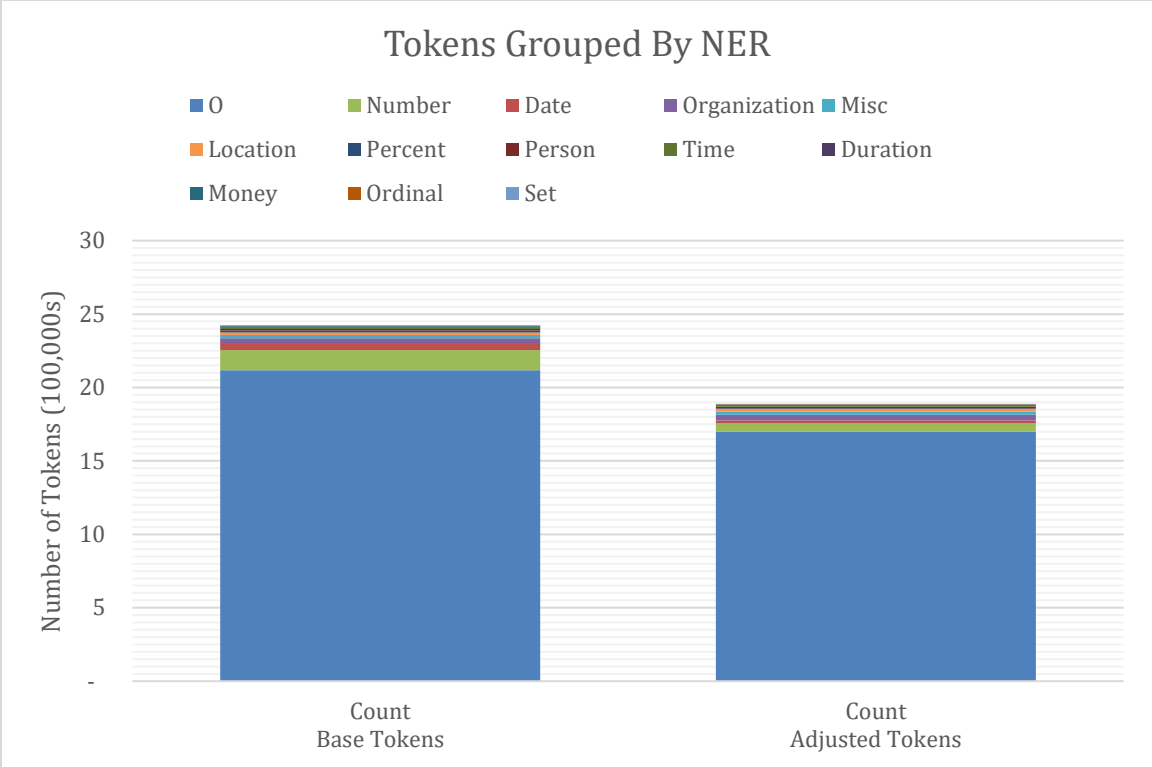

Figure 13: NER Classification Breakdown

NER Misidentification

The first of these issues was misidentification. The software is only as good as the training that the classifier has prior to analysis of a document. Without additional training for the software many tokens were misidentified. For example, Bitdefender, Bitcoin, FT.com, and other tokens were misidentified as locations. Another instance of misidentification can be seen in blog URLs. In Figure 14, below, shows five URLs that contain these dates.

```
1  http://www.reuters.com/article/2015/06/23/us-hackers-insidertrading-idUSKBN0P31M720150623
2  http://blog.shadowserver.org/2012/05/15/cyberespionagestrategicwebcompromisestrustedwebsitesservingdangerousresults/
3  http://www.welivesecurity.com/2015/07/10/sednitaptgroupmeetshackingteam/
4  https://blog.bit9.com/2013/02/08/bit9-and-our-customers-security/
5  http://blogs.wsj.com/chinarealtime/2011/10/05/patriotic-chinese-hacking
```

Figure 14: Sample of URLs Misidentified as Dates

Multi-Token NER Issue

The second issue is not an error as much as it is just the result of the way CoreNLP and BookNLP breakdown the processed documents. As the program breaks the

document down into individual words, any location, organization or individual that spans more than one word is not tracked as such. For example, United Nations or United States each exist as two tokens in the corpus instead of one.

With the data extracted from the documents and summarized within MySQL, we can move forward into the process of laying out the structures of the ontology and identifying terms to fill it.

Building the Ontology

With the documents processed, the data imported and summarized for statistical analysis and identification of terms. The first subsection will focus on the process gone through to select terms. The second will address the ontological structure as designed.

Selecting Terms for the Ontology

In the previous section, you were presented with some of the statistical analysis done to evaluate the nature of the corpus collected. However, apply statistical analysis for the purpose of identifying potential ontological terms. To this end, examining the frequency of terms as a means of doing just that. To this end, I considered five different permutations. They were:

- Appearances
- Appearances per Document
- Appearances per Corpus Documents
- Percentage of Corpus Documents
- Percentage of Corpus Tokens

The reasons for accepting and rejecting these different values are explained in the following two subsections. The third subsection, discusses how the accepted calculations were applied to the corpus.

Statistical Analyses that were Applied

In order to identify potential terms, it is useful to understand how the frequency of appearances relates to the larger whole. There are several ratios that were selected to

identify potentially useful term. They are *Appearances per Document*, *Appearances per Corpus Documents*, and *Percentage of Corpus Documents*.

Appearances per Document is the ratio of total appearances divided by the number of documents the word appears in. This number assumes that the number of appearances within the documents that it appears; indicating it is an important term. The range of values for this statistic ranges from 1 to 716.25.

Percentage of Corpus Documents uses the simple percentage of all corpus documents that the given term appears in. The mean percentage is 0.541% or approximately 2.3 documents while the minimum is 0.2353% (1 document) and maximum is 95.059% (404 documents). The mean percentage represents approximately

Appearances per Corpus Documents divides the number of appearances based on the total of 425 documents in the corpus. This is an effort to normalize the appearances across all documents. Whereas the minimum and maximum of the Appearances per Document ranged from 1 to 716.25, the range for Appearances per Corpus Documents is from 0.0024 to 47.1882. The mean is 0.014 and the mode is 0.0024.

Identifying words based on these three measurements alone would have included terms that would not provide anything meaningful to the ontology. Table 9 shows the top 15 tokens by each of the statistics. The reader will see clear overlap in the three lists as expected but there are differences in these lists.

Table 9: Top 15 Tokens (by Various Factors)

Token	Appearances per Document	Token	Appearances per Corpus	Token	% of Corpus Documents
cid	716.250	cid	47.188	malware	95.059
soft@hotmail.com	206.000	malware	19.969	data	88.235
xfish	197.000	’s	14.421	server	80.235
Ponmocup	170.000	data	11.645	’s	79.529
JinDiQIAO@ hotmail.com	162.000	server	9.713	malicious	77.882
ShimRat	162.000	C	7.875	Windows	75.529
Packrat	146.000	Windows	6.696	Microsoft	71.059
Exposureindicating	143.000	S	5.998	attackers	69.412
TERRACOTTA	137.000	attackers	5.974	IP	68.706
Xing	127.000	com	5.911	further	64.941
gif.dec	126.000	IP	5.741	C	64.000
Mofang	111.000	C2	5.595	email	60.235
GreenSky27	110.000	malicious	5.468	servers	59.529
2_digits	107.000	C&C	5.426	executable	59.059
Invincea	91.000	Microsoft	4.520	detection	58.118

Statistical Analyses that were Rejected

The *Percentage of all Corpus* Tokens is a ratio of the number of appearances divided by the total number of corpus tokens (1,887,700). However, it quickly became apparent that this ratio would not be a useful statistic for screening. Only one word represents more than 0.44959474 % of the corpus and a single appearance in the corpus is 0.00005297% of the whole and only 9,299 words that individually make up 0.00026487% or more of the corpus. The minor variation in the values makes distinguishing the value of this statistic difficult to use.

Figure 15, below, is a graphing of individual token counts as a percentage of the corpus that count represents. While there is measurable change from the smallest (1 appearance) to the largest (20,055 appearances), distinguishing the percentage difference between numbers closer together is subtler. Whereas, Figure 16, illustrates the percentages for all corpus tokens.

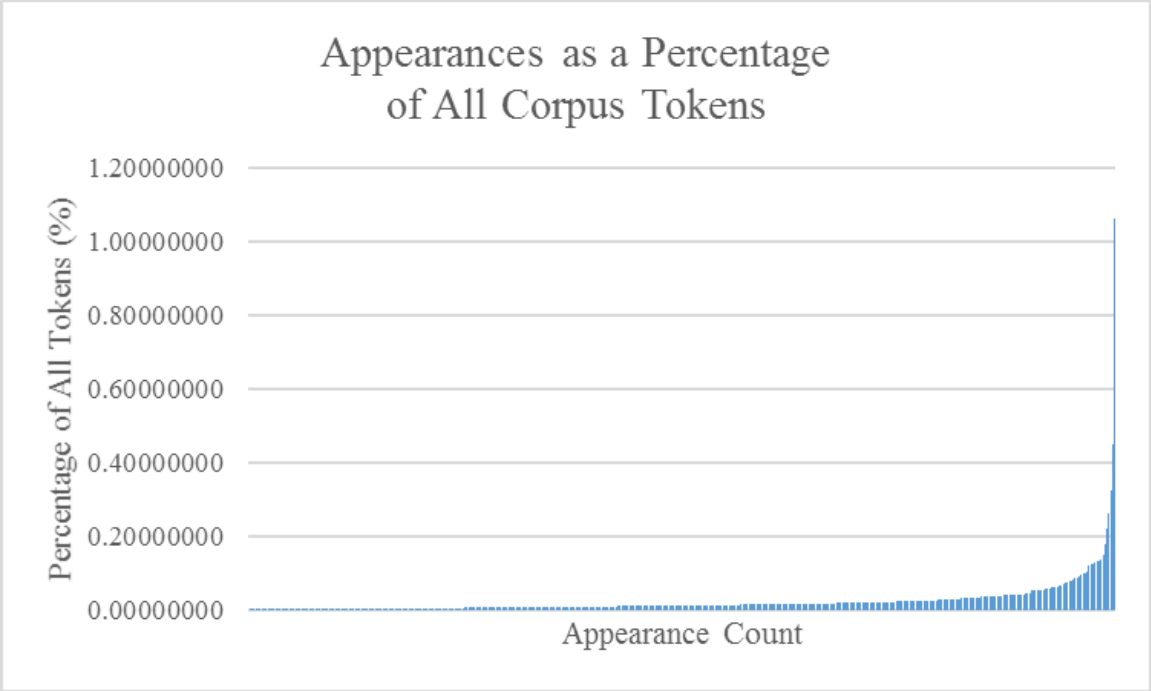

Figure 15: Representation off appearance Count as a Percentage of the Entire Corpus

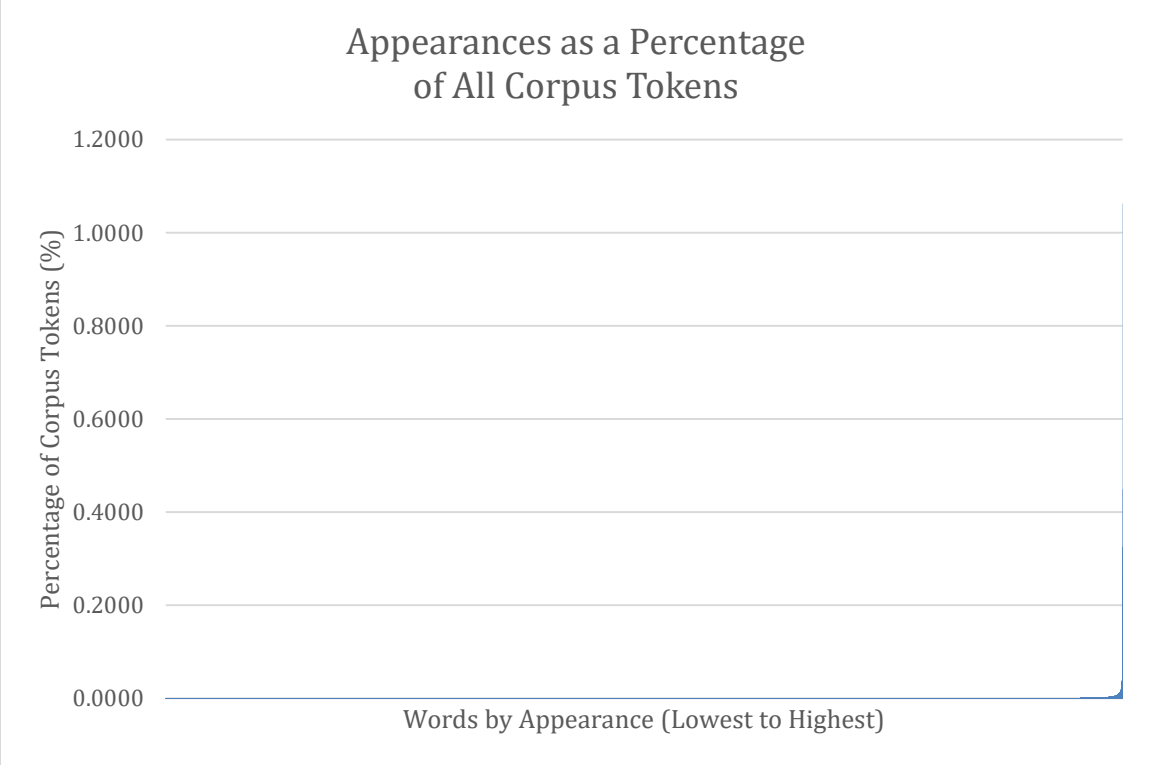

Figure 16: Appearance Ratio for All Tokens in the Corpus

The second statistic rejected for use was the statistic of **Appearances**. As a counting statistic, this number keeps growing without any relationship or correlation to

the overall corpus. Looking at this number by itself does not distinguish between a word appearing many times in one document or numerous times in one or a few documents.

How Analyses were Applied

Using these three measurements in combination with knowledge gained from the manual processing of documents and reviews of the documents terms were extracted from and it took 15-20 hours to identify terms from the entire corpus. In total, the process garnered in excess of 2,000 terms were identified as classes, identities and properties for APTs and their subcomponents. In retrospect, the process was not perfect. Chapter 5 includes a discussion and an assessment of the process.

Building the Ontology Structure

With the terms extracted from the OSINT it was a matter of building and populating the ontology. However, this is not as simple as the statement above might make it appear. The next subsection outlines the first step of organizing and identifying the terms and the role that they serve in the ontology. The second outlines how the organization of the ontology was decided upon and touches on some of the versions that were ultimately rejected. The last subsection provides a brief overview of how the structure moved from concept to reality.

Classifying and Categorizing Terms

The first step in the process required the identification of terms in one of several areas:

Class

Any term that is not a by name reference. Instead it describes a larger group to which subclasses or identities would belong. APT, itself was an easy class to identify and subclasses of APT are Organization and Attack. In some cases, classes were defined by groups of individual entities that share a common relationship. An example of this is the class Protocol. Under protocol were included the various protocols identified by name and/or port in the corpus (e.g. FTP, SSH, HTTP, etc.).

Individual

Individual terms are those terms that represent specific instances of a class. An example of this would be specific nations defined as Targets of a particular APT. Another would be the specific name of an APT Attack or APT Organization. From a programming standpoint, this is equivalent to the instantiation of a class.

Not every class has an individual because there are details that may not have been details that were not contained in a given whitepaper or even known by its author. The best example of these would-be personnel within an organization. The ontology includes the "job positions" with some definition but few whitepapers actually contained names of people within these organization. The way these organizations appear to function, names would probably change so frequently that to maintain that sort of information would be impossible.

Property

A property is simply an attribute of a class or individual entity. Properties of a class are placeholders for values that the individual instance will fill. A variety of variable types were used for these properties. Most properties used strings to specify values but in some instances, it was just as efficient and simple example of this is the port that a protocol uses to communicate. Some of the other properties are Boolean in nature.

There is one property that all classes (and thus individuals have) and that is the property labeled *Alias*. From an ontological standpoint, *aliases* are just a string variable but for purposes of this specific ontology they are actually a very important property. As ontologies are designed to standardize language, I thought it was important to include a property that includes any alternate versions of a term found within the corpus.

Table 10, below, shows just some of the variations for the term Command and Control found in the corpus. As discussed in the previous section on NLP, some of these lines might be the result of NLP processing but it is clear that there is enough variation in the way the term Command and Control is identified that it only seemed logical that these variations be recorded within the ontology.

Table 10: Aliases for Command and Control that Appearing in the Corpus

Lemma	Variation	Documents	Appearances
c&c	C&C	171	2306
c2	C2s	29	72
c2	C2	169	2378
c2host	C2host	1	1
c2host	C2host	1	1
c2infrastructure	C2infrastructure	1	1
c2s	C2s	2	2
commandandcontrol	commandandcontrol	1	2
command-and-control	command-and-control	1	1
command-and-control	command-and-controls	1	1
command-and-control	Command-and-control	81	150

Organizing the Ontology

With the terms identified and categorized, the ontology was ready to be organized into a structure. With such a large group of terms it was easier to work with them in small groups and this worked well with building classes as well. However, while Protégé may be an intuitive tool for some, it was not so in this case. Limited experience with the tool meant is was as much of a hindrance to the process as anything else.

Building the Draft Ontology

Given that data about the terms was already contained within a database and my familiarity with the tools that MySQL offers for database design, it made sense to create the structure within MySQL. Figure 17, below, shows the first draft of the ontology as developed in MySQL Workbench.

A familiarity with relational databases made it easy to conceptualize the work being done and the structure required. Tables represented classes. Columns represented properties that the classes possessed. Rows within the table became individual instantiations of the particular class. Foreign keys indicated relationships between classes and allowed for a many-to-many relationship exist between different classes.

Figure 18, below, shows the Protocol class. It has several properties including the port, aliases, whether it is encrypted, and with what type of encryption. The Protocol

Class has relationships with Legitimate Software (legal software used for illicit means) as well as Malware, Command-and-Control (not shown), etc.

