EXAMINING THREAT GROUPS FROM THE OUTSIDE: GENERATING HIGH-LEVEL OVERVIEWS OF PERSISTENT AND TRADITIONAL COMPROMISES

by

Angela Marie Horneman

B.S., Robert Morris University, 2004

M.S., University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

The University of Pittsburgh in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

School of Information Science

This thesis was presented

by

Angela Marie Horneman

It was defended on

November 18, 2013

and approved by

Prashant Krishnamurthy, PhD, Associate Professor, School of Information Science

Balaji Palanisamy, PhD, Assistant Professor, School of Information Science

Committee Co-Chair: Timur Snoke, MS, Network Defense Analyst, Software Engineering Institute

Committee Co-Chair: James B.D. Joshi, PhD, Associate Professor, School of Information Science

EXAMINING THREAT GROUPS FROM THE OUTSIDE: GENERATING HIGH-LEVEL OVERVIEWS OF PERSISTENT AND TRADITIONAL COMPROMISES

Angela Marie Horneman, M.S.

University of Pittsburgh, 2013

Analyzing threats that have compromised electronic devices is important to compromised organizations, researchers, and law enforcement. Examination of network and host based logs and network traffic is effective in identifying threats, the impact, and how to recover from the compromise. However, this form of analysis is very time consuming and requires technical expertise. This traditional form of analysis also only will provide information concerning organizations that have those logs and network flows. A quick and easy to use methodology for generating a high level overview of threats' targets globally would aid analysts by indicating areas of focus for more in-depth analysis.

In this thesis we propose a methodology for synthesizing information from multiple publicly available, scope limited data sets that allows a rapid and cheap compilation of an overview of a threat. This method has the additional benefits of being available to researchers outside of compromised organizations and of being possible when logs and network flow do not exist. Once the approach has been implemented, it can be used to analyze multiple threats. This is demonstrated by two case studies, one examining a persistent threat called Advanced Persistent Threat 1 and the other overviewing a more traditional threat, the malware family Mabeza Infected.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thank you, Joel and Deana, for proof-reading and moral support. Thank you to all the analysts that provided suggestions and discussions regarding the methodology and case studies. To my fellow SFS students, thank you for attending the defense. I really appreciate it! This work was funded in part by NSF DUE award #1027167.

1.0 INTRODUCTION

The use of electronic devices like personal computers, servers, phones, and routers for storing information, conducting business, and transferring data makes them valuable property and consequently the target of various forms of theft. Unlike with other property, theft can occur even if a device never leaves the possession of its owner. Any device that is ever connected to a network is vulnerable to this type of theft. If an adversary can successfully get software of its choosing, termed malware, installed on a device, the adversary can effectively control the device and use it for its own purposes. For instance they can retrieve information and eavesdrop on communications, which are both thefts of information. They could also process data or use a device as a bridge to access other devices, which are both thefts of computing resources. When these thefts occur, analysts say the device has been exploited.

Exploitations may be the result of the targeting of specific people or organizations by a group of people or could be more opportunistic through some form of automation. Examples of targeting are denial of service (DoS) attacks and cyber espionage of businesses or governments. Viruses, Trojans, worms, and other malware are examples of exploitations that spread manually or automatically and seldom target a specific person or organization. These exploitations are the result of traditional or advanced persistent threats.

Traditional Threats. Malware can be categorized in a number of ways. Most generically categorization is based on how the malware spreads, where we find terms such as virus, Trojan, and worm. Another method of categorization is by code characteristic. Often, one piece of malware will go through rounds of modification to hone its capabilities or to evade malware detection. These variations are detectable with some forms of analysis and are used to create malware groups. For instance, the Conficker worm was not one set of lines of code, but had multiple versions which changed how it spread [1]. Malware groups can also be created by grouping malware into sets that are found together on the same device or originate from the same adversary. For instance, the Russian Business Network tool pack included malware for key logging and botnets [2]. Each malware group can be thought of as an individual threat.

Threats to devices are not just malware, though. People are the most basic threat and are behind all other threats. While some perpetrators will use one particular piece of malware or one malware family, others have an arsenal of malware and other methods that they use to engage in exploitation. The Russian Business Network tool pack mentioned above is an example of a set of malware [2]. Non-malware methods of exploitation are often a form of social engineering, methods such as phishing e-mails asking for account numbers and passwords or phone calls purporting to be the help desk or a business executive requesting log in information. Other non-malware methods include wiretaps, network analysis, and network eavesdropping. All these methods are the traditional threats that have been the main focus of security professionals for years.

Advanced Persistent Threats. In addition to a wide range of modes of operation, people also have varying levels of persistence in conducting exploitations. Those people with specific targets are more likely to be persistent in finding exploits that work against those targets than people who are satisfied with any exploitation that successfully executes against any device. The most persistent people, and certain forms of malware they use, have come to be termed advanced persistent threats. According to the National Institute of Standards and Technology (NIST), an advanced persistent threat (APT) is [3]:

> *[a]n adversary that possesses sophisticated levels of expertise and significant resources which allow it to create opportunities to achieve its objectives by using multiple attack vectors (e.g., cyber, physical, and deception). These objectives typically include establishing and extending footholds within the information technology infrastructure of the targeted organizations for purposes of exfiltrating information, undermining or impeding critical aspects of a mission, program, or organization; or positioning itself to carry out these objectives in the future. The advanced persistent threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives.*

Often APTs use similar tools to exploit devices, but in contrast to the relatively static and opportunistic traditional threats, these threats engage in *on-going and repetitive, adaptive exploitation against a single target*.

Whether people or sets of related malware, both traditional and APT threats are of concern to many people. They threaten productivity, privacy, finances, health, and general well-being. There are threats with the goal of shutting down corporate or government networks. Some threats steal personal information. Others steal money or ideas. As we hear about in the news on occasion, there are threats that can infect medical devices, cars, phones, and the devices that control electric, gas, water, and sewage management. Knowing how these threats propagate, how they behave, what they target, their ultimate goal, and, most importantly, how to recover from them and prevent their future success is very important.

Irrespective of how a device is compromised, it is important to learn how to lessen any impact from the compromise and prevent compromise from similar threats in the future. Internal, detailed analysis through host and network based forensics of compromised machines is important for in-depth understanding of threats. These methods are necessary to determine how an adversary got into a network and obtained control of a machine, as well as to determine what happened after they gained control. Without in-depth analysis, it is hard, if not impossible, to learn the adversary’s goal, gain, and mode of operation—knowledge that is necessary to mitigate the costs of being a victim of compromise and to learn how to prevent further compromise in the future. However, high-level analysis can provide important information and mitigate some of the issues with traditional analysis. This thesis proposes a method that uses multiple data sets to facilitate that high-level analysis.

1.1 MOTIVATION AND RELATED WORK

Analysts with access to detailed device logs and network flow have methods and tools to help them use those sources to understand the threats, but these existing methods have some issues. Two major issues are that traditional analysis is time consuming and inaccessible to external researchers. Logs and network flows detailed enough to provide information on a threat are large and most of the data is irrelevant, being generated from the expected everyday activities on a device or network. This means that analysts must sift through the data to find the information that lends to their analysis. Using logs and flow information to analyze a threat can take multiple analysts days, weeks, months, or even years to fully understand a threat and its impact [3].

Advanced Persistent Threats. There are several short-comings with the existing traditional methods of analysis, especially in the view of advanced persistent threats. Referring back to the NIST definition of APT above, the *on-going and repetitive, adaptive exploitation against a single target* accentuates the issues of analysis time and data access. These issues can limit the ability of organizations to understand the threat and be able to quickly respond as the threat adapts. With these threats, it can be more helpful to know a little bit of relevant information on a threat as soon as possible than it would be to know details of the threat but have to wait weeks or months. For an exploited organization, having an idea of whom and what to monitor could lessen the threat's effectiveness in exploitation success. For academic and industry researchers, the detailed logs and network flow necessary for traditional analysis are seldom made available, even if requested as the information contained, or even the fact that an exploitation occurred, may result in bad publicity. Governments may be able to obtain the logs and flow, but doing so normally requires going through many legalities. The process can be undesirable as it may compromise on-going investigations. Unavailable information could also occur for exploited organizations as the result of misconfigured or non-existent auditing and traffic monitoring or because the threat itself removed the information to prevent analysis, which is one step in the life cycle of an APT [4]. Being able to tell something about a threat in this case is better than not knowing anything.

Another important motivation for this work is to provide data validation of other analysis results. When companies release reports regarding threats, such as the report on Advanced Persistent Threat 1 mentioned in the first case study, it would be helpful for external parties to be able to validate the findings. This is necessary for other organizations to evaluate the risk of the threat to themselves. It may also be necessary so governments can draft policies for mitigating threats that compromise national security or financial stability. To the best of our knowledge there exists no other method for validation that does not require the original data or an audit of a company.

There exists little work related to the specific considerations for evaluating APTs. Literature and handbooks currently assume traditional forensic methods as either a form of detection [5] or as one of the first steps in response [6]. Published research on generating a high-level picture of a threat is lacking, as is using multiple distinct data sets in threat analysis. The only work found presenting a method for either is the prior work done using a method similar to this study for describing infrastructure attributes of the Advanced Persistent Threat 1 group [7]. This study builds on that method, tuning the process for efficiency and automaticity, as well as making it more generic to be applicable to a wider range of threats.

Traditional Threats. In addition to the three short-comings of time, data access, and lack of ability to do data validation, the traditional analysis method suggests another motivation for this thesis. Even after using indicators to identify anomalies, there is still much information that may or may not be relevant to a specific threat, and not all the relevant information is of the same importance or value. With increasing number of devices connected to networks, this is becoming a big data issue. From discussions with analysts, we learned that the analysts find that the more they know before starting an in-depth investigation, the more effectively they can filter the available log and network flow information to find entries that are related to the threat. There are indicators that analysts can use to find anomalies, sometimes suggested in intrusion detection and other network monitoring tools' manuals, but this often still leaves much data that is irrelevant. Thus there is a need for a methodology that can quickly provide useful information specific to a threat that can further assist in filtering logs and network flow.

Much of the existing research work in describing malware pertains to classification. There have been numerous proposed methods to categorize malware by general behavior, for instance [8], [9], and [10], by coding similarities, as in [9] and [11], and state changes that occur on infected machines [12]. These forms of classification are at a very technical level and require the ability to examine malware binaries in detail, in a dedicated environment where the malware cannot interfere with other operations.

Other work in summarizing traditional threats has been done through in-depth analysis of infected systems or network traffic as in [13] and the work and line of business for Mandiant, Symantec, Trend Micro, and other security related firms. Companies such as Mandiant and Symantec do not publish their actual methods of doing analysis, but there are numerous manuals and courses on incident response and forensics that describe and teach methods for analysis that, like malware research, is quite technical and time consuming. All of these methods are valuable for understanding individual threats and for understanding how those threats evolve over time. These forms of detailed analysis help create better functioning anti-virus software, improve understanding of network and device vulnerabilities and security weaknesses, and increase knowledge of how to recover from exploitations. Using these methods for the more traditional threats of the various types of malware have the same issues as they do for APTs.

Multiple Data Sources for Analysis. To effectively analyze a threat, information must be available that pertains to the exploited devices and provides a range of information. It is unlikely that analysts would be able to find one data source that contains all the desired information, so obtaining information from multiple sources is necessary. Combining data sets to answer questions or provide meaningful reports is not a new idea. The field of data integration explores the idea and has led to or coincided with innovations such as adaptive query processing, XML, and enterprise information integration systems [14]. Relational databases also can be used to work with data sets that are related on attributes, but otherwise contain different types of information. Similarly, data warehousing builds on database concepts to optimize data for decision support, where the data may be from several databases or contain historical records [15]. Data integration is also a need with the no-SQL databases such as Hadoop [16].

The methodology presented here is not dependent on any of these technologies, and the simple implementation given as an example in chapter 3.0 is just an example of a set of text parsing scripts, requiring no skills or software other than basic programming skills and a compiler or parser. Any of these technologies could be utilized as a tool within the methodology and may provide functional or efficiency benefits beyond simple scripts.

1.2 THESIS GOALS AND CONTRIBUTIONS

Thesis Goals. As a first step in addressing the four short-comings of the existing traditional analysis approaches mentioned above, this thesis proposes *a methodology for using externally obtained, publicly available data to quickly provide a high-level overview of a threat, based on a known set of machines connected to the internet that are, or were, experiencing exploitation by that threat*. The key goal is to outline a method that tells useful characteristics of machines exploited by a threat, by integrating multiple data sets available to the research community. These characteristics can then be analyzed to provide information that can provide exploit notification and awareness, guide a more in-depth analysis, and inform decision makers for threat response or recovery. This differs from the existing works, which assume the more traditional methods of log and network flow analysis.

Contributions. The contributions of this thesis pertain to the four short-coming discussed above. For the time issue and providing preliminary threat information, the method allows quicker and cheaper compilation of general characteristics of machines connected to the internet that are used in a specific compromise and can assist in short-term threat response and helps provide guidance for more detailed analysis. Using externally obtained information that is not limited to one exploited organization addresses the data access and validation problems—information from a source other than the compromised organizations, means researchers can derive a description of a threat without access to the organizations' network flow data or compromised machines, which organizations and individuals may not be able or willing to provide.

While the methodology presented in this thesis was originally conceived as a method to research one specific advanced persistent threat, it is applicable to other persistent threats and to the traditional threats of malware as well. The two case studies presented in chapter 4.0 are examples of the information that can be obtained. The first case study is an example of applying the method to an advanced persistent threat, while the second applies the method to a traditional type of malware.

Limitations. There are some limitations to this proposed methodology. The three most important to note are: (i) the timeframe of exploit relative to timeframe of data sets, (ii) the reliability of the indicators, and (iii) the analyst's skill. Because the internet is volatile, it is important to choose data sets that contain information from the timeframe of exploitation. As is true for any time sensitive analysis method, if either the time period of exploitation or the time when the information in the source data set was accurate is not known, this method cannot be used. Reliable indicators should reflect an actual exploit, and must all reflect the same threat. The method as described here cannot determine if all the indicators are actually related. An analyst’s skill level is important even in generating a high-level overview. This methodology requires that analysts must have some understanding of how networks and devices function.

1.3 THESIS ORGANIZATION

The remainder of this thesis is organized as follows. Chapter 2.0 discusses the general method, data sources, and tools used. Chapter 3.0 explains the techniques applied to creation of a sample implementation. Chapter 4.0 details two case studies illustrating use of the sample implementation. Section 4.1 applies the method to generalize the in-depth analysis of Advanced Persistent Threat 1 as provided by Mandiant in their report release in February 2013. Section 4.2 applies the method to generate a high-level description of a threat where no description currently exists. Chapter 5.0 presents a summary and conclusions of the study. Section 5.1 presents some limitations of the methodology, while section 5.2 suggests future work.

2.0 THE PROPOSED METHODOLOGY

For an analysis method to lessen the short-comings discussed earlier and provide some indicators of what data may be most relevant for more detailed analysis, information should be available externally to a compromised device or system. No single data source is available to research and network analysts that can provide information equivalent to device logs and flow data. Consequently, information will need to be gathered from multiple sources to provide enough information to create a useful description of threats.

Using multiple data sets allows analysts to build an overview as a mason bricks a house. Each block itself may be useful for something, but only when many are put together is a wall covered. In the same way, data sources with limited scope have applications on their own, but only when multiple sources are combined do they cover enough aspects of a threat to help analysts focus in-depth research or decision makers begin to respond and recover.

The process of generating a description of a threat without direct access to devices, host-based logs, network packet captures, or network, firewall, and intrusion detection system (IDS) logs require exploring available external data sets that could help answer several questions. For an overview of a threat to be useful, it should help researchers determine appropriate responses, prevent further exploitation or provide guidance for effective in-depth research. Hence, before outlining a methodology to describe a threat, it is important to identify a set of information that answers the basic questions about the threat. We have identified the following five questions that when answered provide details that are of value to analysts and others. Together the answers to these five questions can help researchers and network administrators in all three areas.

1. What organization or group of people is infected or being exploited? In other words, ***identify owners***.
2. Where is the threat infrastructure or exploitations located geographically or physically? In other words, ***identify locations***.
3. How are infected or exploited machines connected to the internet? In other words, ***identify connections***.
4. What types of machines are being targeted? In other words, ***identify devices***.
5. What ports appear important to the threat? In other words, ***identify TCP/UDP ports***.

The first two questions allow analysis to know who is targeted and where. This helps in determining what organizations or people may need to be notified of exploitation. The last three questions give analysts insight into connection types, device identification, and important ports. These can indicate that new firewall and IDS rules may need added or may point to issues with software patching. The information obtained can also show researchers if they should focus on certain devices, operating systems, services, or network traffic using specific ports when doing in-depth analysis.

The proposed methodology is shown in Figure 1 and the rest of this chapter presents it in detail. Section 2.1 discusses choosing appropriate sources. Section 2.2 talks about finding initial information to start the analysis process for a threat. Section 2.3 discusses compilation of a process for data extraction and reporting.

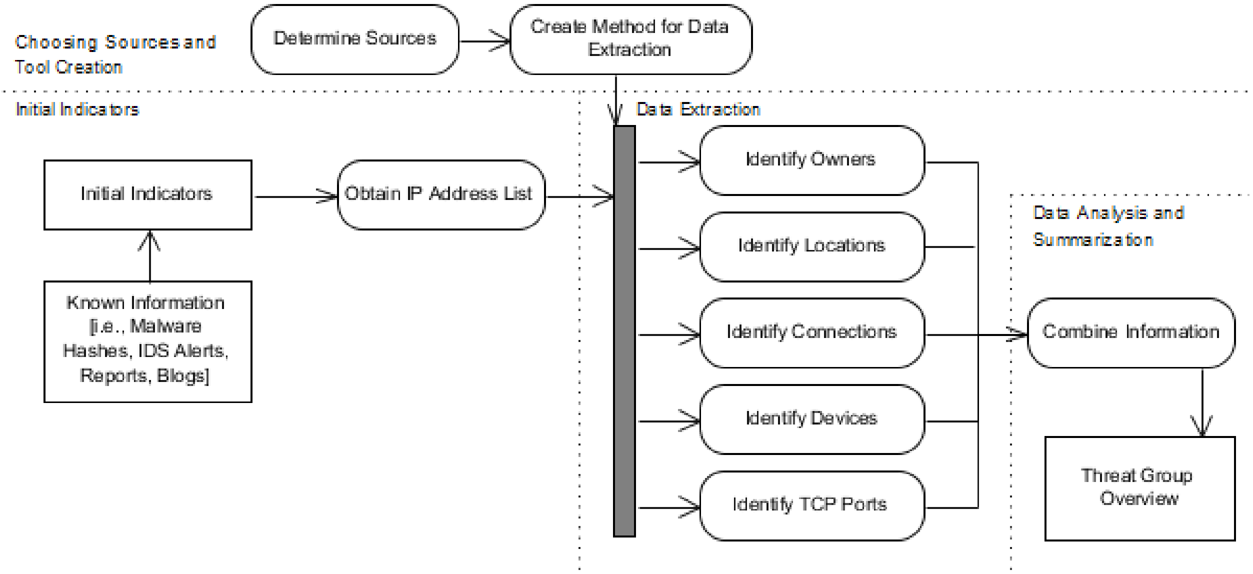

Figure 1. Methodology Overview

2.1 CHOOSING SOURCES

The first requirement in the methodology is to choose sources that can be used to answer the questions identified above. The process for choosing appropriate data sources is shown in Figure 2.

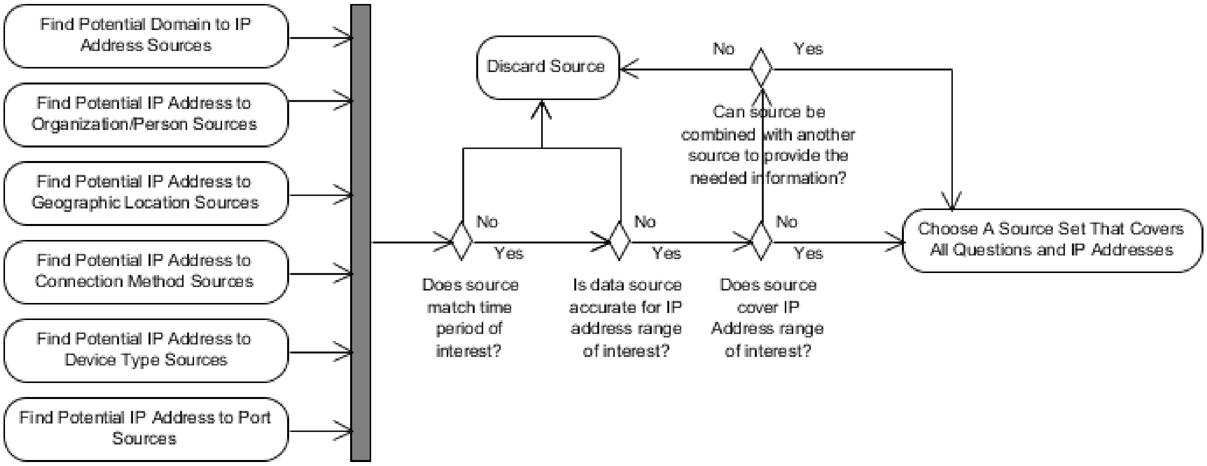
Find Potential Domain to IP Address Sources
Find Potential IP Address to Organization/Person Sources
Find Potential IP Address to Geographic Location Sources
Find Potential IP Address to Connection Method Sources
Find Potential IP Address to Device Type Sources
Find Potential IP Address to Port Sources
Discard Source
No
Yes
Can source be combined with another source to provide the needed information?
No
Yes
No
Yes
No
Yes
Does source match time period of interest?
Is data source accurate for IP address range of interest?
Does source cover IP Address range of interest?
Choose A Source Set That Covers All Questions and IP Addresses

Figure 2. Method to Choose Sources

The actual sources used to answer each question are not as important as the relevance, reliability, and robustness of the information in the sources. A method for generating an overview is useless if the information used is not accurate or if it is not available for the exploited devices. Because the internet is volatile, the method is also futile if the information available does not correspond to the time period of exploitation.

Sources also need to be able to be connected to the information analysts have as a starting point for their analysis. In general, the available information that identifies exploited devices will be a set of domain names, IP addresses, or other information can be transformed into a set of domain names and IP addresses (for instance, sets of related malware hashes). It was observed when exploring data sets that most network-related sources are keyed to IP addresses, so transforming the initial information into a list of IP addresses of interest provides the most flexibility. Note that available sources may answer multiple questions and some questions may need multiple sources. Also be aware that sources may need to differ when the IP addresses of interest are all internal to the analyst's organization versus when some or all are external to the analyst's organization.

Identify Owner. To identify the owners of the targeted or exploited systems, selected sources should be one or more data sets that allow analysts to connect an IP address to an owner. The granularity of ownership that can be obtained will depend on several factors, including whether the IP addresses are external or internal to the analyst's organization. At a minimum, external addresses can be tied to the organization that owns the autonomous system number (ASN) that allocates it by using data sets with whois type information. This information is not very fine grained, but is still valuable. For some IP address sets it may be possible to get a finer level of attribution for ownership with other types of data sets or by doing some manual research. On the other hand, when evaluating addresses from inside the analyst's local organization, it should be possible to link a device to an individual person or department.

Identify Location. Selected sources to identify physical location of targeted or exploited systems should provide some form of geo-location information for IP addresses. For addresses external to the analysts local organization, this means data sets that give country, state (or province), and/or city locations. It is important to understand that geographically locating IP addresses is not an easy matter, especially when they are tied to mobile devices, so accuracy between data sets (and between granularities, or even locations, within data sets) can vary widely. For addresses internal to the local organization, the data set would be an association of IP addresses and physical location, such as office, branch, department, or rack.

Identify Connections. Sources to identify how devices connect to the internet should provide some indication of the type of connections devices have to the internet. Data sets could identify connections by media type such as dial-up, cable modem, or satellite or could identify connections by routing type, such as proxy or mobile gateway.

Identify Devices. Sources to determine what type of devices or operating systems are targeted or exploited should provide some sort of description of the devices that are tied to an IP address. The information could tell the type of physical device, like a router, phone, or personal computer or could tell the operating system that is running on the device. For devices internal to the local organization, the data sets could provide more detailed information, such as types of software installed or when the device was last updated.