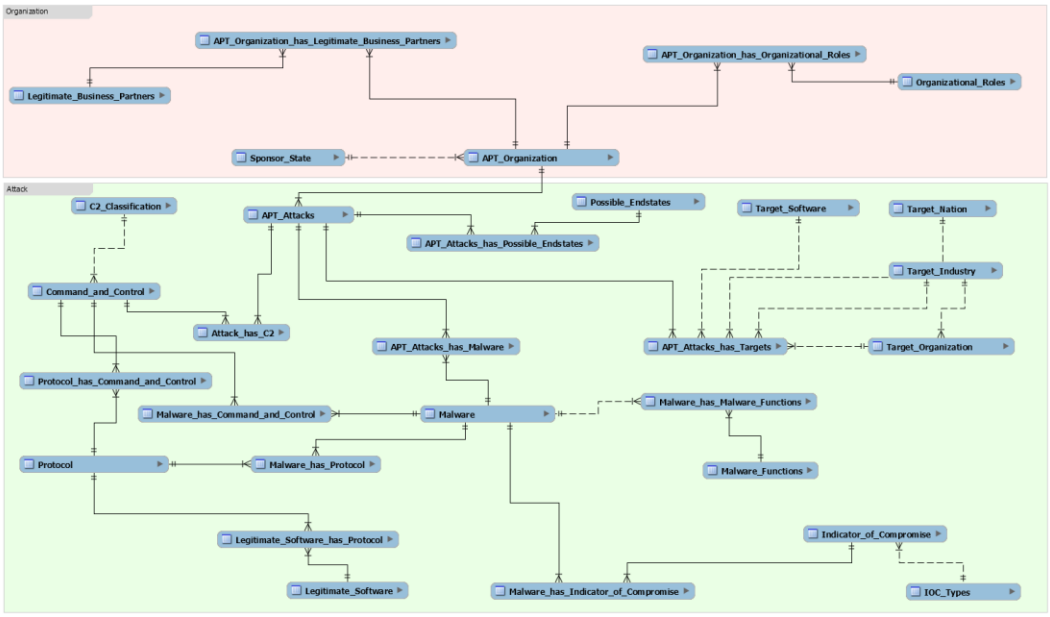

Figure 17: MySQL Database of Ontology Structure

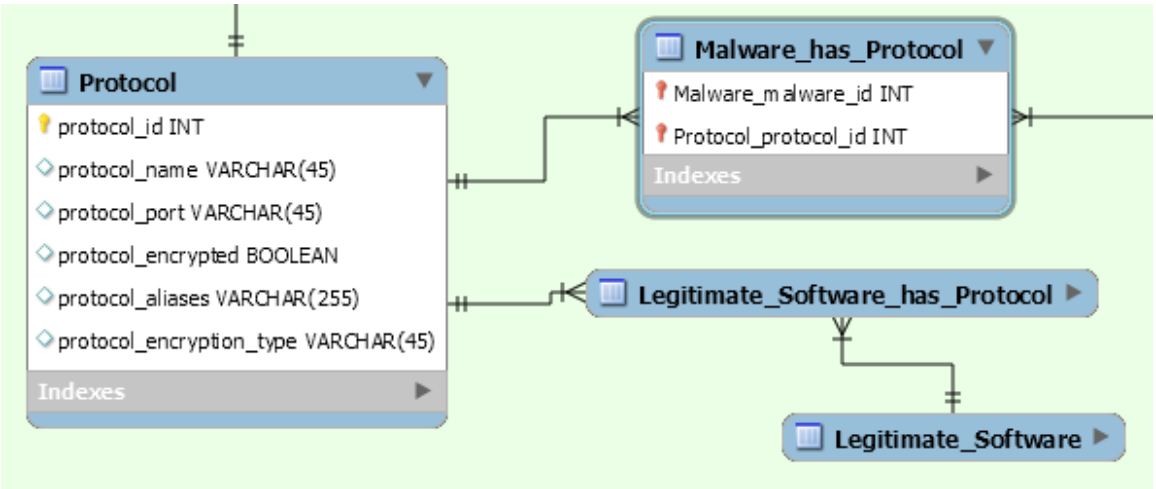

Figure 18: An example of a Class when viewed as a MySQL database table

After the draft was built it was easy to create the database. Records for individual identities were added to the appropriate tables along with all properties. Working with a familiar tool made for more expedient work and made for less issues with assigning information to their proper location. Once the work was complete within the database, the data was exported in XML format which is compatible with Protégé.

I was then able to import that structure into Protégé and make the structural changes required to account for something that I had not considered during the drafting process (e.g. an unforeseen relationship, or a new property).

Basic Structure

Figure 17, above, is divided into two sections as represented by the two background colors. The two regions represent the two subclasses of Advanced Persistent Threat. The upper third (in red) contains elements of the ontology which focus upon the Organizational subclass. The lower two-thirds (in green) contains elements related to the Attack subclass.

As the tables reflect classes it takes a little maneuvering to get them where they need to be when placed inside the ontology. While each class is represented by a table in MySQL, tables are not nested within tables to reflect the subclass nature that they represent in the ontology. Within Protégé it is a relatively easy process to move classes within the structure to create the class-subclass structure desired. Figure 19, below, is an illustration of nesting the table for Malware under the attack class.

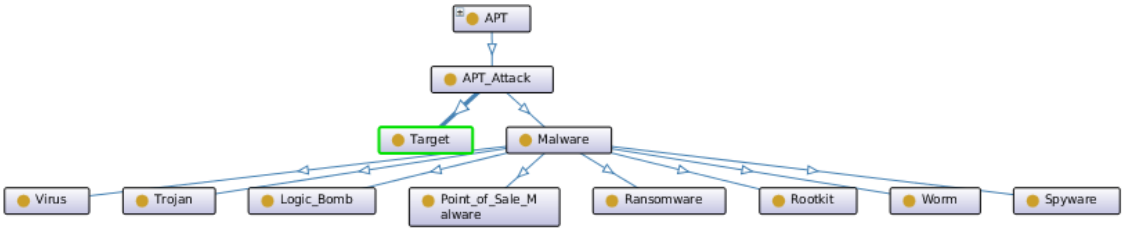

Figure 19: The class-subclass structure in Protégé

Figure 20, illustrates the nesting of subclasses within subclasses as the target tables for nations, industries, organizations and software are nested within a targets class that did not actually exist as a table within the database but made perfect sense for the organization of the ontology.

Figure 20: Nesting of Classes

Summary

In this chapter, we examined this paper's related research. It examined the steps taken to evaluate and merge the two open source knowledgebases to establish the corpus. We explored the various challenges both in processing the documents with CoreNLP and with importing the NLP output into a database for statistical analysis and identification of key terms for the ontology. It concluded with the logic and process used to identify the terms for the ontology and the steps taken to put those terms into the ontological format.

CONCLUSIONS AND RECOMMENDATIONS

The work for this research is done. Its objective was to establish an open source ontology for use in the Advanced Persistent Threat domain. In this chapter, we will review a summary of the study looking at the purpose, significance, and methodology. After which, some future research recommendations will be presented which can use the work done here both to take advantage of this work and to advance it such that its relevance to the cybersecurity community may continue beyond the present effort.

Summary of Study

This section is intended to be a retrospective review of the purpose and significance of the study, and a review of the methodology employed herein.

Purpose of Study

In a world where so much of our commerce, infrastructure, and even national defense depends on the communication pathways provided by the internet, failure to detect and mitigate the threat represented by APTs could impact businesses, civilian infrastructure, and potentially national security of all nations. Current CND measures are unable to reliably and regularly detect APTs that compromise both commercial and government networks. Part of the problem is incomplete understanding of these APTs and to see the various components of a single APT attack as parts of the larger whole.

The goal of this research was threefold. First, to collate OSINT found in the public domain about the individual APTs and from that knowledge build an ontology for the entire domain. Second, was to enable the standardization of language so as to facilitate easier communication about the domain. Third, was that this research could inform cybersecurity professionals with the knowledge that can be incorporated into the improvement of CND.

Significance of Study

As discussed in Chapter 1, the importance of the current research is how it can standardize terminology and definitions as well as providing a comprehensive informational resource for the cybersecurity and IT communities. Whether the ontology supports research designed to protect networks or to provides information that can be used when performing risk assessments and disaster recovery plans. Having a foundation of common terms and definitions for this domain will help share the results of disparate research findings and the development of business plans designed to protect networks from would-be attackers.

The variation of terms used in the small sample size as this was intriguing. In a profession where acronyms are used so frequently, the variation in what terms and acronyms are used and how they are used is analogous to two people speaking in different dialects or using colloquialisms that convey different means to each of the parties. This work has given me (and hopefully my reader) an appreciation for minimizing those differences need to be minimized in order to maximize communication and ultimately results.

Methodology Review

With the work complete and looking back at the path the research has taken, I believe that the methodology was sound but there are some realities that could have been better planned for. Because of challenges faced along the way minor adjustments were made in order to keep the work moving forward. As mentioned in Chapter 4, while MySQL was originally intended to be a solution for collating the output from the NLP process, in the end it offered me a path to get around challenges faced with Protégé and building the ontology.

When issues arose either with software errors or technology limitations, I found alternate paths around the problem. As a technology professional and a leader of Soldiers, it is what is expected when adversity presents itself. So, when the research faced a technological challenge, I found a way around it.

As a person who has worked in the technology field for almost 25 years, the opportunity to learn new software and new technologies has always been an exciting part of the job. I've rarely felt like I was not up to the challenge that comes from exploring new arenas. However, in this instance, the research faced challenges because for several factors which we will explore in the following subsections.

Learning Curve

As mentioned previously, I relish the opportunity to explore new technologies. However, several of the programs used in this research were not intuitive in a manner that foster quick learning and mastery. Learning the finer points of CoreNLP, BookNLP, and Protégé took far more time than I originally envisioned. As discussed in Chapter 4, the challenge of mastering Protégé made me have to rely more heavily on MySQL than I initially planned.

There were points during the research that I looked for other software options both for the processing of documents and building the ontology but with time as an always looming factor, it ultimately came down to plodding ahead with the software the research started with.

Software Challenges

Protégé and CoreNLP are both comprehensive programs and there are numerous plugins that enable greater functionality. With time constraints to complete this work and my limited programming experience in Java and Python, I believed that these numerous extensions of the two programs would enable me find a solution that time and my programming limitations prevented me from addressing myself.

However, as I quickly came to learn, many of these plugins were developed by researchers like myself. The solutions they developed address a particular research challenge they faced and had the time to develop themselves. This meant that some of the plugins were focused to meet a very specific task and thus were not a 100% solution as I hoped and trying to use two or three plugins to address my need became more of a quagmire than a help.

Also, in many cases when the research was done or the funding stopped the developers stopped updating the plugins. While the developers of the main software packages continue to release new versions the plugin developers do not and as the core software changes, the plugin may or may not continue to function with the newer version of the core software.

Another effect of these orphaned plugins was a challenge in finding support. One of the foundations of open source software development is that there is a community of users ready to help one another should one user find himself facing an issue he cannot solve on his own. However, when it comes to these orphaned pieces of software there may or may not be the people in the community who know how to help a user who is trying to get past a cryptic error.

In summary, the assumption that there would be solutions readily available to deal with any problem was a bigger assumption than should have been made. At least given the restraints placed on it. Given what I've learned over the past months, my recommendation to anyone doing this type of work is to allow the research team the time necessary to examine what software has been developed, what state that software is in, and what resources are available to resolve issues when they arise. With the summary of study complete, we will conclude this work by examining some recommendations for future related research.

Recommendations for Future Research

The work done here was designed to be as complete a work as can be done from a small sample of the content available on APTs as has been developed over the last decade. However, it was intended to be just the first step in a larger effort.

New APT organizations will develop new, and likely, more complex attacks and employ new techniques to defeat security measures implemented by cybersecurity professionals. Therefore, this ontology must continue to evolve, to expand, and, if necessary, be restructured in order to meet those new challenges.

The next three subsections recommend different ways in which the work here can be adapted and employed so as to continue to be a product that is useful in the open source domain.

The first section contains recommendations about means to expand this ontology and how to integrate it with other related ontologies in the public domain. The second section proposes the use of this ontology with the science of Link Analysis in order to potentially develop inferred understanding of new APTs in the future as well as means to insulate a network from such an attack. The second section also explores means of measuring the knowledge gained in this manner. The third, and final, subsection looks at applying this ontology in support of Cyber Resiliency Frameworks which were developed to help organizations protect their processes and their IT infrastructure from APT attacks.

Expand the Ontology

The work begun here should be only the first step. Ontologies are intended to be maintained and expanded as the domain they describe changes and grows. To let any ontology to atrophy is to degrade its usefulness and limit its meaningfulness. Therefore, this section, and its subsections, briefly explore ways to keep this ontology meaningful in the APT domain.

Incorporate Other Related Ontologies

The OWL2 standard for ontology development allows for ontologies to build upon one another and expand their usefulness while reducing the amount of work it takes to do so. During this research, a concerted effort was taken to build a unique product in large part because it had to stand on its own for academic scrutiny.

However, as we move past this specific research, efforts should be made to incorporate applicable ontologies. For example, malware ontologies and ontologies that look at the larger cybersecurity domain could provide meaningful information with terms, attributes and properties already written in standardized language.

Apply the Ontology for Further Ontology Development

This research used only two finite knowledgebases with which to establish the corpus but more information about previously known and new APTs are published

regularly. Using the ontology established herein, future research should explore ways for this ontology to enable searching for additional data.

Terms collected here (both general and identifying) should be used to find more published materials available in Open Source. This new information would then be incorporated back into the ontology and to keep it current with the changing face of the APT threat.

Improve Knowledge through Link Analysis

The commonality of terms demonstrated in the current research illustrates that there are similarities between APT campaigns. Whether different attack employ the same botnet, malware, or two campaigns are waged by the same group, inference can reveal information that may not be explicitly stated about an APT through the application of link analysis.

The next section examines the role of link analysis to improve or expand knowledge. The following section examines the role of knowledge management and suggests an approach for scoring or measuring the knowledge known vs. what is inferred through link analysis in the proposed future research.

The Role of Link Analysis

Also referred to as relationship extraction, it is the identification of relationships between two or more entities whether they are individuals, organizations, entities, etc. (Best, 2011). The relationship between them is derived from the context in which the reference is used. For example, if ABC Corporation's Finance Department is managed by Jenny Grier, VP of Finance and JJ and Anna Glass are Billing Managers within the department then there is a boss-employee relationship or association between them as illustrated below in Figure 21: An Example of Link Analysis.

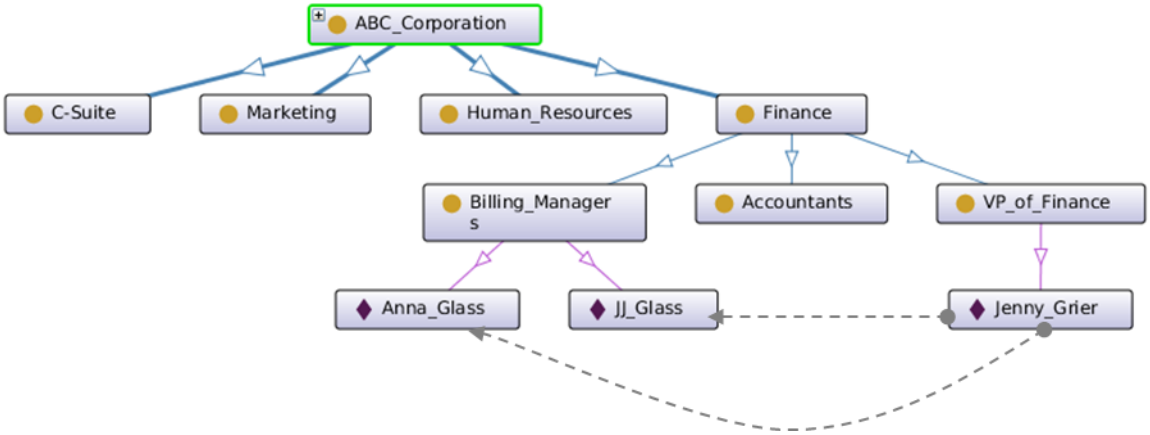

Figure 21: An Example of Link Analysis

The research performed my Ben-Dov, et.al. (Ben-Dov, Wu, Cairns, & Place, 2004) explores the use of link analysis and inference from that analysis can produce increased knowledge. Their efforts were an extension of prior done by Davies, and Swanson and Smalheiser (Ben-Dov et al., 2004). One of the biggest challenges to their research is that contemporary link-analysis tools operated on structured data (Ben-Dov et al., 2004). They overcame the obstacle by processing text through information extraction or text mining as a precursor to the link analysis.

For their research, Ben-Dov, et.al. (Ben-Dov et al., 2004) used two different methods for text mining that the present research must consider. Co-occurrence links uses pattern matching to determine if target phrases exist within a sentence. However, the mechanism does not apply any semantics or syntactic logic to determine how the two terms relate to one another. The second method that they employed Semantic links work by connect noun phrases and verb identification with the application of linguistic and semantic constraints. They accomplished this by using Declarative Information Analysis Language developed at ClearForest Labs (Ben-Dov et al., 2004).

The Ben-Dov et al. research is part of a larger domain referred to as Network Science. It is identified as an emerging discipline which examines the interconnections among diverse entities (Börner, Sanyal, & Vespignani, 2007). The use of the work network is not intended to limit this science to computer networks. It is intended to describe interrelations between any group of nodes or entities. For example, these could be biological entities within a biosphere, or the relationship between cited research within this dissertation (Börner et al., 2007). As Börner et al. notes, Network Science “aims to

develop theoretical and practical approaches and techniques to increase our understanding of natural and man made[sic] networks (Börner et al., 2007, p. 538)."

Measuring Knowledge

Knowledge management is a relatively new field of study that grew out of the change to the American business world from primarily manufacturing and industrial to a structure that focuses on service-oriented business. This change created the demand within business for companies to better measure a wealth built upon the knowledge and experience of its employees.

However, the means for measuring and, more importantly, valuing the knowledge as an asset to a given company was a challenge. Something as tacit as knowledge was difficult to quantify and even harder to put a dollar value on.

The following subsections provide a very brief overview of the thinking that businesses apply in the measurement of knowledge and provides a simple recommendation for measuring knowledge gained through inference in the proposed future research.

A Brief History of Knowledge Management

The shift in business models and the lack of a standard system for valuing knowledge created a gap which needed to be filled. academia and business economics savvy individuals stepped in to fill the need and close the gap. This section presents a review of literature pertaining to these efforts.

Human Resources Accounting (HRA)

Initially developed by accounting theorists, HRA began as an effort to address the value of HC to an organization and as a managerial tool (Flamholtz, Bullen, & Wei, 2002). HRA has seen its popularity grow and wane over the last five decades (Flamholtz et al., 2002).

HRA measures the Human Capital (HC) within an organization. It treats HC as an asset and quantifies the value of the asset in terms of an individual's intelligence, skills, and expertise (Bontis, Dragonetti, Jacobsen, & Roos, 1999; Mahmudul Hoque, 2010).

This model provides input for managerial and financial decisions (Mahmudul Hoque, 2010). Similar to the valuation of equipment or other assets within the company, HRA considers "the historical, acquisition, replacement, or opportunity cost of human assets (Bontis et al., 1999, p. 393)." HRA has three major function (Flamholtz et al., 2002):

- Provides numerical information about the cost and value of people to an organization;
- Serves as an analytical framework to facilitate decision making; and
- Motivates decision-makers to adopt a human-resources perspective.

Balanced Scorecard (BSC)

As with HRA discussed in the previous subsection., Kaplan and Norton designed BSC to function as both a means to calculate knowledge within an organization but also to facilitate the decision making process (Bontis et al., 1999). BSC tracks numerous dimensions within an organization in a systematic way.

BSC was developed from a multi-year, multi-company study sponsored by Harvard Business School (Bontis et al., 1999). Unlike some of the other models discussed, BSC includes factors both internal and external to the organization (Bontis et al., 1999; Moore, Rowe, & Widener, 2001).

The authors intended the model to challenge organizations to reinterpret the vision for the business as well as the business' long-term strategy in terms of four specific perspectives (Bontis et al., 1999). In order to accomplish this reinterpretation, the authors intended top management to communicate across organizational units and develop aligned strategies for the entire organization, and to communicate these new strategies to all levels of the organization.

Kaplan and Norton recommend four specific perspectives for measurement: (1) financial perspective which includes traditional accounting; (2) customer perspective which takes a marketing styled approach toward those groups that the organization targets for its services; (3) internal business which focuses on the concept of the value chain; and (4) learning and growth which includes measurements related to employees and the systems the company has in place to facilitate learning and knowledge sharing (Bontis et al., 1999; Moore et al., 2001; Patton, 2007).

Intellectual Capital (IC)

Although termed as capital there is nothing in IC that encompasses the conventional concept from economics or accounting (Kim, Yoo, & Lee, 2011). Kim, et.al. describe it as a "non-monetary asset without physical substance that can reap economic benefit (Kim et al., 2011, p. 2244)." IC is composed of HC, Organizational Capital (OC), and Customer Capital (CC).

In addition to what was previously discussed, Kim, et.al. notes that another driving force in HC relates to employees' commitment to the organization's mission and their satisfaction with the role that they fill within the organization. Bontis maintains that this is a source of creativity and innovation (Bontis et al., 1999).

OC is comprised of the organization's routines and processes which give it its competitive advantage. Management philosophy, culture, and information technology contribute to OC as well (Kim et al., 2011).

CC refers to the organization's relationships within the market. Of these three components, CC has the most direct effect on the company's value and performance (Kim et al., 2011). In its comparison of several other research efforts into IC, Kim et.al. determined that there are multiple dimensions which contribute to the valuation of IC within an organization and that each of those dimensions has numerous sub-dimensions which contribute to each. Several of which were outlined above.

Skandia's model used 91 IC metrics in addition to 73 traditional metrics to calculate the company's IC assets (Bontis, 2001). When they standardized the model to create a *universal IC report*, the number of metrics expanded to 112 (Bontis, 2001).

Economic Value Added™ (EVA™)

As a knowledge measurement, EVA™ (hereafter referred to as EVA which is a registered trademark) is a tool closely related to the financial concept of Residual Income (RI) (Bontis et al., 1999; Ryan, 2011). RI is the value remaining in an organization after all other capital investors' have been compensated and other factors have been accounted for (i.e. depreciation) (Ryan, 2011). EVA stands as both a common language and means

of benchmarking the value-creation of intangible assets. Proper knowledge management will also increase EVA within the company (Bontis et al., 1999).

Stern Stewart and Company developed EVA in the late 1980s as a "tool to assist corporations to pursue their prime financial directive [and aid] in maximizing the wealth of their shareholders (Bontis et al., 1999, p. 394)." However, the concept was by no means revolutionary. In the 1920s, General Motors applied the concept of RI to measure the performance of divisions within the company (Bontis et al., 1999; Ryan, 2011).

Ryan identifies three reasons for the adjustments that differentiate EVA from RI:

- "To convert from accrual to cash accounting. Investors are interested in cash flows, so many of the accounting adjustments made, such as allowances for doubtful debt, should be eliminated (Ryan, 2011, p. 3)"
- "Spending on 'market building' items such as research, staff training and advertising costs should be capitalised [sic] to the extent that they have not been in the financial statements (Ryan, 2011, p. 3)."
- All items of profit and expenditure should be included (Ryan, 2011).

This concludes the brief overviews of the business world's most common knowledge management systems.

Recommended Solution for Measuring APT Knowledge

The ontology developed in the present research along with any modifications and expansions in previous sections of this chapter provides an excellent template for measuring the knowledge collected about individual APTs and about the APT domain. The next subsection makes a general suggestion regarding any scoring method that might be employed. The second and third subsections look at a simple and weighted scoring approach.

When intelligence about an APT is discovered through link analysis they should be treated as 'unverified.' Therefore, the same weight should not be applied to these data points as to documented data points as this could potentially bias the research and skew the final score of each APT and the overall results of the study.

Therefore, the recommendation would be to modify whatever score is given to these 'unverified' data points. If later expansion of the knowledge base provides

documented proof that an inferred data point is accurate, and thus verified, the modifier should then be removed and the APT's scorecard adjusted accordingly.

Basic Scoring Approach

These attributes may have a one-to-one or many-to-one relationship within the vertical relationships of the ontology. Meaning there may be more than one data point for each attribute for a given APT or there may be none.

As an example of the many-to-one relationship, a given APT could employ multiple means of delivery in order to inject the APT into a target network (e.g., removable media, email attachment, malicious site, etc.). Knowing each of those means of delivery means that there are more data known about a given threat. Therefore, each data point should be counted individually when scoring each APT.

Weighted Scoring Approach

The compelling thought supporting such a scoring methodology rested in the concept that some data points would be more useful for IDS and IPS detection. The researcher even considered different ways to rank the importance of different data points including surveying a subset of the cybersecurity profession with emphasis on experience with APT detection.

A weighted scoring system certainly has a place in further research where one places a value upon the 'usefulness' of data based on detection criteria. Should future research delve into the application of threat intelligence for APT detection, then modifiers based on the usefulness of the data should be employed.

Apply the Ontology to Support Cyber Resiliency

Within the last three years, MITRE and NIST have developed cyber resiliency models with the goal of enabling organizations to build resiliency into their IT and business infrastructure and to address ongoing threats like APTs. As the goal of these methodologies is to ensure the survival and rapid recovery of the entire business, the ontology developed here could provide the kind of standardized language needed to

properly integrate IT risk management plans within the risk management plans for the business and support the application of the Cyber Resiliency methodologies.

The use of an ontology in work like risk management and resiliency planning would reduce the time it takes to develop and implement plans. It would facilitate improved communications and effectiveness of training. Ultimately, should the need arise, the execution of these plans could be done more smoothly both because of the common language and the increased knowledge that the ontology could provide.

Summary

In this chapter, we examined the purpose and significance of the present study. We also reviewed the methodology and critiqued its design. After the critique was concluded, a series of further research efforts were recommended in order to make use of the work begun here. The recommendations for further research include the expansion of the ontology through expanded search for new content and the integration of applicable ontologies available in the public domain. Further recommendations were made, which involved the use of inferential analysis and a means for measuring the amount of knowledge possessed about both the domain and individual APTs by applying tools and techniques from the field of Knowledge Management.

REFERENCES

Abulaish, M., & Dey, L. (2007). Interoperability among distributed overlapping ontologies - A fuzzy ontology framework. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), WI'06*, 397–403. http://doi.org/10.1109/WI.2006.105

Alperovitch, D. (2011). Revealed: Operation Shady RAT. *White Paper*, 1–14.

Armerding, T. (2012). SANS Digital Forensics and Incident Response Blog | Advanced Persistent Threats Can Be Beaten | SANS Institute. misc. Retrieved from https://digital-forensics.sans.org/blog/2012/08/10/advanced-persistent-threats-can-be-beaten

Arsene, L. (2015). The Anatomy of Advanced Persistent Threats - Dark Reading. Retrieved April 28, 2016, from http://www.darkreading.com/partner-perspectives/bitdefender/the-anatomy-of-advanced-persistent-threats/a/d-id/1319525

Ashford, W. (2011). RSA hit by advanced persistent threat attacks. Retrieved April 30, 2016, from http://www.computerweekly.com/news/1280095471/RSA-hit-by-advanced-persistent-threat-attacks

Ask, M., Bloemerus, P., Bondarenko, P., Nordbø, A., Rekdal, J. E., Rekdal, J. E., … Rekdal, J. E. (2013). *Advanced Persistent Threat (APT) Beyond the Hype*. Gjøvik.

Auty, M. (2015). Anatomy of an Advanced Persistent Threat. *Network Security*, *2015*(4), 13–16. article. http://doi.org/10.1016/S1353-4858(15)30028-3

Bamman, D., Underwood, T., & Smith, N. A. (2014). A Bayesian Mixed Effects Model of Literary Character. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 370–379.

Bejtlich, R. (n.d.). TaoSecurity: What Is APT and What Does It Want? Retrieved April 27, 2016, from http://taosecurity.blogspot.se/2010/01/what-is-apt-and-what-does-it-want.html

Bejtlich, R. (2013). China's "Advanced Persistent Threat" to American Computer Networks. *Hampton Roads International Security Quarterly*, (January), 16–19. article. http://doi.org/http://dx.doi.org/10.1108/17506200710779521

Ben-Dov, M., Wu, W., Cairns, P. A., & Place, A. (2004). Improving Knowledge Discovery By Combining Text-Mining And Link-Analysis Techniques Semantics links. *Foundations*, (Davies), 1–7.

Best, C. (2011). Challenges in open source intelligence. *Proceedings - 2011 European Intelligence and Security Informatics Conference, EISIC 2011*, 58–62. http://doi.org/10.1109/EISIC.2011.41

Bird, S., Loper, E., & Klien, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bisson, D. (2015). The OPM Breach: Timeline of a Hack. Retrieved April 30, 2016, from http://www.tripwire.com/state-of-security/security-data-protection/cyber-security/the-opm-breach-timeline-of-a-hack/

Bodeau, D., Graubart, R., Heinbockel, W., & Laderman, E. (2014). *Cyber Resiliency Engineering Aid – Cyber Resiliency Techniques: Potential Interactions and Effects Deborah Bodeau Approved By* (techreport). Bedford.

Bontis, N. (2001). Assessing knowledge assets: a review of the models used to measure intellectual capital. *International Journal of Management Reviews*, *3*(1), 41–60. http://doi.org/10.1111/1468-2370.00053

Bontis, N., Dragonetti, N. C., Jacobsen, K., & Roos, G. (1999). The Knowledge Toolbox: A Review of the Tools Available to Measure and Manage Intangible Resources. *European Management Journal*, *17*(4), 391–402. http://doi.org/16/S0263-2373(99)00019-5

Börner, K. (2007). Making sense of mankind's scholarly knowledge and expertise: Collecting, interlinking, and organizing what we know and different approaches to mapping (network) science. *Environment and Planning B: Planning and Design*, *34*(5), 808–825. http://doi.org/10.1068/b3302t

Börner, K., Sanyal, S., & Vespignani, A. (2007). Network Science. *Annual Review of Information Science & Technology*, *41*, 537–607. Retrieved from http://ivl.cns.iu.edu/km/pub/2007-borner-arist-s.pdf

Bradbury, D. (2010). Shadows in the cloud: Chinese involvement in advanced persistent threats. *Network Security*, *2010*(5), 16–19. article. http://doi.org/10.1016/S1353-4858(10)70058-1

Brand, M., Valli, C., & Woodward, A. (2010). Malware Forensics: Discovery of the Intent of Deception. *Journal of Digital Forensics, Security and Law*, *5*(4), 31–42. article. Retrieved from http://ojs.jdfsl.org/index.php/jdfsl/article/view/142

Brill, A. E. (2010). From Hit and Run to Invade and Stay: How Cyberterrorists Could Be Living Inside Your Systems. *Defence Against Terrorism Review*, *3*(2), 23–36.

Burwell, H. P. (1999). *Online Competitive Intelligence: Increase Your Profits Using Cyber-Intelligence*. (M. Sankey & C. R. Ernst, Eds.). book, Facts on Demand Press.

Card, K. L., & Rogers, M. S. (2012). *Navy Cyber Power 2020* (techreport).

Chandran, S., P, H., & Poornachandran, P. (2015). An Efficient Classification Model for Detecting Advanced Persistent Threat. In *2015 International Conference on Advances in Computing, Communications and Informatics* (pp. 2001–2009). inproceedings, IEEE.

Cole, E. (2012). *Advanced Persistent Threat: Understanding the Danger and How to Protect Your Organization*. *Google Books*. book, Waltham: Syngress.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, *12*, 2493–2537. http://doi.org/10.1.1.231.4614

Coviello, A. (n.d.). Open Letter to RSA Customers. Retrieved April 30, 2016, from https://www.sec.gov/Archives/edgar/data/790070/000119312511070159/dex991.htm

Damballa. (2010). Advanced Persistent Threats (APT). article. Retrieved from https://www.damballa.com/downloads/r{_}pubs/advanced-persistent-threat.pdf

Davies, M. (2016). Word frequency: based on 450 million word COCA corpus. Retrieved from http://www.wordfrequency.info/free.asp?s=y

Dey, L., Rastogi, A. C., & Kumar, S. (2007). Generating concept ontologies through text mining. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 23–29. http://doi.org/10.1109/WI.2006.86

Egele, M., Scholte, T., Kirda, E., & Kruegel, C. (2012). A survey on automated dynamic malware-analysis techniques and tools. *ACM Computing Surveys*, *44*(2), 1–42. article. http://doi.org/10.1145/2089125.2089126

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *In Acl*, (1995), 363–370. http://doi.org/10.3115/1219840.1219885

Flamholtz, E. G., Bullen, M. L., & Wei, H. (2002). Human resource accounting: A historical perspective and future implications. *Management Decision*, *40*(10), 947. http://doi.org/10.1108/00251740210452818

Fleisher, C. (2008). Using open source data in developing competitive and marketing intelligence. *European Journal of Marketing*, *42*(April), 852–866. http://doi.org/10.1108/03090560810877196

Hake, T., & Vaishalideshmukh, P. (n.d.). Building of Domain Ontology using NLP, 220–225.

Hitzler, P., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). OWL 2 Web Ontology Language Primer. *W2C*, (December), 1–123.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroe, C. (2004). *A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools*. *The University Of Manchester*. Manchester: The University of Manchester.

Huang, H. De, Acampora, G., Loia, V., Lee, C. S., & Kao, H. Y. (2011). Applying FML and Fuzzy Ontologies to malware behavioural analysis. *IEEE International Conference on Fuzzy Systems*, 2018–2025. http://doi.org/10.1109/FUZZY.2011.6007716

Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *6th Annual International Conference on Information Warfare and Security*, (July), 1–14. techreport. Retrieved from http://papers.rohanamin.com/wp-content/uploads/papers.rohanamin.com/2011/08/iciw2011.pdf$\$nhttp://www.lockheedmartin.com/content/dam/lockheed/data/corporate/documents/LM-White-Paper-Intel-Driven-Defense.pdf

Information Warfare Monitor, & Shadowserver Foundation. (2010). *Shadows in the Cloud: Investigating Cyber Espionage 2.0*.

International Business Machines. (n.d.). IBM Watson: What is Watson? Retrieved September 1, 2016, from http://www.ibm.com/watson/what-is-watson.html

ISACA. (n.d.-a). APT. Retrieved April 28, 2016, from https://cybersecurity.isaca.org/csx-threats-and-controls/threats/apt

ISACA. (n.d.-b). Malware. Retrieved April 28, 2016, from https://cybersecurity.isaca.org/csx-threats-and-controls/threats/malware

ISACA. (n.d.-c). Social Engineering. Retrieved April 28, 2016, from https://cybersecurity.isaca.org/csx-threats-and-controls/threats/social-engineering

Jones, K. S. (1999). What is the role of NLP in text retrieval. *Natural Language Information Retrieval*, (March 1997), 1–12. Retrieved from http://www.cl.cam.ac.uk/users/sht25/ksj5.pdf

Karoui, L., Aufaure, M. A., & Bennacer, N. (2007). Context-based hierarchical clustering for the ontology learning. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), WI'06*, 420–427. http://doi.org/10.1109/WI.2006.55

Kaspersky Lab Global Research and Analysis Team. (2013). The Icefog APT: A Tale of Cloak and Three Daggers. Retrieved April 29, 2016, from https://securelist.com/blog/research/57331/the-icefog-apt-a-tale-of-cloak-and-three-daggers/

Kim, T. (Terry), Yoo, J. J.-E., & Lee, G. (2011). The HOINCAP scale: measuring intellectual capital in the hotel industry. *The Service Industries Journal*, *31*(13), 2243–2272. http://doi.org/10.1080/02642069.2010.504817

Knublauch, H., Fergerson, R., Noy, N., & Musen, M. (2004). The Protege-OWL Plugin: An Open Development Environment for Semantic Web Applications. *Third International Semantic Web Conference, Hiroshima, Japan*.

Kushner, D. (2013). The Real Story of Stuxnet. *IEEE Spectrum*, (March), 48–53.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics*, 28–34.

Lundquist, D., Zhang, K., & Ouksel, A. (2015). Ontology-driven cyber-security threat assessment based on sentiment analysis of network activity data. *Proceedings - 2014 International Conference on Cloud and Autonomic Computing, ICCAC 2014*, 5–14. http://doi.org/10.1109/ICCAC.2014.42

Mahmudul Hoque. (2010). Methods and Accounting Treatment of HRA. Retrieved July 7, 2016, from http://www.articlesbase.com/international-studies-articles/methods-and-accounting-treatment-of-hra-2773625.html

Mandiant. (2010). M-Trends: The Advanced Persistent Threat. White Paper, Milpitas: FireEye. http://doi.org/10.1049/etr.2014.0025

Mandiant. (2013). APT1 Exposing One of China's Cyber Espionage Units. Milpitas: FireEye. Retrieved from http://intelreport.mandiant.com/Mandiant_APT1_Report.pdf

Mandiant. (2014). M-Trends: Beyond the Breach. White Paper, Milpitas: FireEye.

Manning, C. D., Bauer, J., Finkel, J., Bethard, S. J., Surdeanu, M., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. http://doi.org/10.3115/v1/P14-5010

McAfee. (2011). Global Energy Cyberattacks: "Night Dragon," 19. Retrieved from www.mcafee.com

Meckl, S., Tecuci, G., Boicu, M., & Marcu, D. (2015). Towards an Operational Semantic Theory of Cyber Defense Against Advanced Persistent Threats. In *STIDS 2015 Proceedings* (pp. 58–65).

Mikroyannidis, A., & Theodoulidis, B. (2007). Heraclitus II: A framework for ontology management and evolution. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), WI'06*, 514–521. http://doi.org/10.1109/WI.2006.90

Moore, C., Rowe, B. J., & Widener, S. K. (2001). HCS: Designing a Balanced Scorecard in a Knowledge-Based Firm. *Issues in Accounting Education*, *16*(4), 569–601. http://doi.org/10.2308/iace.2001.16.4.569

Mundie, D. A., & McIntire, D. M. (2013). An Ontology for Malware Analysis. *Proceedings - 2013 International Conference on Availability, Reliability and Security, ARES 2013*, 556–558. http://doi.org/10.1109/ARES.2013.73

Musen, M. A., & The Protégé Team. (2015). The Protégé Project: A Look Back and a Look Forward. *AI Matters*, *116*(4), 4–12. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/

Naval, S., Laxmi, V., Rajarajan, M., Gaur, M. S., & Conti, M. (2014). Employing Program Semantics for Malware Detection, *10*(12), 2591–2604. article. http://doi.org/10.1109/TIFS.2015.2469253

Navigli, R., & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, *30*(2), 151–179. http://doi.org/10.1162/089120104323093276

Patton, J. R. (2007). Metrics for Knowledge-Based Project Organizations. *SAM Advanced Management Journal*, *72*(1), 33–43. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=24688728&site=ehost-live

Perlroth, N. (2013). Chinese Hackers Infiltrate New York Times Computers - The New York Times. Retrieved April 29, 2016, from http://www.nytimes.com/2013/01/31/technology/chinese-hackers-infiltrate-new-york-times-computers.html?_r=0

Ponemon Institute. (2015). 2015 Cost of Data Breach Study: Global Analysis, (June), 1–19.

PR Newswire. (2015, October 15). Advanced Persistent Threat Protection Market Worth 8.7 Billion USD by 2020. *PR Newswire*, p. 2. Pune.

Radev, D. R. (2015). 01.01 - Introduction (8:38) - University of Michigan | Coursera. Retrieved September 1, 2016, from https://www.coursera.org/learn/natural-language-processing/lecture/oX0lw/01-01-introduction-8-38

Radzikowski, P. S. (2016). CyberSecurity: Expanded Look at the APT Life Cycle and Mitigation. Retrieved from http://drshem.com/2016/02/11/cybersecurity-expanded-look-apt-life-cycle-mitigation/

Raj, V. S., Chezhian, R. M., & Mrithulashri, M. (2014). Advanced Persistent Threats & Recent High Profile Cyber Threat Encounters. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(1), 2412–2417.

RSA FraudAction Research Labs. (2011). Anatomy of an Attack - Speaking of Security - The RSA Blog and Podcast. Retrieved April 30, 2016, from https://blogs.rsa.com/anatomy-of-an-attack/

Ryan, N. (2011). *Economic value added versus profit-based measures of performance* (Vol. 900).

Saarinen, M. O. (2013). Developing a Grey Hat C2 and RAT for APT Security Training and Assessment. In *GreHack 2013* (pp. 12–24). Grenoble: GreHack.

Shosha, A. F., James, J. I., Hannaway, A., Liu, C.-C., & Gladyshev, P. (2012). Digital Forensics and Cyber Crime. In M. Rogers & K. C. Seigfried-Spellar (Eds.), *Digital Forensics and Cyber Crime* (pp. 66–80). incollection, Lafayette: Springer.

Sikorski, M., & Honig, A. (2012). *Practical Malware Analysis: The Hands-on Guide to Dissecting Malicious Software*. book.

Sims, S. (2015). Everything You Need to Know about Natural Language Processing. Retrieved September 1, 2016, from http://www.kdnuggets.com/2015/12/natural-language-processing-101.html

Stanford NLP Group. (n.d.-a). Stanford CoreNLP – a suite of core NLP tools | Stanford CoreNLP. Retrieved from http://stanfordnlp.github.io/CoreNLP/

Stanford NLP Group. (n.d.-b). The Stanford Natural Language Processing Group. Retrieved from http://nlp.stanford.edu/

Steele, R. D. (2007). Open source intelligence. In L. K. Johnson (Ed.), *Handbook of Intelligence Studies* (pp. 129–147). New York: Routledge.

TrendLabs. (n.d.). Connecting the APT Dots - TrendLabs Security Intelligence Blog. Retrieved April 30, 2016, from http://blog.trendmicro.com/trendlabs-security-intelligence/connecting-the-apt-dots-infographic/

United States. The White House. (2013). Executive Order 13636: Improving Critical Infrastructure Cybersecurity. *Federal Register*, *78*(33), 1–8. article.