Identify Ports. Sources to identify important ports look for the TCP or UDP ports that are open on a device or suspiciously used by a device. For most devices, ports should be filtered by a firewall or completely closed. Data sets that can tell what ports are not filtered or closed on a device may provide indication of ports that are being used for malware, or what vulnerable applications are running on the exploited machines. Data sets that show ports a device of interest connects to can also give an idea of ports used by the malware.

2.2 FINDING INITIAL INDICATORS

Before analysts can answer any questions about a threat, they have to start with some indicators that are linked to the threat of interest. Indicators come in many forms and from various sources. Indicators may be IP addresses, domain names, malware hashes, lists of file names, etc. To be useful in this method, the indicators to be used for a threat analysis should be able to be converted into IP addresses. For domain names, the conversion is accomplished by doing reverse domain name server (rDNS) lookups, preferably with a data set that contains historical information. It may be possible to convert other indicators to IP addresses as well. For instance, there are research malware catalogs that can be used to make the needed associations with malware hashes as input.

Analysts should use restraint when converting any indicator to IP addresses. It may be tempting to do several rounds of indicator expansion to build a list of IP addresses, but iteratively building a list can cause the list to quickly become irrelevant to the threat of interest. For instance, say an analyst starts with a list of malware hashes, then gets all associated domain names. The analyst should not then repeat the process by taking those domain names and finding all malware hashes that have been associated with them and then using those hashes to find more domain names. After even just one or two rounds, a process like that would lead to a set of indicators that at best an analyst could say were exploited, not that they were exploited by the threat of interest.

Some sources of indicators are malware databases, blogs, forums, and reports by security companies. Indicators could also come from incidents occurring at the analyst’s local organization. Obtaining indicators from an incident may require some preliminary analysis or could mean simply calculating a hash for a piece of malware or taking a domain name and finding others that are related from a database or web search. In the rest of this thesis, these indicators are considered to represent exploited devices. It is important for analysts to understand, though, that there is always the possibility that the device is not actually exploited or that the device has been exploited, but not by the threat of interest. Using this process to identify both instances is mentioned as a possible future work in chapter 5.2. Once analysts have a set of indicators and have determined their sources, they can begin the actual process of generating an overview.

2.3 DATA EXTRACTION AND ANALYSIS

The actual method for creating an overview of a threat has three segments:

1. Tool creation or selection
2. Data extraction
3. Data analysis and summarization

The first segment takes the most time and effort, but should result in a reusable process that can be used for multiple studies.

Tool Creation or Selection. It is unlikely that any data set will only contain information of interest for one particular study. Especially for sources containing information external to an analyst's local organization, it can be infeasible and is inefficient to manually search for the relevant information. If they do not already exist for the chosen data sources, tools should be created for each data set that allows analysts to easily extract and aggregate all relevant information for a set of initial information. The tools should also be created in a manner that allows the extraction to be added to an automated process. The tools will be dependent both on the chosen sources and the hardware and software that analysts can access. Some sources may already come with an effective technique for data extraction in which case analysts may need to do nothing at all or just find a way to tie the existing it into an automated process. Others data sources may require development of programs or scripts to extract the data relevant to a threat.

Data Extraction. Once tools are created for each data set, they should be combined into an automated process that takes as input a set of IP addresses and, if supported by the chosen sources, a date range of interest. This automated process can then be applied each time an analyst needs to evaluate a threat. The analyst will first take the known information and transform it into a set of IP addresses. Then that set should be provided as input to the automated process. The process will output the following information for each IP address in the set:

- Identity of owner
- Geographic or physical location
- Network connection information
- Type of device
- Associated active TCP ports

Data Analysis and Summarization. The output information should be combined and summarized, either as part of the automated process or manually. Analysts should then evaluate the results and create the actual overview of the threat. Depending on context, this overview can then be used to:

- draft first steps in incident response, both cleanup and prevention of further exploitation.
- inform who should be notified of exploitation.
- filter log and network flow data for in-depth analysis.
- analyze the extent of compromised devices.

- create a data set containing real data for theoretical security research.

All of which are possible with the methodology without having to contact an exploited organization or other entity.

3.0 EXAMPLE IMPLEMENTATION

This chapter provides one instance of an implementation of the methodology. Section 3.1 describes the sources used for the section 4.0 case studies and why they were selected. Section 3.2 explains the tools created and how they were combined into an automated process.

3.1 SOURCES USED

Several weeks were spent searching for data sets that could be related on IP addresses and that would help answer the five questions identified in chapter 2.0 . Eventually eight sources were identified that provide a meaningful description by answering the questions for a large portion of the internet, over the timeframes required for the case studies in chapter 4.0 . Sources used for analysis in this study were chosen based on several criteria:

- Historical versions of the contents are available.
- Contents are considered reasonably accurate by researchers or corporations, as shown by their use in other research publications or organizations acknowledging their use.
- Sources are available to others in the research community, either through open source access, purchase, or registration.
- Sources are not confidential to an organization or classified by a government.

- Formats and content lend to process automation.

The eight sources are summarized in Table 1.

Table 1. Example Implementation Sources

Data Source	Use	Availability
Security Information Exchange	Associating domain names with IP addresses	Subscription for analysts working in the public interest
University of Oregon's Route Views Project	Associating IP addresses with autonomous system numbers	Download from link on the Route View Project website
RIPE Network Coordination Centre Routing Data	Associating IP addresses with autonomous system numbers	Download from the RIPE Routing Information Service Raw Data webpage
Potaroo.net Autonomous System Number-to-name Mapping	Associating autonomous system numbers with their owners	Download from the CIDR Reports webpage in the Potaroo.net website
MaxMind GeoLite	Associating IP addresses with geographic locations	The MaxMind GeoLite Free Databases web page
Neustar GeoPoint 7	Associating IP addresses with geographic locations and internet connections	Purchase from Neustar
Internet Census 2012	Associating IP addresses with device types and open ports on the devices associated with the IP addresses	Download from the Internet Census 2012 Bitbucket site
Internet Store Center's All Sources IPs List	Associating IP addresses with	Download from the Internet Storm Center's XML webpage

In some cases, more than one of these data sets provided the same type of information, but were used together to increase accuracy in drawing conclusions. The first instance of this is ASN organization attribution. This was primarily derived from the Oregon Route Views project and Potaroo.net, but then compared with the ASN information from MaxMind to corroborate the results. The second case is geo-location. Two geo-location sources were utilized to get an idea of the consistency of results.

In the remainder of this section each data source used for analysis is described, along with some issues and considerations pertaining to each source. The first three sources, SIE, University of Oregon Route Views Project, and Potaroo.net, are used to identify exploited IP addresses and organizations, which relates to question one, "what organization or group of people is infected or being exploited?" The MaxMind GeoLite and Neustar GeoPoint 7 data sets are used for geo-location of infrastructure, providing the answer to question two, "where is the threat infrastructure or exploitations located geographically or physically?" Neustar GeoPoint 7 also answers question three, "how are infected or exploited machines connected to the internet?" The Internet Census 2012 answers question four, "What types of machines are being targeted?" Finally, the Internet Census 2012 and Internet Storm Center's D Shield collections are used to answer question five, "what ports appear important to the threat?"

With the exception of the Internet Census 2012, all the data sets had versions available for a range of dates. Since these versions do not vary in format, the method of data extraction worked with whatever version of each data set was closest to the time period of interest for a threat. Since the Internet Census 2012 has only one version, this implementation of the method should only be used for time periods of interest during the months when the internet census was conducted or at most a few months before or after.

When choosing these data sources, there were several ethical considerations, specifically in regard to using the Internet Census 2012. Several researchers have questioned the ethicality of using a data set obtained from a botnet that did network scans. See Appendix A: Ethics and the Internet Census 2012 for a discussion.

3.1.1 Security Information Exchange

The Security Information Exchange (SIE) is a collection of network related information for use by researchers and businesses working in the public interest. The collection includes passive domain name services (pDNS), malware, and spam information among other things. This information comes from sensors on participating networks, from places such as universities, internet service providers (ISPs), security companies, and other businesses. As a centralized resource, the SIE allows sharing of sensitive data between disparate entities. [17]

In this sample implementation the SIE contains the data sets used to take the available initial device information and convert it into sets of IP addresses. This was done with the malware and pDNS portions of the SIE database. The malware data set provided correlation of malware hashes to domain names. The pDNS portion was used to associate known domain names, or those obtained from the malware data set, to IP addresses. These IP addresses and the devices attached to them are considered the infrastructure that is being analyzed in the rest of the process.

While the SIE contains much information from many sources, it does not provide a picture of the whole internet. As the SIE data collection is voluntarily provided by individuals and organizations, information that does not traverse the networks of participating individuals and organization are not available for research. The information comes from a wide variety of sources in large volume, making it reasonable to assume the available data is representative of internet traffic. Still, when using this data set, it is important to remember that the information obtained is just a subset and not a comprehensive set of all the activities that have occurred.

3.1.2 University of Oregon's Route Views Project

The University of Oregon's Route Views Project comes from the Advanced Network Technology Center at the University of Oregon. Route Views is a collection of real-time BGP routing information from the view of multiple locations across the internet [18]. Participating organizations provide the project with their BGP routing information, which shows the paths that network traffic can use to travel through autonomous systems to get to its destination IP addresses. Since the paths show what ASN advertises an IP address block, the routing information can be used to determine which IP addresses belong to what ASNs.

This study used the BGP path files from all of the organizations participating in the Route Views project to associate IP addresses with their ASN. The path files are available at http://archive.routeviews.org. These files contain a large portion of the routes that occur on the internet, but not all of them. Combining these routes with routing information from the RIPE NCC provides results that permit attribution of most IP addresses to an ASN.

3.1.3 RIPE Network Coordination Centre Routing Data

RIPE NCC is the European internet registry. In addition, RIPE has a routing information service that collects BGP routing data from multiple locations [19]. Like the Route Views Project, RIPE NCC provides path files, which can be obtained from http://www.ripe.net/data-tools/stats/ris/ris-raw-data.

Both Route Views and RIPE's routing data have the possibility of containing advertised paths that are not from the actual autonomous system owner. This may result from misconfiguration of infrastructure, typographical errors, or deliberate attempts by malicious actors to corrupt routing. While it is not possible to identify all corrupt routes, those routes that blatantly cause red flags are filtered, as discussed in section 3.2.1.2.

3.1.4 Potaroo.net

Potaroo.net is a site ran by Geoff Huston, providing information related to current BGP advertisements, autonomous system numbers (ASNs), and the IPv4 and IPv6 address space. Besides various reports related to address allocation, route updates, and ASN assignment, many updated daily, there is a document that provides the description for registered ASNs, labeled Autonomous System number-to-name mapping. In most cases, the description is the name of the organization owning the autonomous system number. This document has a consistent format and contains information from each of the internet registries making it a convenient choice for determining the organization associated with an ASN. If this document is downloaded regularly, it is possible to create a historical view of the data.

When used with the Oregon Route Views routing information, an association of an IP address to the organization owning its assigned ASN can be built. In this implementation, ASN ownership was used to determine to whom an IP address belonged. In most cases this works well, but for the instances where ASN description does not contain the owning organization's name, it is necessary to find the organization associated with the IP address using a whois look up.

Without access to an archive of the ASN documents, attribution of IP addresses to an organization in the past is not as accurate as looking from data generated during the time period of interest. However, ASN to organization assignment does not change as often as other internet related information, so results would still be informative using the current version of the document. If it is suspected that an ASN assignment has changed, further research can be done on the ASN number to determine its likely assigned organization during the time of interest.

3.1.5 MaxMind GeoLite

MaxMind is a company providing geo-location services for IP addresses. In addition to several databases available for a licensing fee, MaxMind provides a series of free data sets in their GeoLite series. These provide geo-location at the country or city level as comma separated files, and are updated monthly. MaxMind also has a GeoLite database that associates IP address ranges to ASNs which is also updated monthly. For this implementation, information was used from all three data sets.

MaxMind does not detail the process used for determining the geo-location information related to an IP address, nor is there accuracy information for country level attribution. The provided city level accuracy rates vary widely, from 30% accurate attributions in the Ukraine to 95% in Cote D’Ivoire [20]. The stated accuracy rate for the United States city level is 78% correctly identified [20]. It is interesting to note that many IP addresses do not have information other than country information, including information on the ASNs associated with the addresses, which is why ASN organization was determined primarily using Oregon Route Views and Potaroo.net data with MaxMind just providing corroboration when the information was available.