Velardi, P., Fabriani, P., & Missikoff, M. (2001). Using text processing techniques to automatically enrich a domain ontology. *Formal Ontology in Information Systems. IOS Press*, 270–284. http://doi.org/10.1145/505168.505194

Velardi, P., Faralli, S., & Navigli, R. (2013). OntoLearn Reloaded: AGraph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, *39*(3), 665–707.

Villeneuve, N., & Bennett, J. (2012). *Detecting APT Activity with Network Traffic Analysis*. Cupertino. Retrieved from http://www.trendmicro.pl/cloud-content/us/pdfs/security-intelligence/white-papers/wp-detecting-apt-activity-with-network-traffic-analysis.pdf

W3C OWL Working Group. (2012). OWL 2 Web Ontology Language Document Overview. *OWL 2 Web Ontology Language*, (December), 1–7. Retrieved from http://www.w3.org/TR/owl2-overview/

Waldron, M. (n.d.). 10 Common NLP Terms Explained for the Text Analysis Novice - AYLIEN. Retrieved October 1, 2016, from http://blog.aylien.com/10-common-nlp-terms-explained-for-the-text/

Wang, Y., Berant, J., & Liang, P. (2015). Building a Semantic Parser Overnight. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1332–1342. Retrieved from http://www.aclweb.org/anthology/P15-1129

Websense. (2011). Advanced Persistent Threats and Other Advanced Attacks. White Paper, Websense. Retrieved from http://www.websense.com/assets/white-papers/whitepaper-websense-advanced-persistent-threats-and-other-advanced-attacks-en.pdf

APPENDIX A. CORPUS STATISTICS

The data in this appendix represents the comparison of key document characteristics both in their base form (0 words extracted) and their adjusted form with the 5,000 most common words removed.

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
1	2,032	390	161	614	368	661	380	8	7	801	680
2	26,071	4,093	1,012	2,590	1,552	2,996	1,611	12	12	14,314	6,652
3	3,652	790	77	926	397	1,046	412	11	10	2,077	708
4	1,648	150	20	480	130	543	133	10	8	1,199	279
5	1,940	301	43	578	266	626	270	12	10	1,078	518
6	1,549	209	24	506	200	540	204	10	8	979	337
7	604	72	14	285	106	312	108	7	7	404	114
8	5,024	349	155	791	383	880	393	10	9	3,272	1,248
9	1,308	226	105	421	198	445	198	10	8	638	339
10	3,904	532	5	481	173	548	182	10	8	2,491	876
11	1,024	117	18	363	109	408	114	7	6	694	195
12	829	98	13	340	154	369	156	10	9	497	221
13	21,124	2,078	326	2,975	1,446	3,552	1,511	12	12	14,571	4,149
14	3,190	359	45	851	344	928	353	8	8	2,057	729
15	2,288	262	37	775	325	850	335	13	11	1,428	561
16	3,541	786	191	865	452	935	454	12	10	1,655	909
17	2,392	515	177	649	435	685	449	9	8	742	958
18	4,055	857	521	782	432	856	449	12	11	1,426	1,251
20	2,517	266	107	793	416	879	423	11	10	1,397	747

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
21	1,911	211	29	708	295	759	302	10	9	1,227	444
22	5,151	1,140	156	1,205	654	1,330	669	11	10	2,349	1,506
23	12,056	1,459	209	1,972	1,105	2,335	1,164	11	10	7,355	3,033
24	6,086	862	72	1,333	571	1,504	581	11	8	3,820	1,332
25	3,903	527	127	971	429	1,073	443	9	8	2,302	947
26	3,758	444	128	728	336	846	360	11	8	2,228	958
27	6,523	875	127	1,314	769	1,476	787	13	11	3,581	1,940
28	1,652	170	31	638	259	676	258	9	9	1,077	374
29	1,683	239	18	566	244	620	247	12	10	994	432
30	1,998	240	50	638	254	705	262	11	10	1,246	462
31	2,655	630	59	675	341	754	355	8	7	1,241	725
32	1,713	190	15	576	213	635	224	9	8	1,117	391
33	2,454	331	114	608	283	655	284	11	9	1,287	722
34	2,466	500	17	813	452	883	458	11	10	1,200	749
35	1,694	172	8	561	235	634	248	10	8	1,157	357
36	629	63	3	278	80	302	79	7	7	467	96
37	2,748	355	50	695	269	791	280	11	9	1,784	559
38	5,934	935	250	1,310	598	1,475	608	10	9	3,474	1,275
39	7,853	1,281	106	1,188	332	1,441	357	11	10	5,629	837
40	1,756	389	58	486	174	543	190	10	10	1,003	306
41	3,914	450	39	944	391	1,087	404	11	9	2,532	893
42	2,832	619	61	885	548	965	558	12	10	1,357	795
43	5,959	920	99	1,552	907	1,707	930	12	11	3,102	1,838
44	24,136	3,713	1,249	3,252	1,761	3,763	1,814	13	13	13,280	5,894
45	11,058	2,685	772	1,361	817	1,500	842	9	8	4,398	3,203

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
46	15,089	1,955	192	2,505	1,438	2,867	1,481	13	11	9,070	3,872
47	6,557	989	74	1,453	672	1,654	698	10	8	4,152	1,342
48	1,375	150	11	466	166	516	174	10	8	910	304
49	8,725	1,033	193	1,892	1,045	2,108	1,058	11	9	5,011	2,488
50	15,630	1,856	492	2,277	1,272	2,629	1,324	12	12	9,365	3,917
51	1,558	203	43	511	220	569	229	10	10	944	368
52	1,755	257	45	617	294	686	300	11	9	1,001	452
53	2,662	592	59	690	352	746	356	12	11	1,243	768
54	3,438	553	96	968	495	1,085	512	10	9	1,931	858
55	3,260	365	153	787	290	850	297	11	9	2,139	603
56	1,338	158	6	500	186	553	190	10	8	919	255
57	7,315	1,486	146	1,303	681	1,485	713	12	11	3,762	1,921
59	1,066	192	16	402	224	419	223	7	7	475	383
60	3,341	624	67	982	544	1,079	565	11	11	1,639	1,011
61	2,731	550	58	707	307	780	320	12	11	1,523	600
62	3,810	475	28	864	407	997	426	10	9	2,317	990
63	705	150	9	287	125	302	125	6	6	367	179
64	687	75	15	232	85	253	87	9	7	420	177
65	3,913	526	107	814	371	922	383	11	11	2,401	879
66	8,569	1,773	244	1,697	1,060	1,891	1,097	12	11	4,190	2,362
67	9,526	1,738	293	1,421	811	1,587	837	11	10	4,237	3,258
68	3,785	836	63	789	360	875	368	11	9	2,129	757
69	5,399	635	72	1,223	485	1,394	497	11	8	3,594	1,098
70	3,016	566	32	907	473	1,000	488	12	11	1,689	729
71	1,544	168	24	473	147	530	150	8	8	1,037	315

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
72	3,476	479	39	842	373	957	384	13	12	2,063	895
73	5,726	1,173	135	1,267	671	1,452	695	12	11	3,014	1,404
74	2,349	277	75	600	274	671	280	10	10	1,427	570
75	3,166	603	94	808	468	888	481	10	10	1,443	1,026
76	11,325	1,553	242	2,075	1,114	2,417	1,161	12	11	6,784	2,746
77	3,655	791	77	928	399	1,048	414	11	10	2,077	710
78	1,273	153	43	431	185	479	194	10	9	750	327
79	9,372	2,330	400	1,517	865	1,705	895	11	11	4,238	2,404
80	1,553	145	13	588	229	641	234	10	10	1,013	382
81	3,161	343	85	929	430	1,027	445	11	9	1,951	782
82	2,788	457	44	811	369	894	378	10	10	1,615	672
83	14,648	1,913	192	2,469	1,422	2,821	1,461	13	11	8,697	3,846
85	5,550	998	220	1,157	564	1,283	577	11	10	2,783	1,549
86	3,383	784	190	849	437	921	441	12	10	1,581	828
87	3,682	371	44	956	304	1,111	318	11	8	2,666	601
88	11,070	1,962	179	1,939	1,039	2,209	1,063	12	10	6,108	2,821
89	4,166	823	129	898	390	1,006	399	11	10	2,365	849
90	7,471	896	129	1,457	554	1,726	576	12	10	5,130	1,316
91	2,889	535	38	826	406	912	417	12	12	1,622	694
92	5,530	1,138	115	1,253	677	1,374	697	10	10	2,851	1,426
93	1,978	210	50	446	135	506	136	9	8	1,393	325
94	2,038	331	49	656	275	724	283	11	10	1,187	471
95	4,601	1,189	100	855	360	982	373	13	12	2,531	781
96	9,034	1,212	233	1,624	772	1,867	803	12	10	5,458	2,131
97	2,148	246	65	656	280	718	290	11	10	1,376	461

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
98	2,044	148	73	749	358	819	367	10	9	1,271	552
99	9,412	1,505	278	1,927	996	2,217	1,039	10	10	5,315	2,314
100	1,579	261	71	471	234	520	241	11	10	823	424
101	7,645	1,493	175	1,620	970	1,782	995	12	11	3,578	2,399
102	3,230	370	13	796	230	928	245	13	11	2,420	427
103	4,631	723	88	1,014	431	1,145	441	10	10	2,860	960
104	3,732	388	99	887	523	978	532	10	8	1,997	1,248
105	4,522	1,154	274	842	435	923	448	12	11	1,905	1,189
106	3,548	494	55	980	478	1,078	486	11	11	2,116	883
107	7,402	1,477	380	1,393	720	1,588	745	11	10	4,110	1,435
108	1,423	229	21	458	178	506	185	9	7	820	353
109	3,369	411	64	783	369	888	386	12	11	2,015	879
110	10,053	1,600	257	1,635	863	1,883	896	13	11	5,913	2,283
111	1,924	187	62	632	287	691	294	10	9	1,198	477
112	17,694	3,892	864	2,560	1,654	2,877	1,715	13	11	7,673	5,265
113	5,743	940	134	1,194	528	1,344	540	12	11	3,385	1,284
114	7,902	1,622	155	1,476	875	1,656	907	13	12	3,735	2,390
115	11,883	1,815	267	2,084	1,146	2,383	1,180	13	11	6,638	3,163
116	4,596	948	84	826	430	894	432	10	9	2,247	1,317
117	804	90	15	336	104	362	108	8	6	528	171
118	9,910	1,903	337	1,874	1,165	2,134	1,206	12	11	5,119	2,551
119	1,505	140	5	529	188	576	194	10	10	1,009	351
120	743	201	23	279	146	301	153	8	7	313	206
121	2,419	336	265	585	296	652	311	9	8	1,209	609
122	7,528	712	59	1,239	362	1,497	383	12	10	5,713	1,044

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
123	3,423	497	88	1,016	502	1,117	510	11	10	2,020	818
124	1,954	1,131	49	344	293	345	292	8	7	188	586
125	8,038	1,011	89	1,472	788	1,685	820	12	11	4,714	2,224
126	14,378	2,373	479	2,067	1,158	2,377	1,197	13	13	8,099	3,427
127	4,901	678	296	1,176	538	1,324	552	12	10	2,776	1,151
128	13,587	3,046	897	2,511	1,838	2,711	1,867	12	11	4,835	4,809
129	3,404	723	153	963	583	1,056	596	12	11	1,424	1,104
130	6,255	585	160	1,151	439	1,324	463	11	9	4,355	1,155
131	1,474	250	60	531	265	570	271	9	8	736	428
132	1,052	114	7	358	167	393	170	9	9	686	245
133	4,173	927	66	1,120	649	1,217	671	12	12	1,890	1,290
134	1,530	197	41	503	242	554	250	10	10	886	406
135	2,732	1,013	57	643	438	660	439	8	8	763	899
137	6,277	836	258	1,252	661	1,391	672	12	11	3,578	1,605
138	2,414	329	58	571	190	634	192	10	9	1,669	358
139	5,090	803	144	937	378	1,082	399	11	9	3,026	1,117
140	410	43	2	219	81	230	81	6	5	266	99
141	6,289	693	88	1,220	526	1,430	561	13	12	4,091	1,417
143	3,377	304	263	836	258	923	264	11	10	2,295	515
144	11,622	3,605	727	4,075	3,619	4,183	3,629	12	11	2,367	4,923
145	1,636	267	41	611	335	651	336	9	8	849	479
146	4,217	839	155	1,614	1,108	1,689	1,117	11	8	1,661	1,562
147	2,250	298	20	658	199	744	206	11	8	1,539	393
148	1,712	192	14	536	209	588	216	9	7	1,115	391
149	2,957	457	55	770	301	879	316	12	10	1,921	524

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
150	27,122	3,435	1,969	2,809	1,714	3,331	1,802	13	13	16,204	5,514
151	22,225	5,085	451	3,327	2,241	3,761	2,296	12	11	11,482	5,207
152	1,347	141	60	411	197	456	208	9	8	663	483
153	4,100	593	119	985	481	1,089	486	11	11	2,372	1,016
154	4,227	599	117	979	467	1,088	474	11	11	2,475	1,036
155	13,570	3,034	897	2,498	1,825	2,698	1,854	12	11	4,835	4,804
156	1,852	181	53	631	261	700	266	10	10	1,143	475
157	2,783	412	78	698	311	782	316	11	10	1,683	610
158	2,835	425	72	674	299	762	305	11	9	1,714	624
159	6,463	1,125	310	1,354	782	1,508	799	12	10	3,367	1,661
160	8,085	2,080	225	1,723	1,115	1,871	1,136	13	12	3,122	2,658
161	773	158	32	371	166	394	168	8	8	416	167
163	3,686	756	93	862	489	943	507	11	10	1,614	1,223
164	2,099	258	24	671	254	749	260	12	11	1,384	433
165	3,842	795	102	944	367	1,072	375	12	10	2,300	645
167	2,508	222	84	623	258	691	261	10	8	1,506	696
168	13,800	2,062	418	1,979	1,121	2,282	1,158	12	11	7,975	3,345
169	15,391	2,627	482	2,179	1,240	2,519	1,282	13	12	8,693	3,589
170	1,117	149	12	421	140	446	141	9	7	692	264
171	8,495	1,351	195	1,470	699	1,728	729	10	9	5,196	1,753
172	393	32	15	220	76	231	77	10	8	268	78
173	11,846	2,014	1,588	2,151	1,470	2,325	1,504	13	11	4,539	3,705
174	3,050	445	57	762	257	867	267	10	8	1,967	581
175	1,434	119	85	619	263	659	268	9	9	931	299
176	2,197	203	33	662	224	725	222	11	10	1,642	319

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
177	720	118	10	356	153	377	154	11	9	408	184
179	1,453	137	19	517	192	575	200	10	9	1,006	291
180	1,451	145	33	544	239	610	249	9	8	902	371
181	1,404	140	28	522	229	584	236	9	8	878	358
182	4,574	820	63	1,151	585	1,281	598	10	8	2,661	1,030
183	2,717	713	62	675	305	753	315	11	9	1,301	641
184	5,592	896	137	1,195	651	1,306	658	12	10	3,048	1,511
185	2,303	221	36	590	172	665	181	10	9	1,647	399
188	1,858	176	42	608	199	679	206	10	8	1,286	354
189	2,083	314	53	663	280	724	285	11	10	1,241	475
190	4,797	576	84	861	391	1,006	415	12	10	2,927	1,210
191	5,301	772	140	1,062	478	1,202	495	11	10	3,195	1,194
192	1,373	235	27	502	259	527	258	9	8	649	462
194	2,697	316	132	811	403	874	406	11	10	1,575	674
195	2,319	292	70	735	293	808	303	10	9	1,435	522
196	2,894	257	85	718	256	809	264	12	10	1,935	617
197	2,556	270	15	795	280	879	288	12	9	1,692	579
198	2,714	293	100	789	378	850	383	11	10	1,537	784
199	2,735	303	97	769	361	831	366	11	10	1,541	794
200	1,582	153	44	572	209	621	215	9	7	981	404
201	8,790	1,093	208	1,588	829	1,857	867	11	10	5,415	2,074
202	5,664	1,203	145	1,075	584	1,229	612	12	11	2,915	1,401
203	2,885	261	45	862	294	984	306	9	8	2,102	477
204	1,972	169	149	792	328	848	331	10	8	1,141	513
205	10,139	1,991	478	1,968	1,253	2,179	1,290	12	11	4,938	2,732

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
206	2,025	271	77	650	316	722	322	11	9	1,149	528
207	3,300	479	71	717	320	793	322	11	10	2,008	742
208	1,736	201	47	590	230	651	232	11	10	1,152	336
209	3,960	458	62	1,182	627	1,331	648	12	10	2,263	1,177
210	2,650	305	57	632	243	720	260	7	6	1,722	566
211	2,728	540	85	1,231	1,206	1,232	1,207	9	9	36	2,067
212	10,413	705	6,124	981	555	1,067	565	12	11	1,775	1,809
213	2,637	288	45	835	373	941	384	12	11	1,666	638
214	483	79	6	331	268	333	269	8	8	86	312
215	14,831	2,940	865	2,539	1,425	2,817	1,447	12	11	6,992	4,034
216	9,598	2,197	392	1,686	914	1,964	949	11	10	4,867	2,142
217	1,686	247	84	607	340	645	347	12	11	858	497
218	4,646	726	481	1,040	548	1,140	557	12	11	2,149	1,290
219	2,788	861	308	785	507	825	513	10	9	782	837
220	3,742	829	91	946	469	1,070	487	12	10	1,858	964
221	953	100	13	395	164	429	167	9	8	595	245
222	17,639	2,976	471	2,212	1,368	2,555	1,426	13	13	9,735	4,457
223	23,045	4,549	615	2,758	1,574	3,182	1,635	11	10	12,203	5,678
224	36,573	5,411	1,179	3,291	2,175	3,824	2,271	13	13	21,541	8,442
225	27,007	4,316	683	2,444	1,462	2,871	1,531	12	12	15,837	6,171
226	6,250	1,137	133	1,099	535	1,273	556	12	12	3,500	1,480
227	15,453	4,717	659	2,886	1,994	3,153	2,047	12	12	5,865	4,212
228	12,982	1,805	167	1,997	814	2,362	850	12	11	8,913	2,097
229	5,315	1,083	120	1,064	520	1,199	532	10	9	2,664	1,448
230	6,892	1,174	451	1,146	630	1,263	642	11	10	3,333	1,934

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
231	8,589	1,760	318	1,046	504	1,225	528	11	10	5,023	1,488
232	5,071	1,006	104	1,009	456	1,132	467	10	9	2,837	1,124
233	7,144	1,319	172	1,356	705	1,524	718	11	10	3,645	2,008
234	5,985	942	155	1,129	524	1,309	556	13	11	3,713	1,175
235	2,643	375	42	797	396	887	406	11	10	1,571	655
236	2,895	373	69	817	415	904	429	10	7	1,599	854
237	7,963	1,206	155	1,428	727	1,617	744	10	9	4,604	1,998
238	1,036	124	16	374	128	409	132	10	9	695	201
239	3,571	354	224	719	262	827	269	10	8	2,355	638
240	6,861	822	296	1,556	833	1,745	860	11	10	3,954	1,789
241	3,411	776	73	1,142	764	1,218	774	11	8	1,507	1,055
242	2,118	343	106	679	346	738	353	9	9	1,043	626
243	3,435	424	100	792	344	876	348	11	10	2,140	771
244	3,115	343	128	730	263	832	272	11	9	2,069	575
245	3,987	619	172	1,157	633	1,265	649	13	12	2,120	1,076
246	12,127	1,252	321	2,148	879	2,481	904	13	11	8,781	1,773
247	4,638	981	71	946	445	1,066	459	11	10	2,562	1,024
248	8,208	1,858	329	1,088	540	1,252	565	11	8	4,524	1,497
249	4,689	564	85	931	333	1,086	353	9	8	3,265	775
250	9,667	1,866	272	1,764	1,167	1,946	1,194	11	10	4,292	3,237
251	5,448	1,104	155	1,067	530	1,215	550	12	11	2,954	1,235
252	6,597	778	96	1,330	588	1,505	611	11	10	4,215	1,508
253	20,407	5,392	2,732	3,532	2,780	3,759	2,815	13	13	6,008	6,275
254	1,135	297	12	310	157	331	159	6	6	519	307
255	4,651	898	182	1,224	696	1,324	700	11	10	2,310	1,261

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
256	1,502	102	4	423	92	484	97	10	9	1,204	192
257	25,972	3,347	489	3,284	1,807	3,804	1,861	12	12	17,321	4,815
258	4,219	860	158	1,113	637	1,220	648	11	10	2,010	1,191
259	3,744	715	111	774	410	872	422	9	7	1,760	1,158
260	12,315	1,205	218	1,815	760	2,187	798	12	11	8,244	2,648
261	1,978	458	67	489	239	523	241	10	9	947	506
262	10,306	1,924	409	1,507	837	1,714	869	12	10	5,267	2,706
263	16,562	3,633	549	2,498	1,692	2,804	1,748	12	11	7,554	4,826
264	1,948	223	16	606	192	684	203	10	8	1,340	369
265	4,254	875	97	856	376	958	386	11	9	2,449	833
266	1,155	313	12	368	186	383	186	10	10	454	376
267	2,134	659	30	630	391	676	398	12	12	804	641
268	1,196	157	33	450	217	490	221	8	8	688	318
269	13,116	2,181	655	2,156	1,192	2,442	1,240	13	13	6,849	3,431
270	1,572	222	49	506	249	543	250	9	8	867	434
271	1,002	111	10	372	108	414	116	10	8	688	193
272	5,965	932	172	1,121	523	1,282	541	10	9	3,570	1,291
273	1,385	125	8	473	137	535	145	8	7	1,013	239
274	6,531	1,320	187	1,310	660	1,491	692	12	10	3,436	1,588
275	2,407	420	21	757	328	841	341	10	8	1,403	563
276	2,262	470	48	603	250	675	259	11	10	1,314	430
277	2,753	365	54	715	331	787	338	13	11	1,621	713
278	462	46	8	213	67	226	68	6	4	321	87
279	6,073	1,160	302	1,117	530	1,280	547	11	10	3,297	1,314
280	1,768	203	8	519	165	575	173	10	9	1,188	369

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
281	2,842	333	69	772	367	861	377	12	9	1,726	714
282	14,398	2,140	989	2,572	1,582	2,872	1,641	11	10	7,954	3,315
283	7,025	1,032	179	997	532	1,128	549	11	10	4,132	1,682
284	1,043	169	56	386	177	412	176	9	9	553	265
285	1,854	326	19	611	260	657	270	10	8	1,034	475
286	21,340	3,527	615	3,326	2,006	3,772	2,074	13	11	12,048	5,150
287	1,456	437	24	315	156	339	155	11	10	559	436
288	2,751	472	306	853	468	900	472	11	10	1,189	784
289	7,219	1,320	268	1,361	666	1,574	693	12	11	4,072	1,559
290	5,215	583	208	1,006	427	1,149	443	9	8	3,258	1,166
291	1,102	166	12	384	128	431	129	10	8	739	185
292	19,246	5,188	422	3,578	3,058	3,678	3,084	13	12	3,447	10,189
293	5,560	1,296	218	1,146	589	1,274	606	10	10	2,690	1,356
294	1,353	213	54	424	190	466	193	9	8	798	288
295	1,999	290	36	548	236	609	242	10	9	1,163	510
296	17,114	6,239	475	1,349	976	1,413	988	12	11	3,360	7,040
297	6,193	966	196	1,315	729	1,495	760	12	11	3,442	1,589
298	1,687	300	33	504	261	558	273	9	8	864	490
299	1,087	205	6	322	118	348	123	7	6	618	258
300	2,569	647	29	648	323	719	329	9	6	1,286	607
301	3,967	720	90	787	401	862	410	10	9	2,011	1,146
302	2,597	898	22	694	387	750	395	10	9	938	739
303	1,191	193	8	437	173	479	175	9	7	740	250
304	6,735	2,178	173	1,171	697	1,297	719	11	10	2,731	1,653
305	3,700	595	59	967	475	1,075	489	11	10	2,215	831

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
306	986	122	24	360	135	403	137	8	8	683	157
307	1,449	289	59	465	222	512	228	9	8	734	367
308	1,753	330	18	502	226	548	233	9	8	944	461
309	2,719	292	17	761	288	876	298	10	9	1,839	571
311	14,738	2,422	606	2,290	1,273	2,608	1,309	13	12	8,345	3,365
312	5,703	1,078	132	1,203	666	1,346	691	12	11	2,878	1,615
313	6,906	1,314	153	1,581	856	1,798	893	12	11	3,662	1,777
314	7,185	938	218	1,121	528	1,318	560	11	10	4,311	1,718
315	1,035	144	9	332	118	377	127	8	8	679	203
316	1,180	103	21	448	159	496	163	10	8	811	245
317	3,484	710	123	1,063	586	1,135	595	10	8	1,720	931
318	1,222	163	58	436	198	483	208	8	8	729	272
320	11,947	2,291	548	1,862	1,112	2,077	1,140	12	12	5,798	3,310
321	14,074	2,530	979	2,444	1,603	2,758	1,653	12	12	6,712	3,853
322	14,041	2,509	191	1,836	1,041	2,090	1,073	11	9	7,496	3,845
323	13,300	4,883	912	2,200	1,721	2,305	1,730	12	12	2,726	4,779
324	2,187	216	14	602	244	698	261	11	11	1,502	455
325	2,642	448	103	688	393	742	399	10	9	1,269	822
326	7,905	1,245	148	1,339	520	1,578	544	10	10	5,178	1,334
327	3,912	583	173	913	476	984	487	11	10	2,071	1,085
328	1,290	282	24	397	182	437	184	9	8	732	252
329	1,520	180	16	616	313	658	316	11	11	874	450
330	5,268	1,534	209	901	528	991	544	12	12	1,926	1,599
331	4,684	750	42	853	414	981	433	11	11	2,658	1,234
332	1,837	352	20	466	206	506	206	11	9	1,026	439

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
333	1,110	131	50	461	254	489	256	10	9	611	318
334	4,883	942	105	1,197	715	1,323	741	12	11	2,443	1,393
335	509	297	14	183	154	184	155	2	2	38	160
336	1,042	136	9	372	129	417	135	7	6	679	218
337	3,369	442	142	781	396	876	408	11	10	1,820	965
338	1,849	243	45	597	278	673	287	9	9	1,096	465
339	1,522	258	7	481	191	550	196	8	7	872	385
340	918	111	3	375	143	403	145	8	6	613	191
341	6,050	886	306	1,019	446	1,185	464	11	10	3,484	1,374
342	2,765	299	28	695	264	787	273	10	8	1,849	589
343	715	114	2	257	93	282	95	7	5	453	146
344	5,137	612	72	1,121	557	1,288	584	11	10	3,339	1,114
345	2,321	477	74	670	361	727	369	11	10	1,119	651
346	10,318	1,297	635	2,390	1,462	2,658	1,498	13	12	6,106	2,280
347	1,087	147	4	356	115	397	120	9	8	724	212
348	1,232	102	20	427	157	477	164	9	8	841	269
349	23,753	3,497	880	4,086	2,532	4,616	2,595	13	12	13,825	5,551
350	6,306	782	130	1,438	709	1,651	740	12	11	3,880	1,514
351	5,845	2,096	34	979	551	1,088	560	10	9	2,275	1,440
352	3,946	885	84	756	297	858	306	11	9	2,351	626
353	587	110	25	271	191	281	191	10	9	162	290
354	1,589	189	24	439	148	492	156	11	9	1,055	321
355	5,857	1,063	52	1,269	586	1,450	600	11	10	3,434	1,308
356	84,473	39,396	253	1,801	1,102	2,023	1,127	13	13	4,890	39,934
357	7,663	1,979	810	1,320	799	1,450	821	11	9	2,655	2,219

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
358	4,591	1,324	77	863	325	987	339	10	9	2,414	776
359	3,891	575	40	983	436	1,125	453	12	11	2,410	866
360	6,634	1,486	185	1,269	675	1,423	690	11	10	3,475	1,488
361	21,855	4,804	1,198	3,882	2,938	4,183	2,984	13	13	7,942	7,911
362	7,761	1,879	385	1,471	820	1,639	847	12	12	3,703	1,794
363	10,581	1,203	175	1,829	724	2,170	756	13	13	7,724	1,479
364	1,121	162	28	351	144	382	146	10	8	614	317
365	10,981	2,015	266	1,792	877	2,054	904	13	13	6,446	2,254
366	836	168	13	310	128	339	131	8	8	465	190
367	3,835	891	123	1,136	629	1,217	640	10	8	1,822	999
368	7,112	781	306	1,720	994	1,893	1,005	13	11	4,023	2,002
369	1,748	179	27	478	137	551	146	10	8	1,248	294
370	1,616	190	13	561	183	615	185	9	8	1,125	288
371	1,727	111	92	505	213	549	219	8	8	881	643
372	3,612	511	24	1,081	576	1,214	591	11	9	2,052	1,025
373	12,394	2,555	444	2,667	1,676	2,991	1,724	9	9	6,061	3,334
374	15,194	2,142	1,284	2,644	1,648	2,944	1,708	11	10	8,106	3,662
375	3,355	684	116	1,010	536	1,085	543	10	8	1,709	846
376	6,076	798	276	1,227	685	1,362	704	13	13	3,206	1,796
377	3,024	738	192	678	330	748	339	10	8	1,321	773
378	1,315	128	28	391	144	436	148	10	8	875	284
379	9,044	1,225	420	1,787	822	2,047	841	13	11	5,370	2,029
380	2,495	215	13	625	195	710	207	8	8	1,831	436
381	16,367	1,780	417	2,277	1,066	2,652	1,098	12	11	10,814	3,356
382	7,948	858	161	1,263	508	1,467	530	11	10	5,513	1,416

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
383	10,672	1,846	208	1,752	1,046	1,985	1,081	13	11	5,381	3,237
384	7,704	2,319	244	1,363	838	1,534	872	11	9	3,414	1,727
385	2,027	357	36	600	306	653	313	11	10	1,053	581
386	1,813	388	56	524	251	578	261	11	10	925	444
387	13,260	3,091	345	1,988	1,442	2,126	1,462	11	10	4,617	5,207
388	7,181	1,617	157	1,116	464	1,295	478	11	10	4,085	1,322
389	4,908	1,126	136	1,063	517	1,192	537	11	10	2,606	1,040
390	1,760	223	50	493	213	545	220	9	8	1,047	440
392	21,055	2,224	455	3,233	1,715	3,790	1,787	13	12	14,588	3,788
393	2,069	304	47	753	408	819	424	10	9	1,108	610
394	3,862	460	91	980	430	1,084	433	11	10	2,373	938
395	4,836	931	133	1,013	515	1,114	529	12	11	2,539	1,233
396	4,104	966	50	1,195	685	1,294	696	11	10	1,805	1,283
397	1,288	202	57	504	257	550	272	10	10	625	404
398	1,026	176	24	384	188	413	190	11	10	560	266
399	1,203	190	11	410	163	453	172	8	8	730	272
400	14,007	5,594	451	1,906	1,355	2,053	1,380	12	12	3,979	3,983
401	7,545	1,342	139	1,375	622	1,563	644	11	10	4,621	1,443
402	6,609	1,035	387	1,212	606	1,393	635	11	10	3,586	1,601
403	7,144	664	723	1,797	1,440	1,867	1,452	10	10	1,609	4,148
404	7,855	1,325	153	1,700	931	1,903	963	13	12	4,587	1,790
405	11,064	5,033	243	1,696	1,269	1,782	1,284	12	11	2,597	3,191
406	8,804	1,440	260	1,530	808	1,758	846	12	10	4,758	2,346
407	2,883	641	205	905	501	990	512	12	11	1,388	649
408	5,307	1,561	138	1,292	948	1,357	963	9	9	1,357	2,251

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
409	5,214	1,462	128	1,002	513	1,099	526	12	11	2,471	1,153
411	12,927	2,863	287	2,116	1,371	2,389	1,417	12	11	6,218	3,559
412	3,751	979	435	710	453	757	457	10	10	1,076	1,261
413	11,283	1,934	304	1,921	1,181	2,165	1,216	12	12	5,975	3,070
414	1,485	212	43	446	196	497	204	11	10	897	333
415	3,494	581	101	902	433	1,004	441	11	11	1,980	832
416	6,227	904	133	1,323	674	1,484	700	12	10	3,390	1,800
417	807	88	2	350	114	382	117	8	7	561	156
419	3,991	657	66	989	510	1,102	522	11	11	2,186	1,082
420	15,581	5,289	174	3,474	3,160	3,514	3,172	12	10	3,055	7,063
421	2,274	323	37	616	240	695	255	10	9	1,344	570
422	3,149	950	193	953	555	1,026	564	13	11	1,289	717
423	8,693	1,035	297	1,317	580	1,550	602	10	10	5,481	1,880
424	6,027	1,154	134	1,058	486	1,215	506	11	9	3,642	1,097
425	1,453	193	102	545	378	575	382	8	8	470	688
426	727	122	52	388	214	397	214	11	9	302	251
427	3,918	584	447	1,145	657	1,246	666	11	10	1,815	1,072
428	12,050	2,530	807	3,720	3,164	3,861	3,190	12	12	3,297	5,416
429	2,020	279	34	634	298	705	306	10	9	1,220	487
430	2,607	235	60	811	354	892	362	10	9	1,477	835
431	3,611	511	14	900	371	1,033	381	11	9	2,338	748
432	11,357	1,613	424	2,446	1,187	2,737	1,211	12	11	6,542	2,778
433	2,268	419	58	638	269	718	283	10	9	1,300	491
434	1,547	171	39	484	208	528	212	10	8	928	409
435	2,054	298	88	964	751	983	750	9	8	758	910

Document Number	Total Tokens	Punctuation Tokens	Digits Tokens	Unique Lemma (Base)	Unique Lemma (Adjusted)	Unique Original Words (Base)	Unique Original Words (Adjusted)	Unique NER (Base)	Unique NER (Adjusted)	Common Word Tokens Removed	Remaining Tokens
436	2,684	285	31	840	324	924	326	10	9	1,788	580
437	3,125	431	59	868	366	965	376	9	8	2,053	582
438	18,718	2,456	382	3,186	2,076	3,623	2,132	13	13	11,109	4,771
439	26,705	3,555	847	3,293	1,853	3,930	1,930	13	13	16,091	6,212
440	8,893	2,325	856	1,329	881	1,431	893	10	10	2,428	3,284
441	1,393	221	25	465	215	507	223	10	9	771	376

APPENDIX B. FIVE HUNDRED (500) MOST COMMON TOKENS

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
cid	cid	28	20055	716.2500	47.1882	6.5882	1.0624
malware	malware	404	8487	21.0074	19.9694	95.0588	0.4496
s	’s	338	6129	18.1331	14.4212	79.5294	0.3247
datum	data	375	4949	13.1973	11.6447	88.2353	0.2622
server	server	341	4128	12.1056	9.7129	80.2353	0.2187
c	C	272	3347	12.3051	7.8753	64.0000	0.1773
windows	Windows	321	2846	8.8660	6.6965	75.5294	0.1508
s	S	227	2549	11.2291	5.9976	53.4118	0.1350
attacker	attackers	295	2539	8.6068	5.9741	69.4118	0.1345
com	com	214	2512	11.7383	5.9106	50.3529	0.1331
ip	IP	292	2440	8.3562	5.7412	68.7059	0.1293
c2	C2	169	2378	14.0710	5.5953	39.7647	0.1260
malicious	malicious	331	2324	7.0211	5.4682	77.8824	0.1231
c&c	C&C	171	2306	13.4854	5.4259	40.2353	0.1222
microsoft	Microsoft	302	1921	6.3609	4.5200	71.0588	0.1018
cyber	CYBER	209	1865	8.9234	4.3882	49.1765	0.0988
e	E	155	1857	11.9806	4.3694	36.4706	0.0984
server	servers	253	1821	7.1976	4.2847	59.5294	0.0965
email	email	256	1705	6.6602	4.0118	60.2353	0.0903
dll	DLL	206	1649	8.0049	3.8800	48.4706	0.0874
md5	MD5	242	1608	6.6446	3.7835	56.9412	0.0852
china	China	157	1544	9.8344	3.6329	36.9412	0.0818
d	D	180	1496	8.3111	3.5200	42.3529	0.0792
payload	payload	234	1473	6.2949	3.4659	55.0588	0.0780
configuration	configuration	182	1380	7.5824	3.2471	42.8235	0.0731

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
backdoor	backdoor	186	1370	7.3656	3.2235	43.7647	0.0726
byte	bytes	188	1292	6.8723	3.0400	44.2353	0.0684
module	module	126	1275	10.1190	3.0000	29.6471	0.0675
http	HTTP	218	1212	5.5596	2.8518	51.2941	0.0642
o	O	93	1207	12.9785	2.8400	21.8824	0.0639
registry	registry	197	1206	6.1218	2.8376	46.3529	0.0639
infected	infected	231	1175	5.0866	2.7647	54.3529	0.0622
n	N	117	1174	10.0342	2.7624	27.5294	0.0622
b	B	153	1089	7.1176	2.5624	36.0000	0.0577
apt	APT	158	1083	6.8544	2.5482	37.1765	0.0574
attacker	attacker	229	1067	4.6594	2.5106	53.8824	0.0565
r	R	116	1024	8.8276	2.4094	27.2941	0.0542
trojan	Trojan	188	1002	5.3298	2.3576	44.2353	0.0531
binary	binary	195	982	5.0359	2.3106	45.8824	0.0520
executable	executable	251	979	3.9004	2.3035	59.0588	0.0519
t	T	109	976	8.9541	2.2965	25.6471	0.0517
directory	directory	191	974	5.0995	2.2918	44.9412	0.0516
id	ID	172	973	5.6570	2.2894	40.4706	0.0515
appendix	Appendix	140	856	6.1143	2.0141	32.9412	0.0453
website	website	216	846	3.9167	1.9906	50.8235	0.0448
micro	Micro	69	825	11.9565	1.9412	16.2353	0.0437
vulnerability	vulnerability	185	810	4.3784	1.9059	43.5294	0.0429
org	org	95	809	8.5158	1.9035	22.3529	0.0429
detection	detection	247	804	3.2551	1.8918	58.1176	0.0426
functionality	functionality	206	795	3.8592	1.8706	48.4706	0.0421
variant	variant	174	795	4.5690	1.8706	40.9412	0.0421
dropper	dropper	135	779	5.7704	1.8329	31.7647	0.0413
encrypted	encrypted	194	777	4.0052	1.8282	45.6471	0.0412
variant	variants	176	768	4.3636	1.8071	41.4118	0.0407
password	password	168	767	4.5655	1.8047	39.5294	0.0406

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
f	F	111	759	6.8378	1.7859	26.1176	0.0402
exe	EXE	180	748	4.1556	1.7600	42.3529	0.0396
download	download	226	747	3.3053	1.7576	53.1765	0.0396
further	further	276	729	2.6413	1.7153	64.9412	0.0386
url	URL	187	719	3.8449	1.6918	44.0000	0.0381
espionage	ESPIONAGE	155	714	4.6065	1.6800	36.4706	0.0378
temp	temp	147	714	4.8571	1.6800	34.5882	0.0378
korea	Korea	63	707	11.2222	1.6635	14.8235	0.0375
email	emails	154	696	4.5195	1.6376	36.2353	0.0369
decoy	Decoy	98	694	7.0816	1.6329	23.0588	0.0368
filename	filename	142	686	4.8310	1.6141	33.4118	0.0363
hash	hash	162	680	4.1975	1.6000	38.1176	0.0360
l	L	75	673	8.9733	1.5835	17.6471	0.0357
dword	dword	60	661	11.0167	1.5553	14.1176	0.0350
phishing	Phishing	127	660	5.1969	1.5529	29.8824	0.0350
stuxnet	stuxnet	30	658	21.9333	1.5482	7.0588	0.0349
tlp	TLP	31	657	21.1935	1.5459	7.2941	0.0348
p	P	88	643	7.3068	1.5129	20.7059	0.0341
july	July	152	636	4.1842	1.4965	35.7647	0.0337
kaspersky	Kaspersky	112	635	5.6696	1.4941	26.3529	0.0336
symantec	Symantec	88	635	7.2159	1.4941	20.7059	0.0336
compile	compiled	125	618	4.9440	1.4541	29.4118	0.0327
encryption	encryption	168	618	3.6786	1.4541	39.5294	0.0327
blog	blog	201	603	3.0000	1.4188	47.2941	0.0319
m	M	85	591	6.9529	1.3906	20.0000	0.0313
fireeye	FireEye	79	582	7.3671	1.3694	18.5882	0.0308
copyright	Copyright	113	580	5.1327	1.3647	26.5882	0.0307
rights	RIGHTS	183	571	3.1202	1.3435	43.0588	0.0302
google	Google	152	563	3.7039	1.3247	35.7647	0.0298
xor	XOR	132	559	4.2348	1.3153	31.0588	0.0296