3.1.6 Neustar GeoPoint 7

Neustar is a company providing customer intelligence services, including geo-location of IP addresses. The Neustar IP Intelligence web service is available at four levels of detail. The “Where & How” level package provides geo-location and connection information. The “Where, How & What” package also provides ASN and organization information. For a licensing fee files may be available. For this study, a file with the “Where, How & What” level data was used, though the information used can be obtained from “Where & How.”

Neustar describes their process for providing accurate data as collecting data from the source with insights from global partners, validation by analysts, and feedback from user with their GeoFeedback tool [21]. Neustar is able to determine connection information for many IP addresses and they claim that the geo-location accuracy of fixed connections, like DSL and fiber optic cable, are very accurate [22]. Furthermore, for some IP addresses they provide a confidence factor for some of the location fields. This represents the Neustar analysts’ belief of the likelihood that the user of an IP address is at the stated location. Otherwise, Neustar does not provide rates of accuracy. Where IP address location differs between Neustar and MaxMind, further research is necessary to determine the appropriate location.

3.1.7 Internet Census 2012

The Internet Census 2012 is a collection of information from an anonymous scanner of the IPv4 addresses space. A researcher created and deployed a botnet on unprotected devices that used a modified nmap program to map out the internet between March and December of 2012. In early 2013 the resulting data set was posted to bitbucket.org. This data set contains results from nmap

ICMP ping, host probe, sync scan, service probe, and TCP/IP fingerprint tests, as well as some traceroute data and reverse DNS lookup results from 16 of the largest DNS servers. [23]

In this study, the TCP/IP fingerprint and sync scan tests were of interest. The TCP/IP fingerprint data, when compared against the nmap fingerprint database file, gives likely identity of devices, either an actual device like Linksys WRT610Nv3 WAP or an operating system like Microsoft Windows XP SP1. The sync scan data associates IP address with open ports. The other data in the Internet Census 2012 may be helpful in other research, but does not provide relevant information for describing the infrastructure of threats.

The Internet Census 2012 information is relevant, but quickly becomes stale. As the devices tied to IP addresses are frequently updated or patched and IP addresses often are dynamically assigned or become assigned to new devices, the information in this data set is most relevant and accurate for examining infrastructure that was active when the scans were occurring. It is also important to realize that the scans used to obtain this data set were done across the internet. This means there may be some interference of firewalls or proxy servers, which may skew the results [24]. Before using the results in descriptions of infrastructure supporting a threat, it is necessary to evaluate if the information indicates that the results are from the actual devices of interest or reflect protecting devices between the devices of interest and the internet.

3.1.8 Internet Storm Center's All Sources IPs List

The Internet Storm Center (ISC) is an organization with the purpose of detecting and analyzing threats as they arise and then disseminating information on how to respond to those threats to the general internet user population. This volunteer organization is supported by the SANS Institute and collects data around the world from organizations using their DShield intrusion detection system (IDS). Information gathered from the DShield IDS is available on the ISC web site. Available information is concerned with IP addresses generating intrusion detection responses, ports targeted when those connections are blocked, and suspicious domains. [25]

The All Source IPs list available under the XML section of the ISC website is useful for getting an idea of ports used in attempted intrusions. This file is comprised of IP addresses that generated IDS alerts along with the port where each IP address attempted to connect, the protocol used, and information related to the how many times the IP address/port/protocol combination was seen and when. The tab delimited text file is updated daily. [25]

Malicious acting IP addresses are seldom continuously active, so when using this data source it is helpful to have an archive of the information over the whole period of time of suspected activity. Also, like the SIE data, the Internet Storm Center relies on voluntary participation in submission of data. So again, this data set is just a subset of internet activity, which means IP addresses not occurring in the files is not an indication the addresses are not active, just that they are not seen or detected by the ISC participating organizations.

3.2 METHOD AS APPLIED

As outlined above, the method to describe threats consists of three parts. The first part is creating tools to extract relevant information from each of the data sets used in analysis and a process to use them. The second is to apply the resulting processes to obtain information related to an individual threat. The third step is to analyze and summarize the resulting information. This section first discusses the tools and process created for data extraction. It then outlines the steps to utilize them in generating a report and finishes with a discussion of one possible method for reporting on study results.

3.2.1 Tools Creation and Automation

The data sets used to provide information relating to threats and IP addresses contain large amounts of data, most of which is irrelevant in any one study. Creating tools to extract needed information is an important step when the size of data sets makes it impractical to manually search for information. After tool creation, an automated process can be created to invoke the tools with the identifier (i.e. IP addresses or domain names) of the devices of interest. The tools should be designed with flexibility so they are reusable for other studies using the same source data sets. The results of this endeavor are tools that can be used for multiple studies. The time spent developing reasonable tools and the automated process invoking them was the most time consuming portion of this implementation, but worth the effort.

A method for extracting relevant data was created for each of the data sets, with the exception of the SIE. The tools described here are Python and Bash scripts, though there are other ways to efficiently retrieve the desired data. See Appendix B: Alternative Associative Method Using SiLK for an example.

3.2.1.1 Security Information Exchange.

The Security Information Exchange data resides in a database. The database is queried to return the IP addresses associated with domain names of interest, retrieving the information from stored passive DNS records. For instances where there is a timeframe of interest for the domain names, such as when working with data where in-depth analysis has already been done, the queries are limited to data from the time period of interest.

3.2.1.2 Route Views, RIPE routing, and Potaroo.net.

The University of Oregon Route Views, RIPE routing, and Potaroo.net data were combined to produce a new data set. The first step in combining the sources is to obtain all the available routing files for a specific date from the Route Views archive and RIPE Raw Data servers. The routing files consist of entries of IP address netblocks in CIDR notation followed by a list of autonomous system numbers. The list is one path that network traffic can follow to get to the IP address block. The last ASN in each path list is called the advertiser as that is the ASN publishing the path and is supposed to be the address block owner.

In generating the new data set the route files are used as input to a script that outputs a combined file of all the routes, discarding duplicates and unreasonable advertisements. For this tool, unreasonable paths are considered to be those that fall into one of two categories. First are routing paths that are advertising netblocks larger than a /8 in CIDR notation. These are considered unreasonable as no organization has been assigned consecutive /8 netblocks [26]. The second category is unsupported route paths. Normally there will be multiple ways to traverse the internet to get to any particular IP address. This is reflected in routing tables as multiple entries of paths to the advertisers of a netblock. Using a convention suggested by an experienced analyst, when there is only one path telling how to get to the advertiser of a netblock, it is considered an anomaly and assumed to be a bad advertisement.

After a combined file is generated the paths are parsed to result in a list of netblock/advertiser pairs. In most instances, netblocks are only associated with one advertiser in the results. When a netblock is associated with more than one advertiser, this suggests some form of relationship between the advertisers. For instance, internet service providers may advertise IP addresses through multiple ASNs. As these advertisements are often legitimate they are not removed from the data set.

The last step in combining the sources is to take the cidr-report.org report “Autonomous System number-to-name mapping” [27] and combine it with the netblock/advertiser pairs. The final result is a file that associates netblocks to autonomous system numbers and names.

3.2.1.3 MaxMind GeoLite.

There are four MaxMind GeoLite comma separated files used for geo-location and ASN association of IP addresses. The country and ASN information are each in a single file, while the city information is in two files. The first city file associates range of IP addresses with a block, which is a key field used in the second file. The second file associates the blocks with location information. These files are parsed with a Python script, which takes as input a list of target IP addresses and output a file of IP address, location, and ASN tuples.

3.2.1.4 Neustar GeoPoint 7.

All the Neustar GeoPoint data exists in a single compressed, comma separated file. The file is extracted with gzip and then parsed with a Python script. The script takes as input a list of target IP addresses and outputs a file with IP address, location, and connection information tuples.

3.2.1.5 Internet Census 2012 Fingerprints.

The Internet Census 2012 TCP/IP fingerprint files are a set of tab delimited text files. Each file contains IP addresses for one /8 netblock. The file name for each file is the netblock number with no extension, so the file containing addresses from 9/8 is named “9”. Data is extracted from these files in a two-step process. First, a bash script takes as input a list of target IP addresses and outputs a file with the lines that contain matches to the IP addresses from the IC 2012 fingerprint files. Second, a script replicating the nmap fingerprint algorithm takes that file as input and outputs a file with IP addresses and matching identities.

3.2.1.6 Internet Census 2012 Sync Scan.

The Internet Census 2012 sync scan files are a set of tab delimited text files. Like the IC 2012 fingerprint files, each file contains IP addresses for one /8 netblock and again the file name for each file is the netblock number with no extension. Data is extracted from these files with a bash script and then further parsed with Python scripts. One Python script takes as input a list of target IP addresses and produces an output file consisting of IP addresses associated with a list of ports. Ports that are marked as filtered or open/filtered are removed as this indicates the ports are protected by a firewall. For cases where it is helpful to have one port per line, for instance for generating counts and charts, a second Python script is used. It takes in the output from the first script and creates a file where each line is an IP address with a single port, meaning each IP address may occur in multiple entries.

3.2.1.7 Internet Storm Center All Source IPs.

The Internet Storm Center All Source IPs list is a tab separated text file. After obtaining the file, it is saved in ANSI format as the downloaded file format does not work well with Python. Parsing the file with a Python script that takes as input a list of target IP addresses produces an output file consisting of IP addresses associated with a list of ports. As for the IC 2012 sync scan results above, when it is helpful to have one port per line a second Python script is used to output a file where each line is an IP address with a single port.

3.2.2 Data Extraction

The tools above are described in a logical order of execution. As a recap, the ordered steps used in this study to extract data from which to build a description of a threat are as follows.

1. *Determine IP addresses for the threat of interest.*

The first step is to obtain a list of IP addresses associated with a threat. One way this can be accomplished is by direct attribution of IP addresses or domains to a threat by other analysis, for instance from released lists of initial indicators. This method was used in the first case study, where a list of domains was available. A second way to obtain IP addresses is to determine IP addresses that have been observed to originate traffic containing malware from the threat. This method was used in the second case study. When the available information on devices of interest is domain names, IP addresses are obtained with the process for using the SIE data.

2. *Determine organization responsible for the IP addresses.*

The second step is to associate the IP addresses of interest with the organizations that owns the addresses or are responsible for allocating them to others. This is done by executing the Route Views, RIPE routing, and Potaroo.net process.

3. *Determine geo-location and connection data.*

The third step is to determine location and connection information. This is done by running the MaxMind GeoLite and Neustar GeoPoint 7 scripts.

4. *Determine identities for devices associated with the IP addresses.*

The fourth step is to extract information that tells about the devices using the IP addresses and is accomplished with the Internet Census 2012 fingerprint scripts.

5. *Determine active ports.*

The fifth step is to find ports that are used by or on the devices. The Internet Census 2012 sync scan tool outputs ports that are open or listening on the devices of interest. Internet Storm Center tool outputs ports that are being connected to from the devices of interest.

These steps can be scripted into a bash file for automated execution.

3.2.3 Data Analysis and Summarization

The steps above result in four to six files associating IP addresses to the corresponding information from the data sources. These files can be concatenated into a comma separated file for import into Microsoft Excel or some other reporting software, which allows for easy creation of charts and graphs. For a more automated process, several scripts are used to provide summary information of the various components, as described below.

The script that summarizes information regarding autonomous system owners outputs counts and percentages of occurrence for each autonomous system. This includes statistics for the different autonomous system numbers as well as the owners of those ASNs.

Geo-location is summarized by a script that takes in both the MaxMind and Neustar data. This script compares the country locations and if they are different will report on both. If the Neustar data set has state and/or city information it is used even if MaxMind also has the information. If Neustar does not provide the state or city information, the MaxMind information will be used if available. Neustar is preferred over MaxMind for the state and city levels as it has more frequent state and city attribution and the entries are the actual state and city names. At the state level, MaxMind uses state codes. The output of this script tells the frequency of occurrences for the different countries, states, and cities.

The connection information report script outputs counts and percentages for the connection information. This includes statistics for the media type and connection method individually as well as how they occur together.

Identities are overviewed by a script that outputs counts and percentages for the likely identities that occur in the data set. When the identity is the case where two or three possibilities are equally probable, that whole set of possibilities is treated as an individual identity.