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
x	X	127	558	4.3937	1.3129	29.8824	0.0296
pdf	PDF	141	549	3.8936	1.2918	33.1765	0.0291
proxy	proxy	132	549	4.1591	1.2918	31.0588	0.0291
february	February	129	543	4.2093	1.2776	30.3529	0.0288
u	U	94	542	5.7660	1.2753	22.1176	0.0287
website	websites	127	536	4.2205	1.2612	29.8824	0.0284
apt1	APT1	13	519	39.9231	1.2212	3.0588	0.0275
registrant	registrant	67	517	7.7164	1.2165	15.7647	0.0274
intelreports@kaspersky.com	intelreports@kaspersky.com	22	508	23.0909	1.1953	5.1765	0.0269
module	modules	100	508	5.0800	1.1953	23.5294	0.0269
g	G	85	504	5.9294	1.1859	20.0000	0.0267
april	April	138	499	3.6159	1.1741	32.4706	0.0264
h	h	93	496	5.3333	1.1671	21.8824	0.0263
november	November	153	496	3.2418	1.1671	36.0000	0.0263
plugx	PlugX	48	495	10.3125	1.1647	11.2941	0.0262
cnc	CnC	44	491	11.1591	1.1553	10.3529	0.0260
russia	Russia	87	491	5.6437	1.1553	20.4706	0.0260
folder	folder	147	486	3.3061	1.1435	34.5882	0.0257
dn	DNS	132	485	3.6742	1.1412	31.0588	0.0257
contents	contents	198	477	2.4091	1.1224	46.5882	0.0253
system32	System32	112	469	4.1875	1.1035	26.3529	0.0248
byte	byte	104	466	4.4808	1.0965	24.4706	0.0247
offset	offset	88	466	5.2955	1.0965	20.7059	0.0247
utc	UTC	59	466	7.8983	1.0965	13.8824	0.0247
india	India	80	463	5.7875	1.0894	18.8235	0.0245
june	June	157	458	2.9172	1.0776	36.9412	0.0243
explorer	Explorer	147	456	3.1020	1.0729	34.5882	0.0242
august	August	151	449	2.9735	1.0565	35.5294	0.0238
fake	fake	116	449	3.8707	1.0565	27.2941	0.0238
credentials	credentials	148	448	3.0270	1.0541	34.8235	0.0237

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
algorithm	algorithm	124	446	3.5968	1.0494	29.1765	0.0236
crowdstrike	CrowdStrike	32	445	13.9063	1.0471	7.5294	0.0236
eax	eax	26	444	17.0769	1.0447	6.1176	0.0235
january	January	142	442	3.1127	1.0400	33.4118	0.0234
loader	Loader	77	435	5.6494	1.0235	18.1176	0.0230
tcp	TCP	115	434	3.7739	1.0212	27.0588	0.0230
certificate	certificate	80	425	5.3125	1.0000	18.8235	0.0225
mov	mov	26	416	16.0000	0.9788	6.1176	0.0220
adversary	adversary	70	415	5.9286	0.9765	16.4706	0.0220
header	header	117	415	3.5470	0.9765	27.5294	0.0220
y	Y	70	415	5.9286	0.9765	16.4706	0.0220
tiger	TIGeR	13	413	31.7692	0.9718	3.0588	0.0219
sha256	SHA256	49	412	8.4082	0.9694	11.5294	0.0218
pe	PE	115	410	3.5652	0.9647	27.0588	0.0217
rsa	rsa	54	407	7.5370	0.9576	12.7059	0.0216
nt	NT	107	405	3.7850	0.9529	25.1765	0.0215
java	java	70	404	5.7714	0.9506	16.4706	0.0214
php	php	75	404	5.3867	0.9506	17.6471	0.0214
october	October	149	403	2.7047	0.9482	35.0588	0.0213
u.s.	U.S.	98	403	4.1122	0.9482	23.0588	0.0213
duqu	Duqu	19	401	21.1053	0.9435	4.4706	0.0212
info	Info	128	401	3.1328	0.9435	30.1176	0.0212
zero-day	Zero-Day	82	399	4.8659	0.9388	19.2941	0.0211
compile	COMPILE	79	397	5.0253	0.9341	18.5882	0.0210
located	located	171	396	2.3158	0.9318	40.2353	0.0210
buffer	buffer	74	395	5.3378	0.9294	17.4118	0.0209
ff	FF	35	394	11.2571	0.9271	8.2353	0.0209
vulnerability	vulnerabilities	135	393	2.9111	0.9247	31.7647	0.0208
botnet	botnet	56	387	6.9107	0.9106	13.1765	0.0205
packet	packet	72	387	5.3750	0.9106	16.9412	0.0205

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
states	States	150	387	2.5800	0.9106	35.2941	0.0205
encode	encoded	146	385	2.6370	0.9059	34.3529	0.0204
shellcode	shellcode	82	385	4.6951	0.9059	19.2941	0.0204
september	September	139	384	2.7626	0.9035	32.7059	0.0203
adobe	Adobe	109	375	3.4404	0.8824	25.6471	0.0199
attribution	Attribution	104	373	3.5865	0.8776	24.4706	0.0198
appdata	APPDATA	75	370	4.9333	0.8706	17.6471	0.0196
ukraine	Ukraine	47	369	7.8511	0.8682	11.0588	0.0195
int	int	46	367	7.9783	0.8635	10.8235	0.0194
api	API	98	361	3.6837	0.8494	23.0588	0.0191
plugin	plugin	68	361	5.3088	0.8494	16.0000	0.0191
currentversion	CurrentVersion	114	360	3.1579	0.8471	26.8235	0.0191
mcafee	McAfee	58	357	6.1552	0.8400	13.6471	0.0189
persistence	persistence	135	351	2.6000	0.8259	31.7647	0.0186
installer	installer	75	347	4.6267	0.8165	17.6471	0.0184
spear	spear	100	347	3.4700	0.8165	23.5294	0.0184
iran	Iran	55	345	6.2727	0.8118	12.9412	0.0183
accord	according	147	343	2.3333	0.8071	34.5882	0.0182
default	default	131	343	2.6183	0.8071	30.8235	0.0182
parameter	Parameters	115	341	2.9652	0.8024	27.0588	0.0181
targeting	targeting	116	341	2.9397	0.8024	27.2941	0.0181
december	December	126	340	2.6984	0.8000	29.6471	0.0180
embedded	embedded	135	340	2.5185	0.8000	31.7647	0.0180
download	downloaded	152	339	2.2303	0.7976	35.7647	0.0180
encrypt	encrypted	144	336	2.3333	0.7906	33.8824	0.0178
html	HTML	127	336	2.6457	0.7906	29.8824	0.0178
hacking	Hacking	104	334	3.2115	0.7859	24.4706	0.0177
upload	upload	128	332	2.5938	0.7812	30.1176	0.0176
browser	browser	116	331	2.8534	0.7788	27.2941	0.0175
inc.	Inc	93	329	3.5376	0.7741	21.8824	0.0174

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
panda	Panda	35	328	9.3714	0.7718	8.2353	0.0174
tibetan	Tibetan	36	328	9.1111	0.7718	8.4706	0.0174
delete	delete	124	327	2.6371	0.7694	29.1765	0.0173
clr	CLR	33	326	9.8788	0.7671	7.7647	0.0173
hash	HASHES	128	325	2.5391	0.7647	30.1176	0.0172
attachment	attachment	116	324	2.7931	0.7624	27.2941	0.0172
node	node	32	322	10.0625	0.7576	7.5294	0.0171
os	OS	101	321	3.1782	0.7553	23.7647	0.0170
username	username	94	321	3.4149	0.7553	22.1176	0.0170
antivirus	AntiVirus	122	319	2.6148	0.7506	28.7059	0.0169
bot	bot	53	319	6.0189	0.7506	12.4706	0.0169
downloader	Downloader	62	319	5.1452	0.7506	14.5882	0.0169
v	V	77	315	4.0909	0.7412	18.1176	0.0167
e.g.	e.g.	103	314	3.0485	0.7388	24.2353	0.0166
identifier	identifier	85	313	3.6824	0.7365	20.0000	0.0166
services	Services	117	312	2.6667	0.7341	27.5294	0.0165
persistent	persistent	144	310	2.1528	0.7294	33.8824	0.0164
screenshot	screenshot	106	310	2.9245	0.7294	24.9412	0.0164
inc.	Inc.	62	305	4.9194	0.7176	14.5882	0.0162
blackenergy	BlackEnergy	17	301	17.7059	0.7082	4.0000	0.0159
hong	Hong	64	300	4.6875	0.7059	15.0588	0.0159
hxxp	hxxp	53	300	5.6604	0.7059	12.4706	0.0159
parameter	parameter	94	299	3.1809	0.7035	22.1176	0.0158
http/1	HTTP/1	80	295	3.6875	0.6941	18.8235	0.0156
twitter	Twitter	73	294	4.0274	0.6918	17.1765	0.0156
sav	SAV	5	293	58.6000	0.6894	1.1765	0.0155
kong	Kong	65	292	4.4923	0.6871	15.2941	0.0155
packrat	Packrat	2	292	146.0000	0.6871	0.4706	0.0155
exploitation	exploitation	113	291	2.5752	0.6847	26.5882	0.0154
keylogger	keylogger	77	291	3.7792	0.6847	18.1176	0.0154

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
8b	8b	23	289	12.5652	0.6800	5.4118	0.0153
blockbuster	BLOCKBUSTER	6	288	48.0000	0.6776	1.4118	0.0153
desktop	desktop	97	288	2.9691	0.6776	22.8235	0.0153
backdoor	backdoors	69	287	4.1594	0.6753	16.2353	0.0152
naikon	NAIKON	9	287	31.8889	0.6753	2.1176	0.0152
taiwan	Taiwan	79	287	3.6329	0.6753	18.5882	0.0152
hacker	hackers	72	285	3.9583	0.6706	16.9412	0.0151
gh0st	gh0st	36	284	7.8889	0.6682	8.4706	0.0150
hklm	HKLM	60	284	4.7333	0.6682	14.1176	0.0150
startup	Startup	103	284	2.7573	0.6682	24.2353	0.0150
archive	archive	95	283	2.9789	0.6659	22.3529	0.0150
kernel	kernel	62	280	4.5161	0.6588	14.5882	0.0148
plugin	plugins	55	280	5.0909	0.6588	12.9412	0.0148
apus	API	80	279	3.4875	0.6565	18.8235	0.0148
registration	REGISTRATION	93	276	2.9677	0.6494	21.8824	0.0146
timer	Timer	13	276	21.2308	0.6494	3.0588	0.0146
backspace	BACKSPACE	7	274	39.1429	0.6447	1.6471	0.0145
doc	DOC	90	274	3.0444	0.6447	21.1765	0.0145
earlier	earlier	152	274	1.8026	0.6447	35.7647	0.0145
vector	Vector	121	274	2.2645	0.6447	28.4706	0.0145
x00	x00	25	274	10.9600	0.6447	5.8824	0.0145
latest	latest	144	273	1.8958	0.6424	33.8824	0.0145
base64	Base64	95	272	2.8632	0.6400	22.3529	0.0144
k	K	65	271	4.1692	0.6376	15.2941	0.0144
sha1	SHA1	70	271	3.8714	0.6376	16.4706	0.0144
aurora	AURORA	27	270	10.0000	0.6353	6.3529	0.0143
spear-phishing	spear-phishing	88	268	3.0455	0.6306	20.7059	0.0142
de	de	76	267	3.5132	0.6282	17.8824	0.0141
hostname	hostname	75	267	3.5600	0.6282	17.6471	0.0141
networks	Networks	67	267	3.9851	0.6282	15.7647	0.0141

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
password	passwords	103	267	2.5922	0.6282	24.2353	0.0141
gmt	GMT	42	266	6.3333	0.6259	9.8824	0.0141
6f	6f	23	264	11.4783	0.6212	5.4118	0.0140
regin	Regin	13	264	20.3077	0.6212	3.0588	0.0140
additionally	Additionally	119	263	2.2101	0.6188	28.0000	0.0139
poison	Poison	45	261	5.8000	0.6141	10.5882	0.0138
decrypt	decrypted	118	260	2.2034	0.6118	27.7647	0.0138
lazarus	Lazarus	6	260	43.3333	0.6118	1.4118	0.0138
cc	cc	44	259	5.8864	0.6094	10.3529	0.0137
theft	Theft	93	257	2.7634	0.6047	21.8824	0.0136
binary	binaries	101	256	2.5347	0.6024	23.7647	0.0136
inquire	inquire	5	256	51.2000	0.6024	1.1765	0.0136
user-agent	User-Agent	87	256	2.9425	0.6024	20.4706	0.0136
w	w	74	256	3.4595	0.6024	17.4118	0.0136
z	Z	56	255	4.5536	0.6000	13.1765	0.0135
authentication	Authentication	90	253	2.8111	0.5953	21.1765	0.0134
ivy	Ivy	43	253	5.8837	0.5953	10.1176	0.0134
timeline	timeline	94	253	2.6915	0.5953	22.1176	0.0134
ukrainian	Ukrainian	21	252	12.0000	0.5929	4.9412	0.0133
compilation	compilation	77	251	3.2597	0.5906	18.1176	0.0133
file	file	33	251	7.6061	0.5906	7.7647	0.0133
ip	IPs	98	251	2.5612	0.5906	23.0588	0.0133
miniduke	MiniDuke	19	251	13.2105	0.5906	4.4706	0.0133
carbanak	Carbanak	6	250	41.6667	0.5882	1.4118	0.0132
debug	debug	77	250	3.2468	0.5882	18.1176	0.0132
interface	interface	96	249	2.5938	0.5859	22.5882	0.0132
intrusion	intrusion	93	249	2.6774	0.5859	21.8824	0.0132
attachment	attachments	100	248	2.4800	0.5835	23.5294	0.0131
ca	CA	88	247	2.8068	0.5812	20.7059	0.0131
char	char	40	247	6.1750	0.5812	9.4118	0.0131

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
cybersecurity	cybersecurity	66	247	3.7424	0.5812	15.5294	0.0131
virustotal	VirusTotal	85	246	2.8941	0.5788	20.0000	0.0130
decryption	decryption	87	245	2.8161	0.5765	20.4706	0.0130
decrypt	decrypt	101	243	2.4059	0.5718	23.7647	0.0129
url	URLs	80	243	3.0375	0.5718	18.8235	0.0129
64-bit	64-bit	58	242	4.1724	0.5694	13.6471	0.0128
firewall	Firewall	78	239	3.0641	0.5624	18.3529	0.0127
fidelis	Fidelis	12	238	19.8333	0.5600	2.8235	0.0126
syrian	Syrian	20	237	11.8500	0.5576	4.7059	0.0126
cozyduke	CozyDuke	6	236	39.3333	0.5553	1.4118	0.0125
infect	infect	107	236	2.2056	0.5553	25.1765	0.0125
mbr	MBR	15	236	15.7333	0.5553	3.5294	0.0125
rootkit	rootkit	51	235	4.6078	0.5529	12.0000	0.0124
compatible	compatible	77	234	3.0390	0.5506	18.1176	0.0124
usa	USA	64	234	3.6563	0.5506	15.0588	0.0124
windir	WINDIR	34	234	6.8824	0.5506	8.0000	0.0124
msie	MSIE	69	233	3.3768	0.5482	16.2353	0.0123
ebp	ebp	20	231	11.5500	0.5435	4.7059	0.0122
timestamp	timestamp	79	230	2.9114	0.5412	18.5882	0.0122
retrieve	retrieved	59	228	3.8644	0.5365	13.8824	0.0121
ecx	ecx	22	227	10.3182	0.5341	5.1765	0.0120
gen	Gen	12	227	18.9167	0.5341	2.8235	0.0120
implant	implant	48	227	4.7292	0.5341	11.2941	0.0120
n/a	n/a	38	227	5.9737	0.5341	8.9412	0.0120
vpn	VPN	53	227	4.2830	0.5341	12.4706	0.0120
rc4	RC4	66	226	3.4242	0.5318	15.5294	0.0120
sha-256	SHA-256	18	225	12.5000	0.5294	4.2353	0.0119
configure	configured	97	223	2.2990	0.5247	22.8235	0.0118
asia	Asia	88	222	2.5227	0.5224	20.7059	0.0118
certificate	certificates	54	222	4.1111	0.5224	12.7059	0.0118

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
mutex	mutex	80	222	2.7750	0.5224	18.8235	0.0118
infect	infected	101	220	2.1782	0.5176	23.7647	0.0117
download	downloads	119	219	1.8403	0.5153	28.0000	0.0116
gif	gif	26	219	8.4231	0.5153	6.1176	0.0116
cybercrime	CybERCRImE	61	218	3.5738	0.5129	14.3529	0.0115
apt30	APT30	4	217	54.2500	0.5106	0.9412	0.0115
inject	injected	86	217	2.5233	0.5106	20.2353	0.0115
linux	Linux	47	217	4.6170	0.5106	11.0588	0.0115
overview	Overview	110	217	1.9727	0.5106	25.8824	0.0115
payload	payloads	86	217	2.5233	0.5106	20.2353	0.0115
rar	RAR	67	217	3.2388	0.5106	15.7647	0.0115
intel	InTel	49	216	4.4082	0.5082	11.5294	0.0114
advisory	advisory	31	214	6.9032	0.5035	7.2941	0.0113
6c	6c	30	213	7.1000	0.5012	7.0588	0.0113
login	login	103	213	2.0680	0.5012	24.2353	0.0113
av	AV	73	212	2.9041	0.4988	17.1765	0.0112
injection	injection	67	212	3.1642	0.4988	15.7647	0.0112
var	VAR	22	211	9.5909	0.4965	5.1765	0.0112
fa	fA	27	210	7.7778	0.4941	6.3529	0.0111
japan	Japan	74	209	2.8243	0.4918	17.4118	0.0111
hardcoded	hardcoded	87	208	2.3908	0.4894	20.4706	0.0110
threatconnect	ThreatConnect	12	207	17.2500	0.4871	2.8235	0.0110
controller	controller	36	206	5.7222	0.4847	8.4706	0.0109
reconnaissance	reconnaissance	79	206	2.6076	0.4847	18.5882	0.0109
soft@hotmail.com	soft@hotmail.com	1	206	206.0000	0.4847	0.2353	0.0109
destructive	destructive	33	205	6.2121	0.4824	7.7647	0.0109
trademark	trademarks	86	205	2.3837	0.4824	20.2353	0.0109
xxx	XXX	15	205	13.6667	0.4824	3.5294	0.0109
config	config	57	204	3.5789	0.4800	13.4118	0.0108
breach	breach	69	203	2.9420	0.4776	16.2353	0.0108