Port information is summarized by a script that calculates statistics for individual IP addresses as well as across the data set. Statistics include average open ports per IP address, total count of unique ports, and the number of IP addresses where each port is associated.

4.0 CASE STUDIES

The two case studies presented in this chapter are meant to illustrate the usefulness of the methodology described earlier. They present real information from real threats, but should be looked at with a view toward the type of information that can be obtained as opposed to an actual analysis of the threats. What is presented for the threats described is not meant to be analysis showing everything that can be learned or inferred from the output information.

The first case study applies the method to Advanced Persistent Threat 1. This case study illustrates how the methodology can be used to validate the output of more traditional, in-depth analysis that has already been reported. It also illustrates how the methodology can be used to determine information about a set of exploitations by one threat that is a persistent people group and not just one piece of traditional malware or malware family.

In contrast to a people group threat, the second case study applies the method to a set of related malware, namely, the *Mabeza Infected* family. The exploits associated with this threat may or may not originate from the same person or set of people, but are all variations on the same piece of malware. This case study also illustrates how the methodology can work when nothing else is known about a threat other than machines exploited by it.

4.1 CASE STUDY 1: ADVANCED PERSISTENT THREAT 1

Advanced Persistent Threat 1 (APT 1) is purported to be a Chinese cyber-espionage unit of the People's Liberation Army of China, operating out of Shanghai [28]. Details of this group's operations, along with indicators of activity were released by the information security company Mandiant in their report "APT1: Exposing One of China's Cyber Espionage Units". The indicators of activity included a list of domain names that have been associated with APT 1. These are used as the initial input for this case study, as they are the most accessible identifiers of exploited machines and they presumably correspond to the devices analyzed by Mandiant when they carried out their analysis.

This first case study shows how the methodology described above can generalize an in-depth analysis. The results can be compared with the rest of the information in the Mandiant report to gauge the accuracy and see if any other details can be found from this method. First the implementation is reviewed, followed by a discussion of the findings. This section ends with a comparison of the results of this study to the Mandiant report.

4.1.1 Applying the Methodology to APT 1

Following the method outlined above, the first step to gather relevant information is obtaining a list of IP addresses. The initial indicator for this case study were the domain names in the Mandiant report. These were used to query the SIE database, returning a list of 622 unique addresses. These addresses were then used to extract relevant information from each of the other data sources. With the available data, all IP addresses were able to be associated with autonomous system numbers and owners as well as with a country of origin. At least two-thirds of the addresses had one or more pieces of information related to state (or province, as the case may be), connection type, and open ports. Approximately one third of the addresses could be associated with fingerprint identities from the nmap database. Just over 10% of the IP addresses occurred in the ISC data in the year 2012, indicating they were sources of blocked connections to devices protected by firewalls.

4.1.2 APT 1 Results

Using the method described above allows creation of a general picture of the Advanced Persistent Threat 1 infrastructure—the devices used in the various methods of exploitation, such as command and control servers, distribution systems, or possibly even espionage targets. This section discusses each of the components of the overview individually. First presented are the autonomous system related findings, then geo-location information, connection information, identities from fingerprints and finally information on relevant TCP ports. The section closes with an overall discussion of the results.

ASN and Owner Information. There are 205 unique autonomous system numbers associated with the IP addresses in this case study. This includes 10 ASNs that occur in pairs, meaning that both are announcing the corresponding IP addresses and, presumably, both have some ownership of the addresses. Manually checking the owners of these ASNs that occur in pairs supports the presumption. The most frequent ASN occurs 73 times. Only 13 occur ten or more.

Associated with the 205 different autonomous system numbers are 197 owners. Reading through the list shows many well-known companies. For instance, the three owners with the most addresses in the data set are Google, Hurricane Electric, and GoDaddy. Other well-known organizations also occurring are Amazon.com, AT&T, Comcast Cable, and Yahoo!. Most of the autonomous system owners are hosting companies or internet service providers. There are a few instances of organization that are related to education, such as Hong Kong University of Science and Technology and Riverside County Office of Education, as well as one instance related to government, the Florida Department of Management Services Technology Program. There are 145 ASN owners that only have one or two IP addresses in this data set and only 13 that have 10 or more. Figure 3 shows the autonomous system owners having at least five IP addresses in the data set. Related organizations, for instance those with a parent/child relationship, are not combined, as relationships cannot be easily inferred just by looking at ASN numbers or organization names.

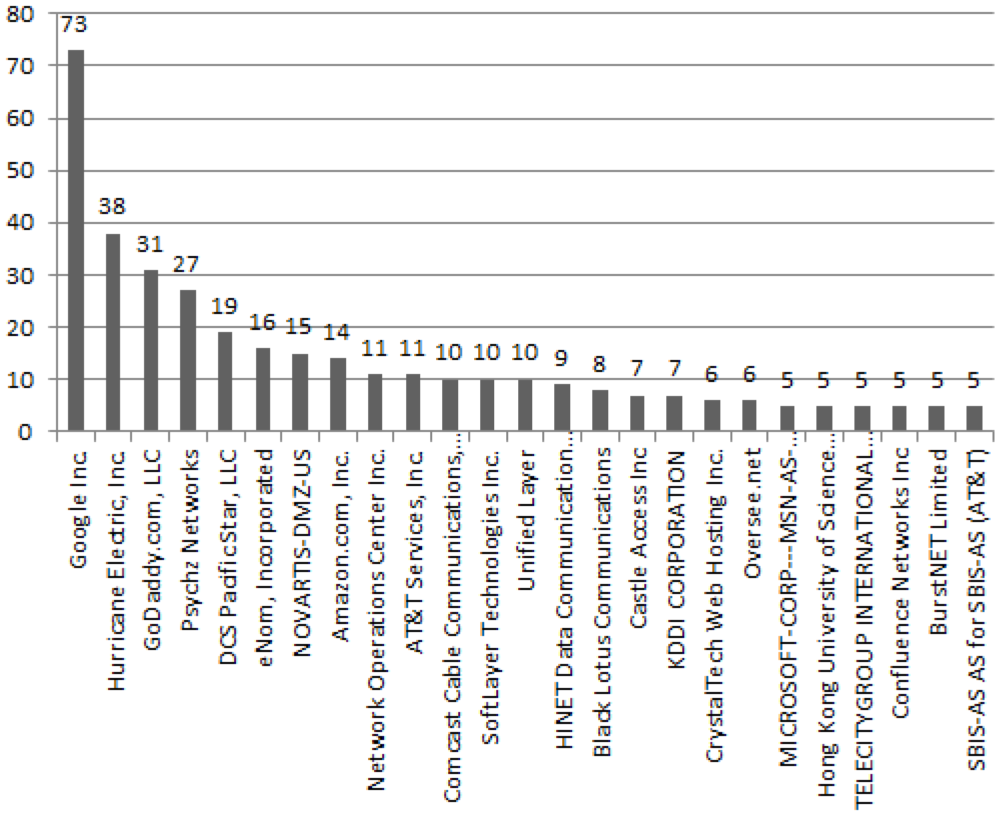

Figure 3. Count of Autonomous System Owners Occurring Five or More Times

The autonomous system owners occurring in this case study suggests questions for further research. Were these IP addresses the end targets of APT 1 or were they just used to enable exploitation of others? Does APT 1 target hosting providers for their infrastructure? With so many organizations only having one or two addresses in the data set, how many other IP addresses belonging to these organizations have also been exploited by APT 1, but are not associated with the specific domain names provided in the Mandiant report? These questions cannot be answered by the method presented here, but the information obtained can guide researchers by providing an initial list of organizations of interest.

Geo-location Information. Examining the geographic location of the IP addresses it is found that there are 59 different countries represented. The locations are heavily skewed to the United States, with 432 which is about 69.4% of all the addresses. The next most frequent country is Great Britain with 15 and Taiwan with 14, both just over 2% of the addresses. All the other countries have less than 10 addresses. Figure 4 shows the number of IP addresses for all non-U.S. countries.

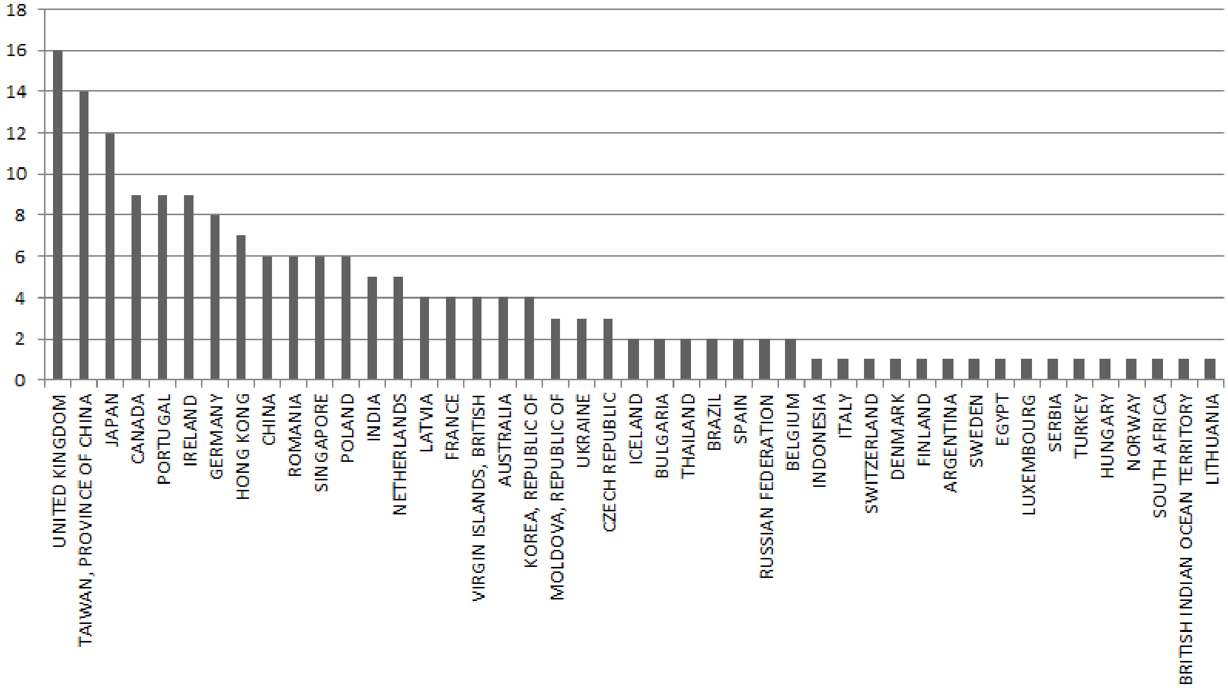

Figure 4. Count of IP Addresses by Non-U.S. Country

Within the U.S., 34 states plus the District of Columbia occur. Almost half of the U.S. addresses resolve to California. Arizona has 44 addresses, Texas has 24, and Virginia has 21. Figure 5 shows the number of IP addresses for each U.S. state.

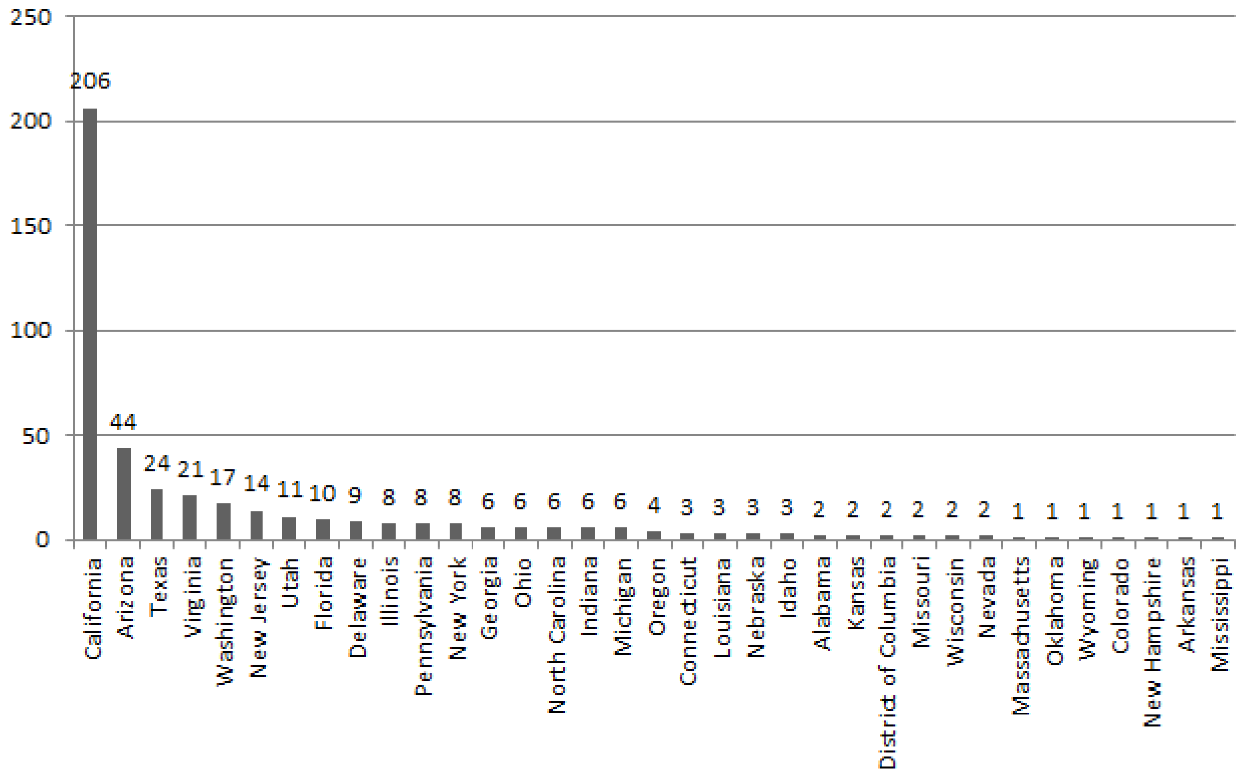

Figure 5. Count of IP Addresses by U.S. State

The concentration of the various locations is interesting in this data set. Such a high occurrence of locations in the United States would seem to indicate that U.S. IP addresses are valuable to APT 1. Especially as no other country has even 3% of the addresses. The concentration within California is even more interesting. Does this indicate that APT 1 was deliberately targeting Californian IP addresses because of the location? Or is this a by-product of the top occurring ASNs being located in California and the ASNs were the target. Or is it a combination of both? Either way, determining why California is important could provide useful insight during further research.

Internet Connection Method. Exploring the internet connection methods of the IP addresses is done by looking at two components of the connection. The first is the connection media and the second is connection routing method.

Connection media is the physical connection type used by an IP address, such as fiber optic cable or satellite. The media types that occur in this data set are leased lines (tx), fiber optic cable (ocx), DSL, cable, dialup, mobile wireless, frame relay, and integrated services digital network (isdn). Leased lines and fiber optic cable are dedicated media, indicating always on connections. According to the Neustar documentation, these connection methods usually indicate IP addresses that belong to medium size or larger organizations, or organizations hosting web content [22]. These are the most common media type with leased lines being associated with about 45% of IP addresses and fiber optic cable associated with around 11% of addresses. No connection information is known for about 32% of the addresses, which means the number of addresses connecting through any of the methods could be higher. Figure 6 shows the count of IP addresses for each connection media.

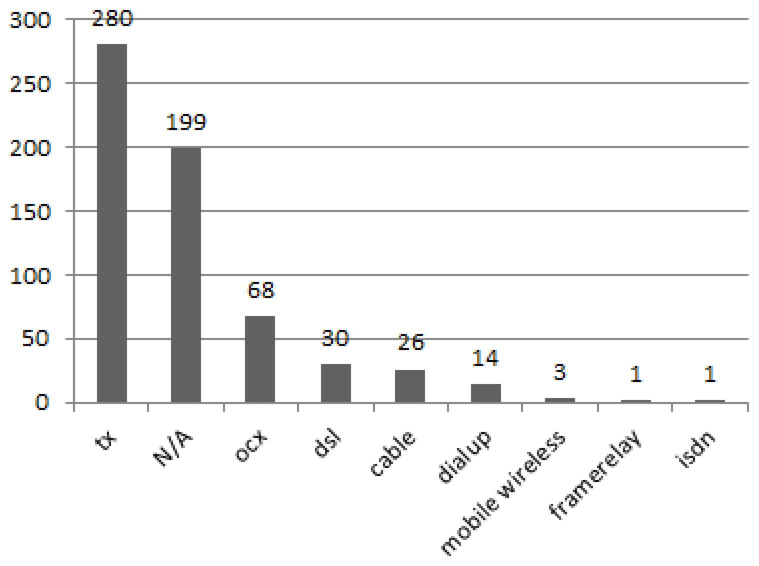

Figure 6. Count of IP Addresses by Media Type

Routing type is how internet connections are routed onto the internet, for instance through a regional proxy or point of presence. The routing types that occur for IP addresses in this data set are fixed, point of presence (pop), AOL network (aol), mobile gateway, international proxy, and regional proxy. The most common routing type is through a fixed connection. Figure 7 shows the count of IP addresses by routing type.

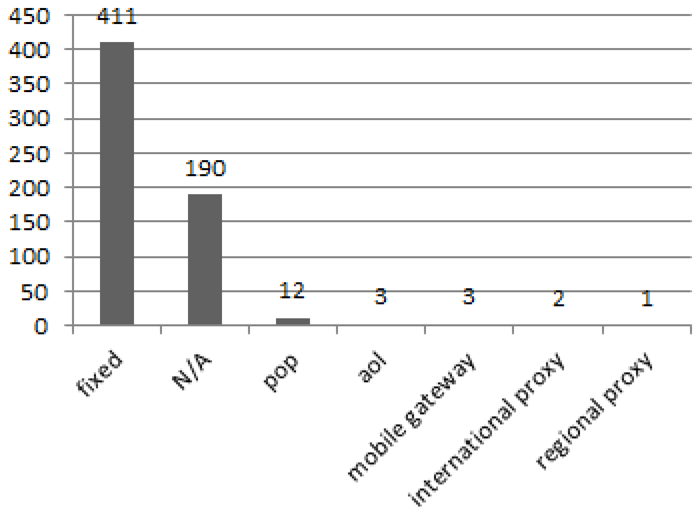

Figure 7. Count of IP Addresses by Routing Type

Combining the connection media and routing type information shows that APT 1 infrastructure is mostly located on fixed connections, with approximately 55% being fixed, dedicated type connections. This seems to indicate that the exploited IP addresses do not belong to home users, even for organizations known for being internet service providers. Figure 8 shows the count of IP addresses when the fixed routing type and dedicated media types are grouped into three categories, one for fixed dedicated, one for fixed non-dedicated, and one for other dedicated. To emphasize the dominance of fixed, dedicated connections, Figure 9 shows the information as proportions.

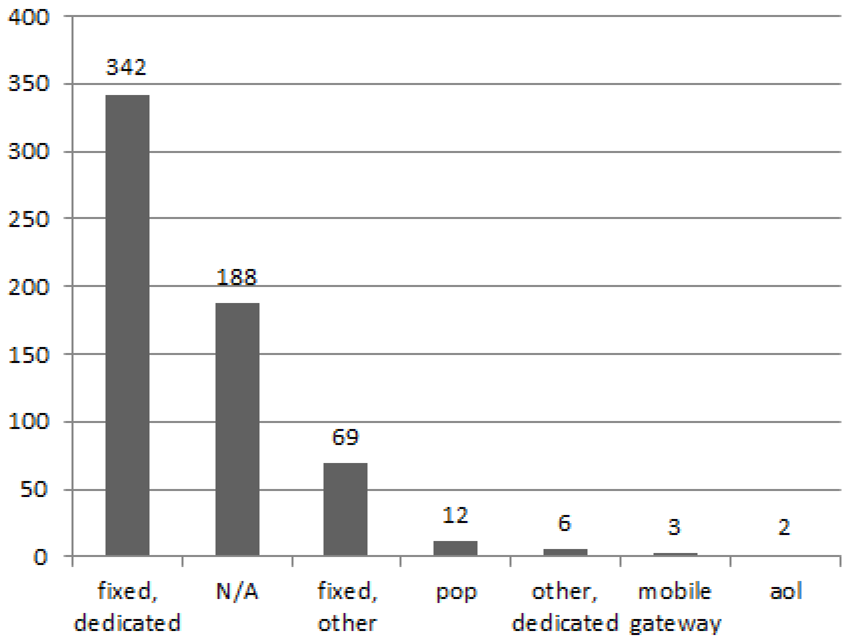

Figure 8. Count of IP Addresses by Media Connection Types

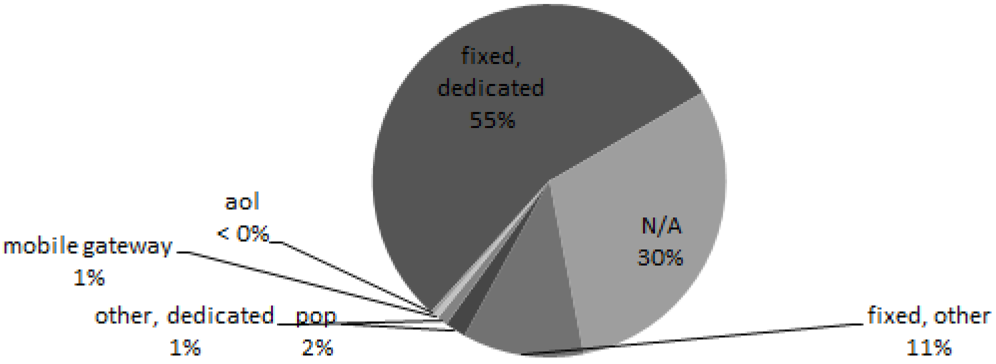

Figure 9. Proportions of IP Addresses by Media Connection Type

Infrastructure with fixed type routing and/or dedicated type media would be beneficial to threats as these indicate that connections are stable and usually continuous. Seeing instances of dialup connections is less expected. As dialup connections are more likely to be disconnected from the internet, it would be interesting for further research to try and determine if there was something special about these IP addresses that made these attractive to APT 1.

Device Identity. Fingerprints were available for 261 IP addresses. Of these 11 had no matching identity in the nmap fingerprint dictionary and 43 had more than three exact matches. Various Microsoft Windows machines and Linux kernels make up most of the remaining IP addresses. Seventy-four machines matched both Windows Server 2003 and Windows XP while 34 additional machines matched only Windows Server 2003. Windows Server 2003 was also one possibility for 21 other devices. Other Windows identities include Windows 2000 and Windows Server 2008. Linux kernels in the range 2.6-3 were the most frequent occurrence for Linux with 40 devices identified as having operating systems somewhere within that range and an additional 19 matching exactly Linux Kernel 2.6. Other identities include F5 Networks embedded, HP embedded, Juniper embedded, Fortinet embedded, and Citrix embedded. Figure 10 shows the count of IP addresses for each identity or set of identities. See Table 2 for details on interpretation of the identities.