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
cyberespionage	cyberespionage	27	203	7.5185	0.4776	6.3529	0.0108
ftp	FTP	55	203	3.6909	0.4776	12.9412	0.0108
ic	ICS	25	202	8.0800	0.4753	5.8824	0.0107
endpoint	endpoint	62	201	3.2419	0.4729	14.5882	0.0106
solutions	Solutions	30	201	6.7000	0.4729	7.0588	0.0106
usb	USB	62	201	3.2419	0.4729	14.5882	0.0106
6e	6e	24	199	8.2917	0.4682	5.6471	0.0105
currentcontrolset	CurrentControlSet	46	199	4.3261	0.4682	10.8235	0.0105
ge	Ge	8	199	24.8750	0.4682	1.8824	0.0105
j	J	49	199	4.0612	0.4682	11.5294	0.0105
usage	usage	99	199	2.0101	0.4682	23.2941	0.0105
beacon	beacon	62	198	3.1935	0.4659	14.5882	0.0105
cmd.exe	cmd.exe	59	198	3.3559	0.4659	13.8824	0.0105
upload	uploaded	95	198	2.0842	0.4659	22.3529	0.0105
xfish	xfish	1	197	197.0000	0.4635	0.2353	0.0104
32-bit	32-bit	64	196	3.0625	0.4612	15.0588	0.0104
older	older	91	193	2.1209	0.4541	21.4118	0.0102
operational	operational	96	193	2.0104	0.4541	22.5882	0.0102
apt28	APT28	10	191	19.1000	0.4494	2.3529	0.0101
hkcu	HKCU	69	191	2.7681	0.4494	16.2353	0.0101
static	static	81	189	2.3333	0.4447	19.0588	0.0100
android	Android	39	188	4.8205	0.4424	9.1765	0.0100
cybercriminal	cybercriminals	48	188	3.9167	0.4424	11.2941	0.0100
svchost.exe	svchost.exe	71	188	2.6479	0.4424	16.7059	0.0100
warfare	warfare	33	188	5.6970	0.4424	7.7647	0.0100
interestingly	Interestingly	108	187	1.7315	0.4400	25.4118	0.0099
mandiant	Mandiant	23	187	8.1304	0.4400	5.4118	0.0099
myanmar	Myanmar	22	187	8.5000	0.4400	5.1765	0.0099
whois	whois	60	187	3.1167	0.4400	14.1176	0.0099
facebook	Facebook	72	186	2.5833	0.4376	16.9412	0.0099

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
ghostnet	GhostNet	17	185	10.8824	0.4353	4.0000	0.0098
anti-virus	anti-virus	64	184	2.8750	0.4329	15.0588	0.0097
georgia	Georgia	22	183	8.3182	0.4306	5.1765	0.0097
callback	callback	48	182	3.7917	0.4282	11.2941	0.0096
monitoring	monitoring	89	182	2.0449	0.4282	20.9412	0.0096
mozilla/4	Mozilla/4	62	182	2.9355	0.4282	14.5882	0.0096
obfuscation	obfuscation	87	182	2.0920	0.4282	20.4706	0.0096
redacted	Redacted	37	182	4.9189	0.4282	8.7059	0.0096
ae	AES	29	181	6.2414	0.4259	6.8235	0.0096
chopper	Chopper	7	179	25.5714	0.4212	1.6471	0.0095
query	query	73	179	2.4521	0.4212	17.1765	0.0095
removable	Removable	44	179	4.0682	0.4212	10.3529	0.0095
vector	vectors	69	179	2.5942	0.4212	16.2353	0.0095
luxembourg	Luxembourg	14	178	12.7143	0.4188	3.2941	0.0094
ssl	SSL	60	178	2.9667	0.4188	14.1176	0.0094
2e	2e	23	177	7.6957	0.4165	5.4118	0.0094
exfiltration	exfiltration	75	177	2.3600	0.4165	17.6471	0.0094
tg-3390	TG-3390	2	177	88.5000	0.4165	0.4706	0.0094
™	™	58	177	3.0517	0.4165	13.6471	0.0094
compress	compressed	75	176	2.3467	0.4141	17.6471	0.0093
txt	txt	51	176	3.4510	0.4141	12.0000	0.0093
circl	CIRCL	6	175	29.1667	0.4118	1.4118	0.0093
cryptography	cryptography	28	175	6.2500	0.4118	6.5882	0.0093
handler	handler	30	175	5.8333	0.4118	7.0588	0.0093
toolset	toolset	43	175	4.0698	0.4118	10.1176	0.0093
turla	Turla	21	175	8.3333	0.4118	4.9412	0.0093
backup	backup	62	174	2.8065	0.4094	14.5882	0.0092
notable	notable	70	173	2.4714	0.4071	16.4706	0.0092
alert	alert	68	172	2.5294	0.4047	16.0000	0.0091
bypass	Bypass	82	172	2.0976	0.4047	19.2941	0.0091

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
c7	C7	20	172	8.6000	0.4047	4.7059	0.0091
earliest	earliest	58	172	2.9655	0.4047	13.6471	0.0091
javascript	Javascript	52	172	3.3077	0.4047	12.2353	0.0091
serial	Serial	75	172	2.2933	0.4047	17.6471	0.0091
incorporated	Incorporated	29	171	5.8966	0.4024	6.8235	0.0091
mz	MZ	31	170	5.4839	0.4000	7.2941	0.0090
ponmocup	Ponmocup	1	170	170.0000	0.4000	0.2353	0.0090
delete	deletes	78	169	2.1667	0.3976	18.3529	0.0090
edx	edx	19	169	8.8947	0.3976	4.4706	0.0090
greenfor	GreenFor	3	169	56.3333	0.3976	0.7059	0.0090
operations	Operations	48	169	3.5208	0.3976	11.2941	0.0090
socket	socket	49	169	3.4490	0.3976	11.5294	0.0090
www.fireeye.com	www.fireeye.com	17	169	9.9412	0.3976	4.0000	0.0090
aspx	aspx	40	168	4.2000	0.3953	9.4118	0.0089
delete	deleted	75	168	2.2400	0.3953	17.6471	0.0089
iranian	Iranian	30	168	5.6000	0.3953	7.0588	0.0089
admin	admin	52	167	3.2115	0.3929	12.2353	0.0088
cn	CN	43	167	3.8837	0.3929	10.1176	0.0088
embed	Embedded	100	167	1.6700	0.3929	23.5294	0.0088
finfisher	FinFisher	8	167	20.8750	0.3929	1.8824	0.0088
pivy	PIVY	6	167	27.8333	0.3929	1.4118	0.0088
scanning	scanning	69	167	2.4203	0.3929	16.2353	0.0088
5a	5A	20	166	8.3000	0.3906	4.7059	0.0088
hacker	hacker	58	166	2.8621	0.3906	13.6471	0.0088
ddo	DDoS	38	165	4.3421	0.3882	8.9412	0.0087
executable	executables	82	165	2.0122	0.3882	19.2941	0.0087
phish	phishing	70	165	2.3571	0.3882	16.4706	0.0087
fakem	FakeM	6	164	27.3333	0.3859	1.4118	0.0087
hangover	Hangover	7	164	23.4286	0.3859	1.6471	0.0087
retrieve	retrieve	72	164	2.2778	0.3859	16.9412	0.0087

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
4d	4d	29	163	5.6207	0.3835	6.8235	0.0086
m	ms	43	163	3.7907	0.3835	10.1176	0.0086
extract	extracted	100	162	1.6200	0.3812	23.5294	0.0086
jindiqiao@hotmail.com	JinDiQIAO@hotmail.com	1	162	162.0000	0.3812	0.2353	0.0086
shimrat	ShimRat	1	162	162.0000	0.3812	0.2353	0.0086
dec	dec	53	161	3.0377	0.3788	12.4706	0.0085
embassy	embassy	31	161	5.1935	0.3788	7.2941	0.0085
arbor	Arbor	17	160	9.4118	0.3765	4.0000	0.0085
zip	zip	68	160	2.3529	0.3765	16.0000	0.0085
adversary	adversaries	47	159	3.3830	0.3741	11.0588	0.0084
winnti	Winnti	5	159	31.8000	0.3741	1.1765	0.0084
content-type	Content-Type	37	158	4.2703	0.3718	8.7059	0.0084
longer	longer	104	158	1.5192	0.3718	24.4706	0.0084
a1	a1	32	157	4.9063	0.3694	7.5294	0.0083
obfuscate	obfuscated	84	157	1.8690	0.3694	19.7647	0.0083
sandbox	sandbox	54	157	2.9074	0.3694	12.7059	0.0083
tools	Tools	37	157	4.2432	0.3694	8.7059	0.0083
extract	extract	91	156	1.7143	0.3671	21.4118	0.0083
cyberattack	cyberattack	22	155	7.0455	0.3647	5.1765	0.0082
feb	FEB	42	155	3.6905	0.3647	9.8824	0.0082
larger	larger	94	155	1.6489	0.3647	22.1176	0.0082
www.fidelissecurity.com	WWW.FIDELISSECURITY.COM	8	155	19.3750	0.3647	1.8824	0.0082
crysys	CrySyS	9	154	17.1111	0.3624	2.1176	0.0082
europe	europe	78	154	1.9744	0.3624	18.3529	0.0082
kitten	Kitten	12	154	12.8333	0.3624	2.8235	0.0082
directory	directories	77	153	1.9870	0.3600	18.1176	0.0081
po	PoS	19	153	8.0526	0.3600	4.4706	0.0081
sql	SQL	50	153	3.0600	0.3600	11.7647	0.0081
tmp	tmp	53	153	2.8868	0.3600	12.4706	0.0081

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
troy	Troy	7	153	21.8571	0.3600	1.6471	0.0081
pdb	PDB	40	152	3.8000	0.3576	9.4118	0.0081
decode	decoded	83	151	1.8193	0.3553	19.5294	0.0080
sophistication	sophistication	62	151	2.4355	0.3553	14.5882	0.0080
command-and-control	Command-and-control	81	150	1.8519	0.3529	19.0588	0.0079
germany	GeRMANy	72	150	2.0833	0.3529	16.9412	0.0079
randomly	randomly	66	150	2.2727	0.3529	15.5294	0.0079
tech	TECH	37	150	4.0541	0.3529	8.7059	0.0079
unsigned	unsigned	39	150	3.8462	0.3529	9.1765	0.0079
db	db	37	149	4.0270	0.3506	8.7059	0.0079
indicators	Indicators	89	148	1.6629	0.3482	20.9412	0.0078
institute	institute	52	148	2.8462	0.3482	12.2353	0.0078
mac	MAC	62	148	2.3871	0.3482	14.5882	0.0078
pitty	PITTy	4	148	37.0000	0.3482	0.9412	0.0078
seconds	seconds	62	147	2.3710	0.3459	14.5882	0.0078
worm	Worm	39	147	3.7692	0.3459	9.1765	0.0078
0x00	0x00	18	146	8.1111	0.3435	4.2353	0.0077
generic	generic	74	146	1.9730	0.3435	17.4118	0.0077
lnk	lnk	34	146	4.2941	0.3435	8.0000	0.0077
cve-2012-0158	CVE-2012-0158	59	145	2.4576	0.3412	13.8824	0.0077
jan	Jan	54	145	2.6852	0.3412	12.7059	0.0077
ptr	ptr	23	145	6.3043	0.3412	5.4118	0.0077
rundll32.exe	rundll32.exe	56	145	2.5893	0.3412	13.1765	0.0077
underground	underground	41	145	3.5366	0.3412	9.6471	0.0077
xp	XP	69	145	2.1014	0.3412	16.2353	0.0077
7e	7E	13	144	11.0769	0.3388	3.0588	0.0076
overlap	overlap	68	144	2.1176	0.3388	16.0000	0.0076
p.	p.	22	144	6.5455	0.3388	5.1765	0.0076
x509v3	X509v3	3	144	48.0000	0.3388	0.7059	0.0076
yahoo	Yahoo	50	144	2.8800	0.3388	11.7647	0.0076

Lemma	Original Word	Documents Appeared In	Appearances	Appearances per Document	Appearances per Corpus Document	% of Corpus Documents	% of Corpus Tokens
c++	C++	53	143	2.6981	0.3365	12.4706	0.0076
emissary	eMISSARy	5	143	28.6000	0.3365	1.1765	0.0076
exposureindicating	Exposureindicating	1	143	143.0000	0.3365	0.2353	0.0076
handshake	handshake	24	143	5.9583	0.3365	5.6471	0.0076
labs	Labs	59	142	2.4068	0.3341	13.8824	0.0075
www.threatgeek.com	www.threatgeek.com	7	142	20.2857	0.3341	1.6471	0.0075
lure	Lure	56	141	2.5179	0.3318	13.1765	0.0075
means	means	79	141	1.7848	0.3318	18.5882	0.0075
dukes	Dukes	4	140	35.0000	0.3294	0.9412	0.0074
affairs	Affairs	45	139	3.0889	0.3271	10.5882	0.0074
ch	Ch	20	139	6.9500	0.3271	4.7059	0.0074
ltd.	Ltd	31	139	4.4839	0.3271	7.2941	0.0074
georgian	Georgian	9	138	15.3333	0.3247	2.1176	0.0073
newer	newer	81	138	1.7037	0.3247	19.0588	0.0073
nov	Nov	43	138	3.2093	0.3247	10.1176	0.0073
beijing	Beijing	36	137	3.8056	0.3224	8.4706	0.0073
cve	CVE	56	137	2.4464	0.3224	13.1765	0.0073
encrypt	Encrypt	78	137	1.7564	0.3224	18.3529	0.0073
filename	filenames	81	137	1.6914	0.3224	19.0588	0.0073
terracotta	TERRACOTTA	1	137	137.0000	0.3224	0.2353	0.0073
tibet	TIBeT	21	137	6.5238	0.3224	4.9412	0.0073
timestamp	timestamps	48	137	2.8542	0.3224	11.2941	0.0073
www.crysys.hu	www.crysys.hu	3	137	45.6667	0.3224	0.7059	0.0073
ee	ee	25	136	5.4400	0.3200	5.8824	0.0072
fox-it	fox-it	3	136	45.3333	0.3200	0.7059	0.0072
kunming	Kunming	2	136	68.0000	0.3200	0.4706	0.0072