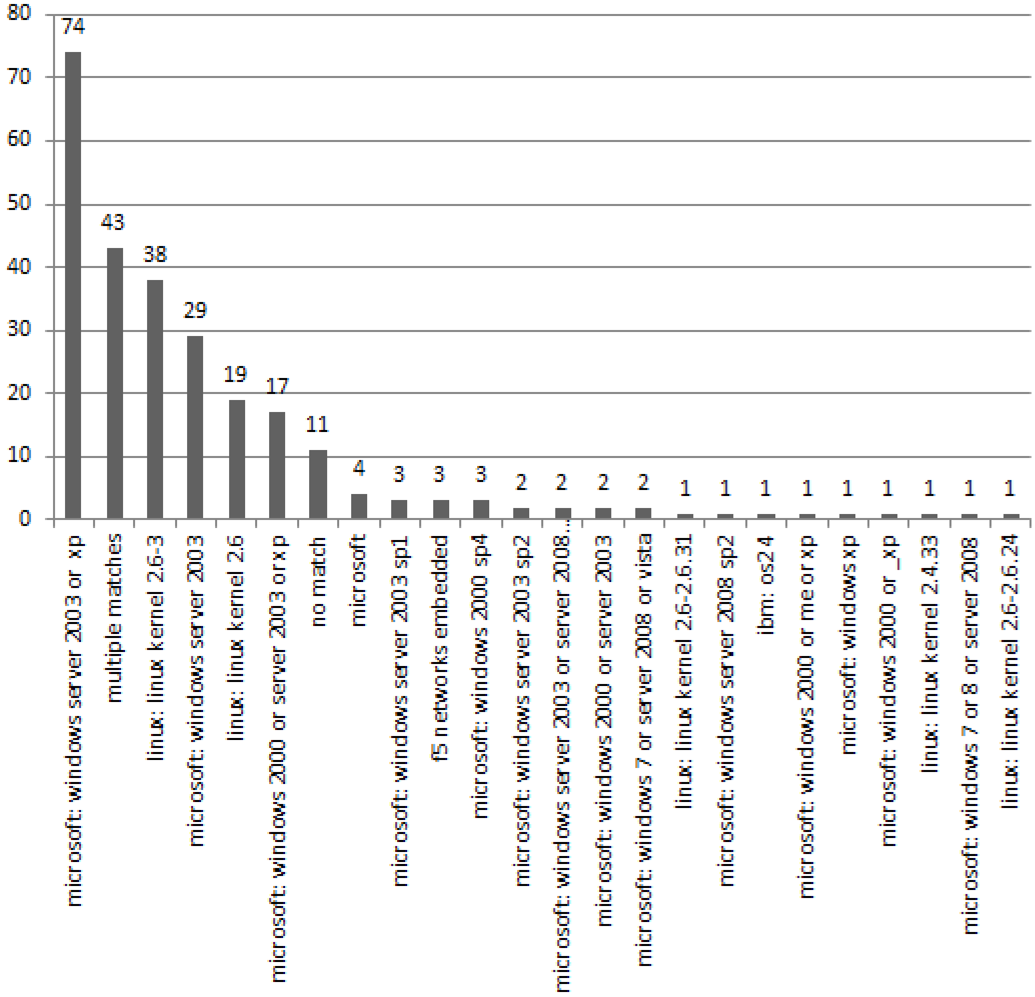

Figure 10. Count of IP Addresses by Identity

Table 2. Parsed Fingerprint Outputs

Format	Meaning
`no_match`	The fingerprint has no reasonable matches in the nmap fingerprint database file.
`multiple_matches`	The fingerprint has more than three possible matches.
`field1`	The algorithm could only resolve the fingerprint to the most generic level, usually Microsoft or Linux.
`field1:field2`	The algorithm resolved the fingerprint to a family of products.
`field1:field2!field3`	The algorithm resolved the fingerprint to a specific verson within a family of products.
`field1:field2a?field2b?field2c`	The algorithm resolved the fingerprint to one of up to three families of products.

From the identity results, it seems that Advanced Persistent Threat 1 was either targeting Microsoft Windows Server 2003 or XP and Linux kernels 2.6-3 or their exploits were most successful against devices with those operating systems. The lists of possible identities of the devices that had more than three matches or that are identified only as Microsoft machines could provide a starting point for further research into the operating systems used. And with insight from the other identities, the most likely options in the list could be explored first.

Important Ports. TCP/IP ports exist as both source and destinations for network traffic. In this study there was no data set that provided insight into source ports. Both the Internet Census 2012 and Internet Storm Center information is regarding destination ports. In the IC 2012, the destination ports belong to the associated IP addresses. They are the ports that are open—meaning the ports are available for incoming connections. In the ISC, the destination ports do not belong to the associated IP addresses, but rather show the ports where some service on the IP addresses are attempting to connect with another device. First the open ports on IP addresses of interest are discussed, followed by the ports where IP addresses of interest were attempting to connect.

Four hundred and ten different addresses had at least one open port, with the average count being five. There were 101 unique open ports, of which 16 ports were open for at least 10% of the addresses. As would be expected, port 80 (HTTP) is the most frequent open port. The second most frequent port is 3389, which is often used for a remote desktop protocol. Figure 11 shows the count of IP addresses by open port from the IC 2012.

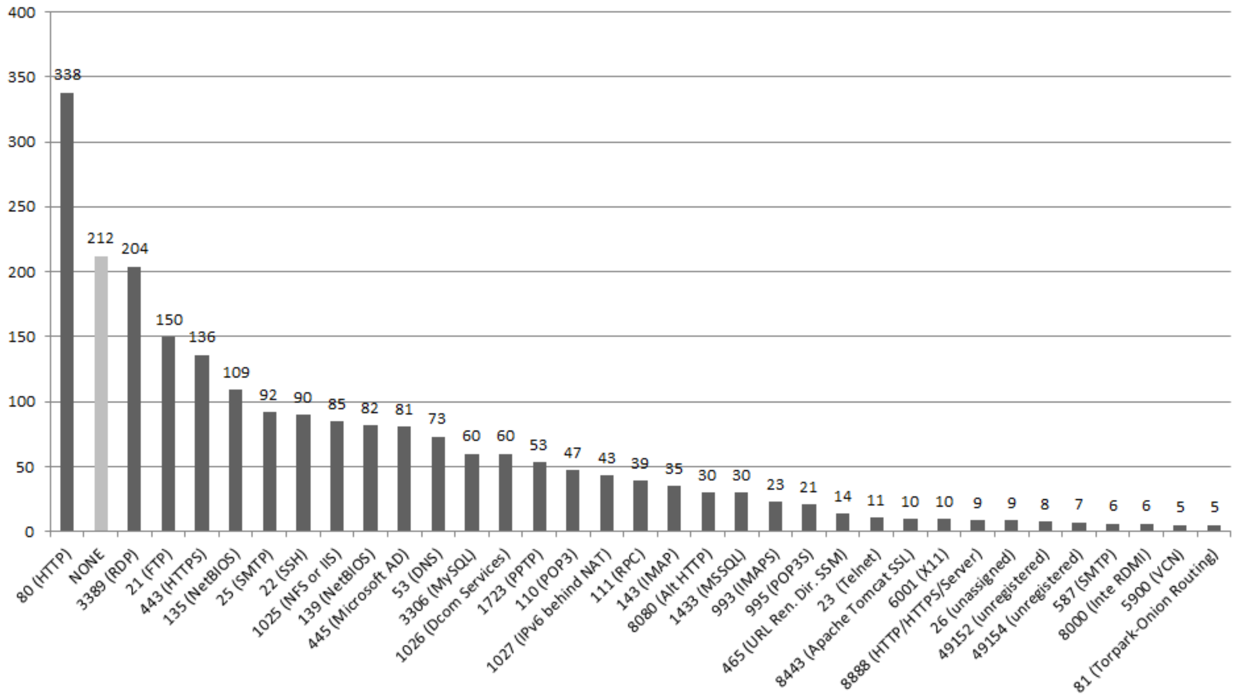

Figure 11. Count of IP Addresses by Open Ports

From the ISC, 65 different IP addresses had blocked connections. Across the whole 2012 calendar year, the average number of connections to different port that were blocked for each address was over 1000. This high number is caused by one IP address that attempted connections over thousands of ports. After removing that IP address from the statistics, the average number of connections to different ports changes to 85. Again, port 80 is the most frequently occurring port. Figure 12 shows the count of IP addresses by blocked connection ports from the ISC.

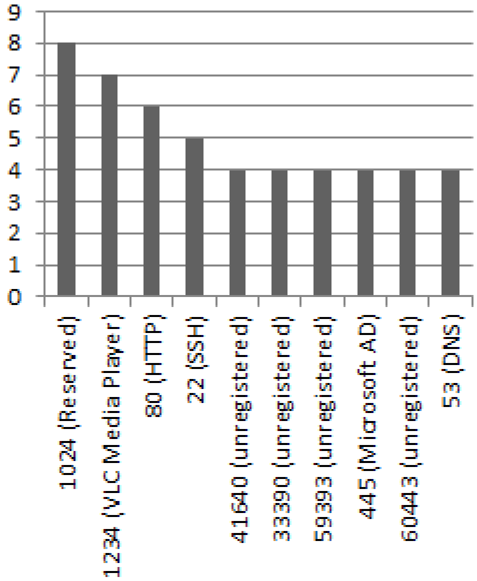

Figure 12. Count of IP Addresses by Blocked Connection Port

From the open ports, it seems that some of APT 1 infrastructure had need to allow connections to specific ports. Many of the ports that occur are frequently used in exploitations. For instance port 3389 is used for remote control of devices and port 21 is used for file transfers.

All the APT 1 infrastructure information retrieved can be analyzed and combined in several ways. For instance, looking at the connection information with country, it is seen that connection type is not dependent on country. Fixed leased lines are common across the board and the few dialup connections occur in Singapore and the U.S. The above results are just brief highlights of the data to illustrate the type of information that was retrieved and is not meant to be a comprehensive presentation of everything that can be derived from the data.

4.1.3 Comparison of Findings with Mandiant Report

Comparing the results of this case study to the Mandiant report provides validation of the accuracy of the method of reporting described here. Without the in-depth information, length of time, interaction with exploited organizations, or man power used by Mandiant, several useful points were extracted.

Table 3. Case Study 1 Comparison with Mandiant Report

Mandiant Report Information	Case Study 1 Findings
Targets are in the United States	Exploited devices are in the United States, mostly California
Targets are government, education, and big business	Exploited devices are hosting providers
Command and control servers communicate over port 443	Many instances of unfiltered port 443 on exploited devices
Many of the used malware communicate using remote desktop and similar protocols	Many instances of unfiltered, fully opened ports 3389 (the standard remote desktop protocol port), 1723 (the standard point to point tunneling protocol port), and 111 (the standard remote procedure call protocol port)
Used compromised mail servers to engage in social engineering	Many instances of ports 25, 110, 587, and 993, 995 (standard mail protocol ports)

First, the exploited targets are primarily located in the United States. This is presented in the Mandiant report, but beyond the Mandiant report, location of much infrastructure was further pinpointed to California. Other countries mentioned in the report also show in this method, but not nearly to the extent of the United States.

Second, the exploited targets are primarily IP addresses belonging to hosting services. From the Mandiant report, it was expected to see more government, education, and big business than occur in the data set. While there are several instances of each, what was unexpected is the number of hosting providers. This information did not lend to drawing the conclusion that this threat is interested in cyber-espionage, but it does make a point about how APT 1 operates that is not included in the Mandiant report. It also makes the threat relevant to more people—people outside of government, education, and large corporations are being exploited as well.

Third, the APT 1 infrastructure appears to have several TCP ports that are necessary for operation. In general non-HTTP ports should not show as open to random network scans. In most networks the ports should at least be filtered by a firewall, if not blocked completely. The most frequent open ports in this case study are used for protocols such as HTTP, HTTPs, remote desktop programs, e-mail, and DNS. Open HTTP related ports, such as ports 80 and 443 might be expected for IP addresses related to hosting providers, but it is normal procedure for network administrators to filter or block external traffic to other ports. Protocols mentioned in the Mandiant report, as well as descriptions of how infrastructure worked matched the frequent ports that were identified as open in the Internet Census 2012 data, as well as the ports showing blocked connections in the Internet Storm Center data.

There is a lot of information in the Mandiant report that cannot be obtained from external sources, but the method described here provided useful information. Specifically for APT 1, the method here reinforced some findings, as well as gave additional information on geographic location of exploitation, whose devices were being targeted, and what operating systems those devices use. Beyond APT 1, this case study illustrates the practicality of the method for use in other situations where a report from in-depth internal information is available, but the report is missing certain generalizations of the threat. It also shows that information provided could have been useful for identifying likely relevant characteristics of data from internal sources, such as network traffic connecting to or from specific ports.

4.2 CASE STUDY 2: MABEZA INFECTED MALWARE GROUP

Mabeza Infected is a family of related malware. Domain names associated with this family of malware from January through August 2013 were used as the initial set of indicators for this case study. This malware was known to be in existence before January 2013, but no other information on this group of malware was known to the author before applying the methodology overviewed in section 3.2.

This second case study shows how the described method can provide information about a threat where no information other than domain names are known at the start. Since there is no existing description to compare against the results for *Mabeza Infected*, the results must be evaluated with regards to reasonableness. The rest of this section first reviews the method and then discusses the findings.