APPENDIX C. SAMPLE OF THE ONTOLOGY

This section contains a partial sample of the ontology in RDF/OWL format. The full ontology is available digitally.

```
<?xml version="1.0"?>
<rdf:RDF
xmlns="http://www.semanticweb.org/coreythomasholzer/ontolog
ies/apt#"

xml:base="http://www.semanticweb.org/coreythomasholzer/onto
logies/apt"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
    <owl:Ontology
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt"/>

    <!--

//////////////////////////////////////////////////////////
//////////////////////////
    //
    // Data properties
    //

//////////////////////////////////////////////////////////
//////////////////////////
     -->


    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Aliases -->

    <owl:DatatypeProperty
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Aliases"/>

    <!--

//////////////////////////////////////////////////////////
//////////////////////////
```

```
    //
    // Classes
    //

///////////////////////////////////////////////////////////
//////////////////////////
     -->


    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
# -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#"/>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#APT -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#APT">
        <owl:disjointWith
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Protocol"/>
        <rdfs:comment xml:lang="en">An advanced persistent
threat (APT) is the name given to a network attack in which
an unauthorized organization, also identified as an APT,
gains access to a network and stays there undetected for a
long period of time in order to execute a complex
attack.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#APT_Attack -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#APT_Attack">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT"/>
        <rdfs:comment>An advanced persistent threat (APT)
is a network attack in which an unauthorized person gains
```

```
access to a network and stays there undetected for a long
period of time. The intention of an APT attack is to steal
data rather than to cause damage to the network or
organization.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#APT_Organization -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#APT_Organization">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT"/>
        <rdfs:comment></rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Actions_on_Objective_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Actions_on_Objective_Phase">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment xml:lang="en">During this phase the
APT is actively going after the data that they originally
identified as their target (Hutchins et al., 2011). The APT
uses previously installed software to determine the network
layout including, but not limited to, mapping the hosts of
networked drives, database servers, domain controllers,
PKI, etc. (Ask et al., 2013). The goal here is to footprint
the network and to establish a network account and elevate
the privileges for that account (Ask et al., 2013). During
this phase, the APT will also seek to compromise more hosts
in order to strengthen its foothold in the target network.
The extraction of the target data may also be accomplished
using custom encryption and/or tunneling within other
protocols to hide the data from security professionals
(Websense, 2011).</rdfs:comment>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Administrators -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Administrators">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Those who manage or administer
networks and servers.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Affiliates -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Affiliates">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Anti-Forensics -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Anti-Forensics">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Hardening_Techniques"/>
        <rdfs:comment>Anti-forensic techniques try to
frustrate forensic investigators and their techniques to
analyze and understand malware. This can include refusing
to run when debugging mode is enabled, refusing to run when
running inside of a virtual machine, or deliberately
overwriting data.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Anti-Malware -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Anti-Malware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Anti-Spyware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Anti-Spyware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Antivirus -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Antivirus">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Backdoor -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Backdoor">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Bank_Accounts -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Bank_Accounts">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Individuals"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#BotNet_Managers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#BotNet_Managers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Botnet -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Botnet">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>a network of private computers
infected with malicious software and controlled as a group
without the owners&apos; knowledge, e.g., to send spam
messages.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#C2 -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#C2">
```

```
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Command-and-Control"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>The centralized computer that issues
commands to a botnet (zombie army) and receives reports
back from the coopted computers.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#C2_Server -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#C2_Server">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Command-and-Control"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>The centralized computer that issues
commands to a botnet (zombie army) and receives reports
back from the coopted computers.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#CA -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#CA">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Certification_Authority"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
        <rdfs:comment>In cryptography, a certificate
authority or certification authority (CA) is an entity that
issues digital certificates. A digital certificate
certifies the ownership of a public key by the named
subject of the certificate.</rdfs:comment>
```

```
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#CNC -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#CNC">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Command-and-Control"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Certificate_Authority -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Certificate_Authority">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Certification_Authority"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
        <rdfs:comment>In cryptography, a certificate
authority or certification authority (CA) is an entity that
issues digital certificates. A digital certificate
certifies the ownership of a public key by the named
subject of the certificate.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Certification_Authority -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Certification_Authority">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
```

```
        <rdfs:comment>In cryptography, a certificate
authority or certification authority (CA) is an entity that
issues digital certificates. A digital certificate
certifies the ownership of a public key by the named
subject of the certificate.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Command-and-Control -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Command-and-Control">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#C&amp;amp;C"/>
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#C&amp;amp;C_Server"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>The centralized computer that issues
commands to a botnet (zombie army) and receives reports
back from the coopted computers.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Command_and_Control_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Command_and_Control_Phase">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment xml:lang="en">The Command and Control
phase begins once the infected host beacons the C2 server
(Hutchins et al., 2011). Attackers need to maintain access
to the victim's network means that each communication with
a compromised system (Auty, 2015). During this phase the
APT will seek to obtain elevated privileges on the system
and will install additional software to facilitate the
attack (i.e., encryption) on compromised system and network
(Ask et al., 2013). While the initial installation is
```

```
achieved by software designed to exploit a zero-day
vulnerability, the additional software is likely to be
commonly known software that may even be approved to
operate on the network for legitimate activities (e.g.,
SSH, SecureFTP, etc.) (Ask et al., 2013).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Common_Vulnerabilities_and_Exposures -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Common_Vulnerabilities_and_Exposures">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
        <rdfs:comment>CVE is a list of information security
vulnerabilities and exposures that aims to provide common
names for publicly known cyber security issues. The goal of
CVE is to make it easier to share data across separate
vulnerability capabilities (tools, repositories, and
services) with this &quot;common
enumeration.&quot;</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Communication_Pathway -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Communication_Pathway">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Communications"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Communications -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Communications">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Components"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Components -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Components">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Attack"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Compression_Algorithm -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Compression_Algorithm">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Compression is a reduction in the
number of bits needed to represent data. Compressing data
can save storage capacity, speed file transfer, and
decrease costs for storage hardware and network
bandwidth.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Compression_Program -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Compression_Program">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>An application which employs one or
more Compression Algorithms for the purpose of compressing
files or directory structures.</rdfs:comment>
```

```
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Credit_Cards -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Credit_Cards">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Individuals"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Criminal_Services -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Criminal_Services">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#DDoS_Attackers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#DDoS_Attackers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Responsible for managing the DDoS
attacks that are part of the APT Campaign.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Data_Evaluators -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Data_Evaluators">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Personnel who filter through data
looking for the information that has value.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Database_Server -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Database_Server">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Service"/>
        <rdfs:comment>Database server is the term used to
refer to the back-end system of a database application
using client/server architecture. The back-end, sometimes
called a database server, performs tasks such as data
analysis, storage, data manipulation, archiving, and other
non-user specific tasks.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Defensive_Counter_Measures -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Defensive_Counter_Measures">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Steps taken by APT Organizations to
reduce the chance of detection</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Defensive_Techniques -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Defensive_Techniques">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Defined_Path -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Defined_Path">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Delivery_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Delivery_Phase">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment xml:lang="en">In the Delivery phase,
the APT transmits the weapon to the targeted system
(Hutchins et al., 2011). Lockheed Martin identifies the
most common delivery methods as email attachments, websites
and removable media. In addition to those three, Ask,
et.al. (Ask et al., 2013) identified social media as
another means for launching at attack against an individual
within the target organization. For the attack to move
beyond this phase, the targeted individual most click on
the link, attachment, or application for the attack to move
into the next phase (Auty, 2015).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Domain_Name -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Domain_Name">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Domain_Name_Server -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Domain_Name_Server">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Service"/>
        <rdfs:comment>Domain Name Servers (DNS) are the
Internet&apos;s equivalent of a phone book. They maintain a
directory of domain names and translate them to Internet
Protocol (IP) addresses. This is necessary because,
although domain names are easy for people to remember,
computers or machines, access websites based on IP
addresses.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Downloader -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Downloader">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment>An application that will download and
install other Trojans onto your computer.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Dropper -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Dropper">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Encryption_Algorithm -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Encryption_Algorithm">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>What we all call Triple DES is EDE
(encrypt, decrypt, encrypt). The way that it works is that
you take three 56-bit keys, and encrypt with K1, decrypt
with K2 and encrypt with K3. There are two-key and three-
key versions. Think of the two-key version as merely one
where K1=K3. Note that if K1=K2=K3, then Triple DES is
really Single DES.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Encryption_Program -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Encryption_Program">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Programs designed to encrypt data at
rest (DAR) on a local machine or server.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Exploit_Kit -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Exploit_Kit">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Exploitation_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Exploitation_Phase">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment xml:lang="en">Exploitation involves
compromising the host machine on the network. It is where
the weaponized tool is triggered (Hutchins et al., 2011).
The exploitation can be of a flaw in the operating system
or an individual application on the host (Ask et al., 2013;
Hutchins et al., 2011).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#FTP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#FTP">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#File_Transfer_Protocol"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TCP"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Fences -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Fences">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Converts stolen data into
money</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#File_Transfer_Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#File_Transfer_Protocol">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Transmission_Control_Protocol"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Firewall -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Firewall">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#For_Profit_Organization -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#For_Profit_Organization">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Forensics_Tools -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Forensics_Tools">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Funding -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Funding">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
        <rdfs:comment>Sources of funding for conducting
criminal or espionage activities.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Government_Agency_or_Department -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Government_Agency_or_Department">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#HTTP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#HTTP">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#HyperText_Transfer_Protocol"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TCP"/>
```

```
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#HTTPS -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#HTTPS">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Secure_HyperText_Transfer_Protocol"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TCP"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Hardening_Techniques -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Hardening_Techniques">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>hardening is usually the process of
securing a system or program by reducing its surface of
vulnerability, which is larger when a system performs more
functions</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#HyperText_Transfer_Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#HyperText_Transfer_Protocol">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Transmission_Control_Protocol"/>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#ICS -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#ICS">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Incident_Command_System"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Software"/>
        <rdfs:comment>The Incident Command System (ICS) is
a standardized approach to the command, control, and
coordination of emergency response[1] providing a common
hierarchy within which responders from multiple agencies
can be effective.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#IOC -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#IOC">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
        <rdfs:comment>IOC (indicator of compromise) - a
list of threat data (e.g., strings defining file paths or
registry keys) which can be used to detect a threat in the
infrastructure using automated software-based
analysis.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#IP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#IP">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Impersonated_Software -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Impersonated_Software">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Malicious software with a name
matching legitimate software on a computer.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Incident_Command_System -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Incident_Command_System">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Software"/>
        <rdfs:comment>The Incident Command System (ICS) is
a standardized approach to the command, control, and
coordination of emergency response[1] providing a common
hierarchy within which responders from multiple agencies
can be effective.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Indicators_of_Compromise -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Indicators_of_Compromise">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
```

```
        <rdfs:comment xml:lang="en">IOC (indicator of
compromise) – a list of threat data (e.g., strings defining
file paths or registry keys) which can be used to detect a
threat in the infrastructure using automated software-based
analysis.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Injection -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Injection">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Injection_File"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Injection_File -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Injection_File">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>This appears to be a legitimate
document in one of many common file formats (DOC, PPT, XLS,
PDF, etc.) which contains a payload of malicious code. When
the document is opened the code is executed and the host
computer is infected.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Installation_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Installation_Phase">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment>The next phase of the attack is the
Installation phase. Installation refers to the installation
of a Remote Administration Tool (RAT) or backdoor that the
APT can use to gain control of the target’s computer (Ask
et al., 2013; Hutchins et al., 2011). Once the victim
triggers the malicious code (e.g. by clicking the malicious
link, opening the infected file, or visiting the
compromised site, etc.) the code reaches back to its
Command and Control (C2) server and provides the attacker
with useful information about the target network’s
environment that could be useful in executing the later
stages of the APT attack (Ask et al., 2013). Once installed
the RAT can also lay dormant until the C2 server connects
to it (Ask et al., 2013; Sikorski &amp;amp; Honig,
2012).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Installer -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Installer">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment xml:lang="en">A self-extracting,
self-installing file that delivers a payload of malicious
software to an unsuspecting user&amp;apos;s
computer.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Installers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Installers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
```

```
        <rdfs:comment>Software designed to build installing
applications for other programs.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Intrusion_Detection_System -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Intrusion_Detection_System">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
        <rdfs:comment>Intrusion Detection System (IDS) is a
type of security management system for computers and
networks. An IDS gathers and analyzes information from
various areas within a computer or a network to identify
possible security breaches, which include both intrusions
(attacks from outside the organization) and misuse (attacks
from within the organization).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Intrusion_Prevention_System -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Intrusion_Prevention_System">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
        <rdfs:comment>An Intrusion Prevention System (IPS)
is a network security/threat prevention technology that
examines network traffic flows to detect and prevent
vulnerability exploits.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Keylogger -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Keylogger">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment>A keylogger is a type of surveillance
software (considered to be either software or spyware) that
has the capability to record every keystroke you make to a
log file, usually encrypted. A keylogger recorder can
record instant messages, e-mail, and any information you
type at any time using your keyboard.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Legitimate_Applications -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Legitimate_Applications">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>These are applications that are
developed for legimate and legal purposes.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Legitimate_Business_Partners -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Legitimate_Business_Partners">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Logic_Bomb -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Logic_Bomb">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
```

```
        <rdfs:comment xml:lang="en">A logic bomb is
malicious code embedded within an application that executes
based on certain events. The logic bomb lies dormant until
that event occurs.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#MD5_Hash -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#MD5_Hash">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#MUTEX -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#MUTEX">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#MaaS -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#MaaS">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware_as_a_Service"/>
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TaaS"/>
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Trojan_as_a_Service"/>
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Mail_Server -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Mail_Server">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Service"/>
        <rdfs:comment>A mail server (also known as a mail
transfer agent or MTA, a mail transport agent, a mail
router or an Internet mailer) is an application that
receives incoming e-mail from local users (people within
the same domain) and remote senders and forwards outgoing
e-mail for delivery.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malicious_Domain -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malicious_Domain">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Domains which are registered and
maintained to conduct malicious operations. Usually used to
impersonate legitimate sites but which are designed to
distribute malicious programs to unsuspecting
victims.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malicious_Host -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malicious_Host">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malicious_IP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malicious_IP">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Components"/>
        <rdfs:comment xml:lang="en">Malware refers to any
type of malicious software that tries to infect a computer
or mobile device.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malware_Analysis -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malware_Analysis">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malware_Packers -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malware_Packers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malware_Testers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malware_Testers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Malware_as_a_Service -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Malware_as_a_Service">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TaaS"/>
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Trojan_as_a_Service"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment>Cybercriminals are increasingly
offering malware as a cloud-based on-demand service. ...
Rather than turning a profit just once by selling a
security exploit as a one-off, authors of malicious
software are now selling malware as a cloud-based
service.</rdfs:comment>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Manager -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Manager">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Runs the APT Organization and
develops its infrastructure</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Metamorphic -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Metamorphic">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Polymorphic"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment>This type of malware is constructed
in such a manner that is can re-engineer or recode itself
(Raj et al., 2014; Sikorski &amp;amp; Honig, 2012). This
recoding can take place each time it propagates or is
distributed through a network. This type of malware hinders
the use of signature-based protection tools (Raj et al.,
2014).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Money_Flow_Managers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Money_Flow_Managers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Motivation -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Motivation">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
        <rdfs:comment>The Organization&amp;apos;s motivation
for conducting their APT attacks.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Mules -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Mules">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#NFP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#NFP">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Not_for_Profit"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#NGO -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#NGO">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Non-Government_Organization"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Non-Government_Organization -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Non-Government_Organization">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Not_for_Profit -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Not_for_Profit">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#OS_Component -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#OS_Component">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System_Component"/>
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#OS_Tool -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#OS_Tool">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System_Tool"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Obfuscation -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Obfuscation">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Hardening_Techniques"/>
        <rdfs:comment>Obfuscation and its subset, packing,
are techniques used by malware developers to make static
analysis more difficult for the forensics experts (Brand et
al., 2010; Sikorski &amp;amp; Honig, 2012). Obfuscation is a
means of hiding or disguising code (Sikorski &amp;amp; Honig,
2012).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#On-the-fly-Encryption -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#On-the-fly-Encryption">
```

```
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Real-time_Encryption"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
        <rdfs:comment>On-the-fly encryption (OTFE), also
known as real-time encryption and transparent encryption,
is a method used by some disk encryption software. ... In
general, every method in which data is transparently
encrypted on write and decrypted on read can be called on-
the-fly encryption.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Operating_System -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Operating_System">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Software"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Operating_System_Component -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Operating_System_Component">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Operating_System_Tool -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Operating_System_Tool">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Operating_System"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#POSMalware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#POSMalware">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Point_of_Sale_Malware"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Packing -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Packing">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Hardening_Techniques"/>
        <rdfs:comment>Obfuscation and its subset, packing,
are techniques used by malware developers to make static
analysis more difficult for the forensics experts (Brand et
al., 2010; Sikorski &amp;amp; Honig, 2012). Obfuscation is a
means of hiding or disguising code (Sikorski &amp;amp; Honig,
2012). Packing uses compression and a wrapping program as a
means of disguising the true purpose of program (Sikorski
&amp;amp; Honig, 2012). Even more challenging for analysts and
malware detection is recursive packaging which obfuscates
code in multiple layers of recursive compression (Egele et
al., 2012).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Password_Recovery -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Password_Recovery">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
        <rdfs:comment>Password Recovery Software to help
recover lost and forgotten passwords</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Personally_Identifiable_Information -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Personally_Identifiable_Information">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Individuals"/>
        <rdfs:comment xml:lang="en">Personally Identifiable
Information (PII) is data that can be used to impersonate
someone or used to steal their identity and commit
fraudulent acts while pretending to be that
person.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Personnel -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Personnel">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
        <rdfs:comment>The different skillsets of
individuals working for the group.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Phases -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Phases">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Attack"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Phishing -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Phishing">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>the activity of defrauding an online
account holder of financial information by posing as a
legitimate company.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Point_of_Sale_Malware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Point_of_Sale_Malware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment xml:lang="en">A Point of Sale Malware
(POSMalware) is designed to attack and exploit POS
systems.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Polymorphic -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Polymorphic">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment>This type of malware is constructed
in such a manner that is can re-engineer or recode itself
(Raj et al., 2014; Sikorski &amp;amp; Honig, 2012). This
recoding can take place each time it propagates or is
distributed through a network. This type of malware hinders
the use of signature-based protection tools (Raj et al.,
2014).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Programmers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Programmers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Programming_Language -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Programming_Language">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Software"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Protectors -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Protectors">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Hardening_Techniques"/>
```

```
        <rdfs:comment>Obfuscation and its subset, packing,
are techniques used by malware developers to make static
analysis more difficult for the forensics experts (Brand et
al., 2010; Sikorski &amp;amp; Honig, 2012). Packing uses
compression and a wrapping program as a means of disguising
the true purpose of program (Sikorski &amp;amp; Honig, 2012).
Even more challenging for analysts and malware detection is
recursive packaging which obfuscates code in multiple
layers of recursive compression (Egele et al.,
2012).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Protocol">
        <rdfs:comment xml:lang="en">A protocol defines
rules and conventions for communication between network
devices.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Ransomware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Ransomware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment xml:lang="en">Ransomware holds a
user’s data hostage. The latest ransomware variants encrypt
the user’s data, thus making it unusable until a ransom is
paid to retrieve the decryption key.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Real-time_Encryption -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Real-time_Encryption">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Legitimate_Applications"/>
        <rdfs:comment>On-the-fly encryption (OTFE), also
known as real-time encryption and transparent encryption,
is a method used by some disk encryption software. ... In
general, every method in which data is transparently
encrypted on write and decrypted on read can be called on-
the-fly encryption.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Reconnaissance_Phase -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Reconnaissance_Phase">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phases"/>
        <rdfs:comment xml:lang="en">Reconnaissance is the
selection and identification of the desired target. In this
stage the APT is footprinting the target organization and
collecting information including but not limited to names,
positions, email addresses, physical locations, operating
systems, etc. (Hutchins et al., 2011). Through this
information gathering process the APT determines who has
ownership of the desired information that they seek to
steal (2013). The APT will determine which employee to
compromise as well as a potential means for doing
so.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Recruiters -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Recruiters">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Find individuals with the talents
need to complete a specific task or develop a specific
component of the APT Attack.</rdfs:comment>
```

```
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Recursive_Packaging -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Recursive_Packaging">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Hardening_Techniques"/>
        <rdfs:comment>Obfuscation and its subset, packing,
are techniques used by malware developers to make static
analysis more difficult for the forensics experts (Brand et
al., 2010; Sikorski &amp;amp; Honig, 2012). Obfuscation is a
means of hiding or disguising code (Sikorski &amp;amp; Honig,
2012). Packing uses compression and a wrapping program as a
means of disguising the true purpose of program (Sikorski
&amp;amp; Honig, 2012). Even more challenging for analysts and
malware detection is recursive packaging which obfuscates
code in multiple layers of recursive compression (Egele et
al., 2012).</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Registry_Key -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Registry_Key">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Remote_Access_Toolkit -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Remote_Access_Toolkit">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Remote_Access_Trojan"/>
```

```
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Remote_Administration_Toolkit"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Remote_Access_Trojan -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Remote_Access_Trojan">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Remote_Administration_Toolkit"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Remote_Administration_Toolkit -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Remote_Administration_Toolkit">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Rootkit -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Rootkit">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment xml:lang="en">A rootkit is the
combination of programs designed to infect your computer
```

```
without being detected. Your antivirus application
communicates with your operating system to identify
threats. However, rootkits breaks down this communication
process.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#SFTP -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#SFTP">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Secure_File_Transfer_Protocol"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TCP"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#SHA1_Hash -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#SHA1_Hash">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#SHA256_Hash -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#SHA256_Hash">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>
```

```
    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#SSH -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#SSH">
        <owl:equivalentClass
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Secure_Shell_Protocol"/>
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#TCP"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Secure_File_Transfer_Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Secure_File_Transfer_Protocol">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Transmission_Control_Protocol"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Secure_HyperText_Transfer_Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Secure_HyperText_Transfer_Protocol">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Transmission_Control_Protocol"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Secure_Shell_Protocol -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Secure_Shell_Protocol">
```

```
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Transmission_Control_Protocol"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Security_Firm -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Security_Firm">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Security_and_Protective_Measures -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Security_and_Protective_Measures">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Attack"/>
        <rdfs:comment xml:lang="en">Tools and techniques
that cybersecurity professionals can use to protect
networks and systems against various hacking attacks
including elements of APT Attacks.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Security_as_a_Service -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Security_as_a_Service">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Security_and_Protective_Measures"/>
        <rdfs:comment xml:lang="en">Security-as-a-service
(SaaS) is an outsourcing model for security management.
Typically, Security as a Service involves applications such
as anti-virus software delivered over the Internet but the
```

```
term can also refer to security management provided in-
house by an external organization.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Service -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Service">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Target_Software"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Service_Renters -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Service_Renters">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>A person who acts on the
organization&apos;s behalf to secure
services.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Signed_Certificate -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Signed_Certificate">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Indicators_of_Compromise"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Social_Engineering -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Social_Engineering">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Techniques_and_Tools"/>
        <rdfs:comment>Social engineering is an attack
vector that relies heavily on human interaction and often
involves tricking people into breaking normal security
procedures.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Social_Engineers -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Social_Engineers">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Individuals who specialize in
conducting social engineering campaigns</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Spam_Distributors -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Spam_Distributors">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Personnel"/>
        <rdfs:comment>Personnel who distribute spam
messages to targeted individuals or groups of
individuals.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Spearphishing -->
```

```
    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Spearphishing">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Phishing"/>
        <rdfs:comment>Spear phishing is an email that
appears to be from an individual or business that you know.
But it isn&apos;t. It&apos;s from the same criminal hackers
who want your credit card and bank account numbers,
passwords, and the financial information on your
PC.</rdfs:comment>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Sponsor_State -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Sponsor_State">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#APT_Organization"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/coreythomasholzer/ontologies/apt
#Spyware -->

    <owl:Class
rdf:about="http://www.semanticweb.org/coreythomasholzer/ont
ologies/apt#Spyware">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/coreythomasholzer/
ontologies/apt#Malware"/>
        <rdfs:comment xml:lang="en">Spyware is installed on
a machine without the user’s awareness or consent. Spyware
attempts to gather specific user information and send it to
a third party.</rdfs:comment>
    </owl:Class>
```

VITA

Corey T. Holzer
Graduate School, Purdue University

Education

2009 - MS, Networking and Communications Management, Keller Graduate School of Management, Downers Grove, IL
2009 - MBA, Information Security Management, Keller Graduate School of Management, Downers Grove, IL
1994 - MA, Government and Politics, St. John's University, Jamaica NY
1992 - BA, Government and Politics, St. John's University, Jamaica NY

Military Education

2009 - Captain's Career Course (Signal), Fort Gordon, GA
2007 - Basic Officer's Leader Course III, Fort Gordon, GA
2006 - Basic Officer's Leader Course II, Fort Benning, GA
2006 - Basic Officer's Leader Course I (Officer Candidate School), Fort Benning, GA

Certifications

2016 - Certified Ethical Hacking, EC Council
2014 - Security+ ce, CompTIA
2014 - Certified Information Systems Security Professional, $(ISC)^2$

Publications

Accepted - Employing Link Analysis for the Improvement of Threat Intelligence Regarding Advanced Persistent Threats, Holzer, Dietz, and Yang

Presentations

2016 - Guest Speaker - CERIAS Security Seminar - The Application of Natural Language Processing to Open Source Intelligence for Ontology Development in the Advanced Persistent Threat Domain, Holzer
2016 - IEEE Homeland Security and Technology - The Ethics of Hacking Back, Holzer and Lerums
2016 - CERIAS Symposium - Using Link Analysis to Improve Advanced Persistent Threat Intelligence and Detection, Holzer
2015 - CERIAS Symposium - Assessing Risk and Cyber Resiliency, Holzer and Merritt

Military Experience

2015-2016 - Student, US Army Student Detachment, West Lafayette, IN
2012-2014 - Network Branch Deputy Chief, Capabilities Development and Integration Directorate, Fort Sill, OK
2010-2012 - Company Commander, 507th Signal Company, Fort Wainwright, AK
2010 - Alaska Regional Network Operations and Security Center Officer-in-Charge, Joint Base Elmendorf-Richardson, AK
2008-2009 - Battalion Communications Officer-in-Charge, 2/320 FAR, Balad, Iraq and Fort Campbell, KY
2008 - Battalion Communications Officer-in-Charge, 1-101 STB, Tikrit, Iraq
2007-2008 - Signal Platoon Leader, C Company, Fort Campbell, KY and Tikrit, Iraq

Personal Awards

Meritorious Service Medal
Army Commendation Medal (3 Awards)
Overseas Service Ribbon (2 Awards)

Work Experience

2004-2006 - Systems Administrator, MCSP, Inc., Winter Haven, FL
2000-2004 - Manager of Internet Technologies, J. Walter Thompson, New York, NY
1997-2000 - Network Administrator, Organization Resources Counselors, New York, NY
1996-1997 - Helpdesk Technician, World Wrestling Federation, Stanford, CT
1992-1995 - PROFS Administrator, Donovan Data Systems, New York, NY