4.2.1 Applying the Methodology to *Mabeza Infected*

Following the method outlined above, the first step to gather relevant information is obtaining a list of IP addresses. The domain names known to be associated with the *Mabeza Infected* malware family from 2013 were used to query the SIE database. This returned a list of 2,690 unique addresses, which included the reserved IP address 0.0.0.0, which had no information in any of the data sources. These addresses were then used to extract relevant information from each of the other data sources. With the available data, 2,547 IP addresses were able to be associated with autonomous system numbers and 2,546 were associated with owners. The one with an ASN but no owner was looked up manually in the RIPE NCC database. All but the reserved IP address was associated with a country of origin. At least two-thirds of the addresses had one or more of state (or province, as the case may be), while connection types occurred for three-fifths. Over 80% of the IP addresses had open ports, but very few could be associated with fingerprint identities. Not quite 30% of the IP addresses showed up in the ISC data as having blocked connections during January through August of 2013.

4.2.2 *Mabeza Infected* Findings

Using the same method as was applied to APT 1, some general information was extrapolated for *Mabeza Infected* and put together in a short overview. As in the APT 1 Results section above, this section discusses each of the components of the overview individually—the autonomous system findings, geo-location, connection information, identities, and information on relevant TCP ports. It closes with the short overview that describes the malware group.

ASN and Owner Information. There are 53 unique autonomous system number associated with the IP addresses in this case study. The most frequent ASN occurs 2376 times, accounting for over 88% of the addresses. Only four ASNs in total occurred ten or more times. About 5.3% of the IP addresses were not associated with ASNs through the process. The ASNs that were identified are associated with 52 owners; two of the ASNs belong to Amazon, which is the owner with the most occurrences of IP addresses in the data set. Most of the other owners are not as well-known. All but three of the unknown addresses fall in the 54.240.188.0 – 54.255.190.255 block. A quick whois query on one of these addresses showed these addresses are all part of a block belonging to one ASN registered to Amazon. The descriptions for the ASNs with two or more occurrences are show in Figure 13.

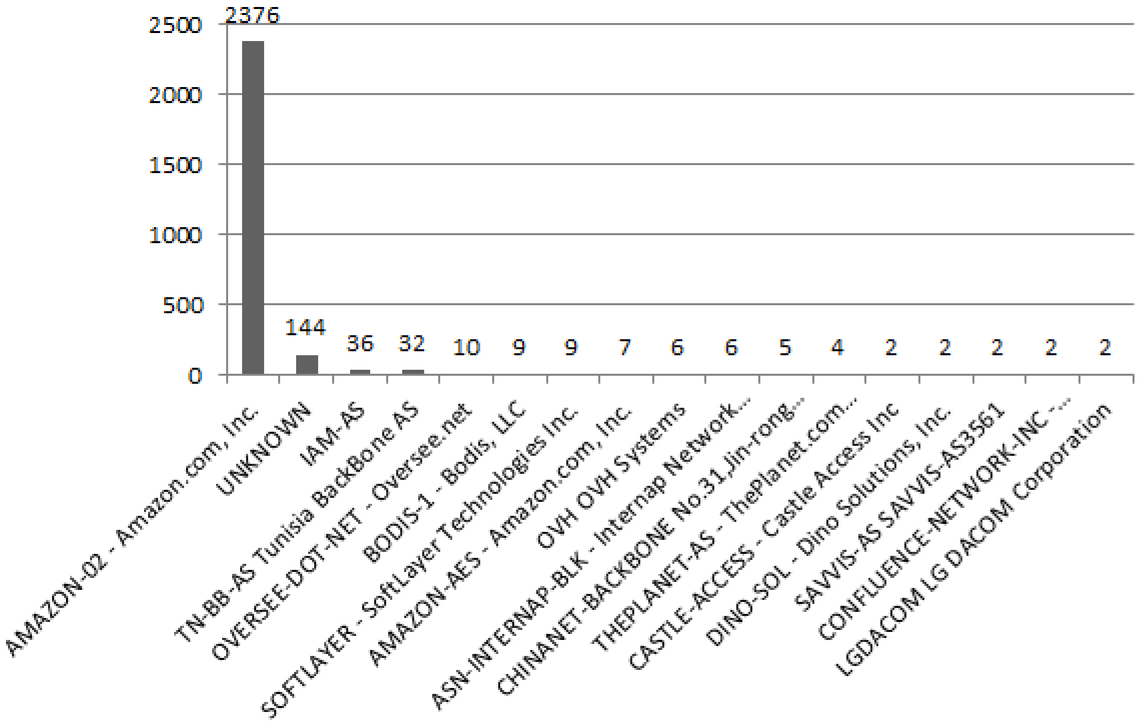

Figure 13. Count of Autonomous System Owners Occurring Two or More Times

So much of the attribution being for one owner within one ASN is interesting. This may indicate that the sensors for detection of the malware from the data source used to obtain IP addresses were biased toward that ASN. It could also indicate that this was the earliest exploitation or that this was the most vulnerable to exploitation. To get an idea which, if any of these, is the case, analysts could widen the time period of analysis or look for documentation of infections on other machines by other sources. For instance, does this malware exist in one of the other available malware database with associations to other IP addresses?

Geo-location Information. The geographical location information for this data set provided location information for all but one IP address. There were some conflicting locations between the MaxMind and Neustar data sources. For simplicity, the Neustar information is presented as it seems the more accurate data source, based on some manual research into IP address location done on other projects. For a less automated process, it would make sense to do some manual research on the addresses. The process results show 16 different countries, with 86.8% of the IP addresses resolving to the United States, with Canada and France following far behind with 5.0% and 2.6% respectively. Non-U.S. countries are shown in Figure 14.

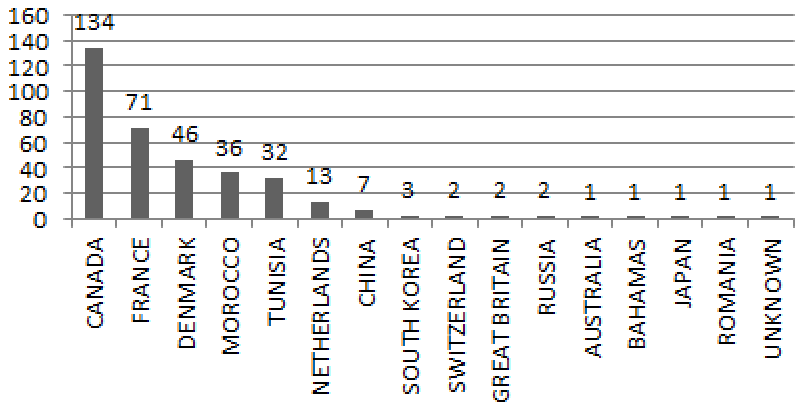

Figure 14. Count of IP Address by Non-U.S. Country

For the IP addresses that belong to the U.S., all had information at the state level. Texas had the highest occurrence, with 24.1% of the total data set addresses. Washington and California also had over 10% of the addresses from the total data set. The count of IP addresses by U.S. state is shown in Figure 15.

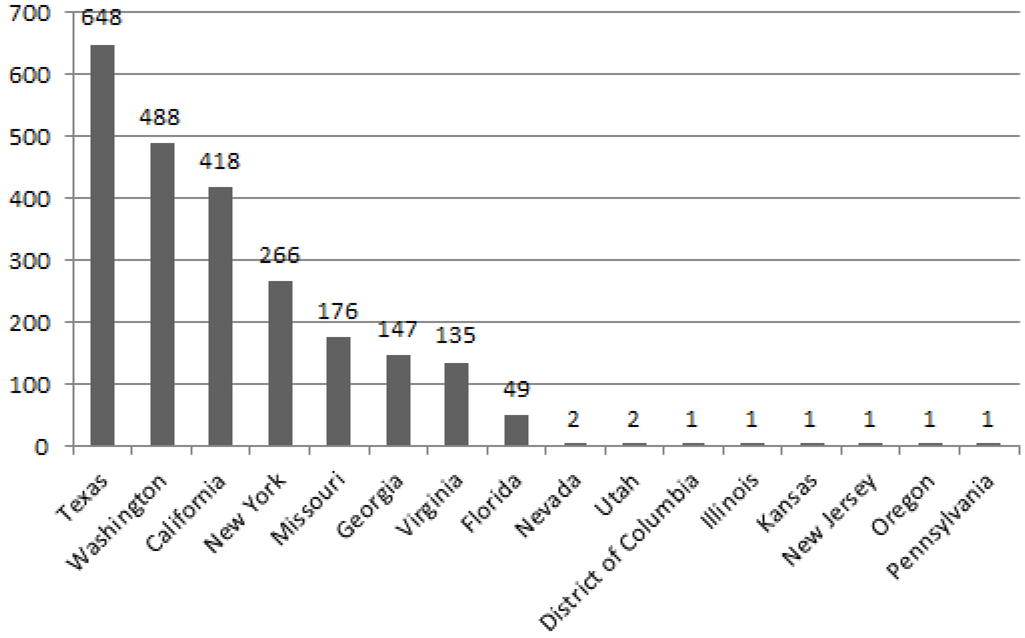

Figure 15. Count of IP Address by U.S. State

Considering that few of the IP addresses in this data set did not belong to Amazon, having a concentration of addresses in one country makes sense. It is interesting though that the state level is much more distributed. This implies that whatever caused the high occurrence of exploitation, for instance deliberate targeting or ease of exploitation, was not location dependent.

Internet Connection Method. The connection media and connection routing methods identified for this data set provided information for 62% of the IP addresses. The media types that occur for this data set are leased lines (tx), mobile wireless, DSL, dialup, and cable. Leased lines account for 59.2% of the address connection media. Figure 16 shows the count of IP addresses for each type.

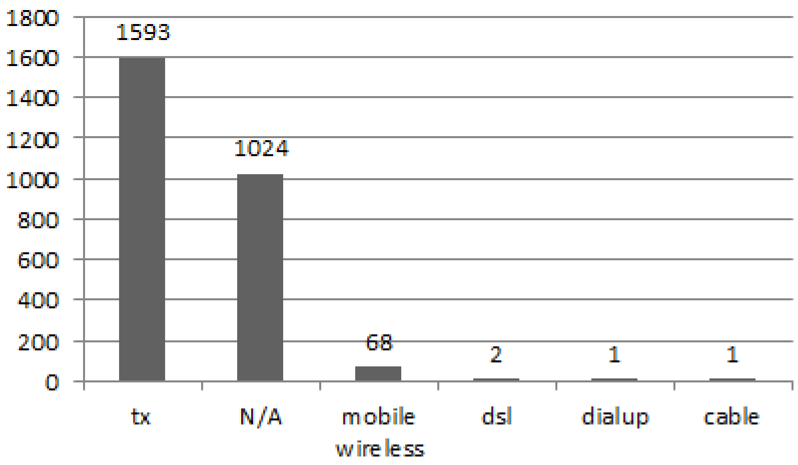

Figure 16. Count of IP Addresses by Media Type

In this data set, the routing types are fixed, mobile gateway, and point of presence (pop). Figure 17 shows the count of IP addresses by routing type. Figure 18 shows the combination of connection media and routing type.

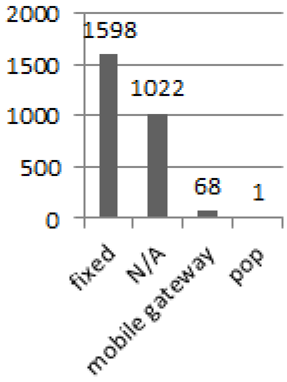

Figure 17. Count of IP Addresses by Routing Type

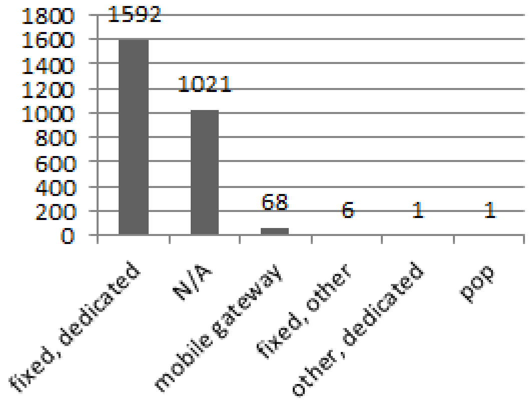

Figure 18. Count of IP Addresses by Media Connection Types

The connections that occur are what would be expected for devices related to Amazon. The fixed, leased lines are what would be expected for servers, or other devices, that require 24/7 access.

Device Identity. This data set had very few IP addresses with fingerprint information in the Internet Census 2012 data source and many of those that occurred had no matching identities or an identity that matches more than three possibilities. Of the devices that were identified, over half were Linux kernels in the range of 2.6-3. The results for device identification for the few addresses with fingerprints are shown in Figure 19.

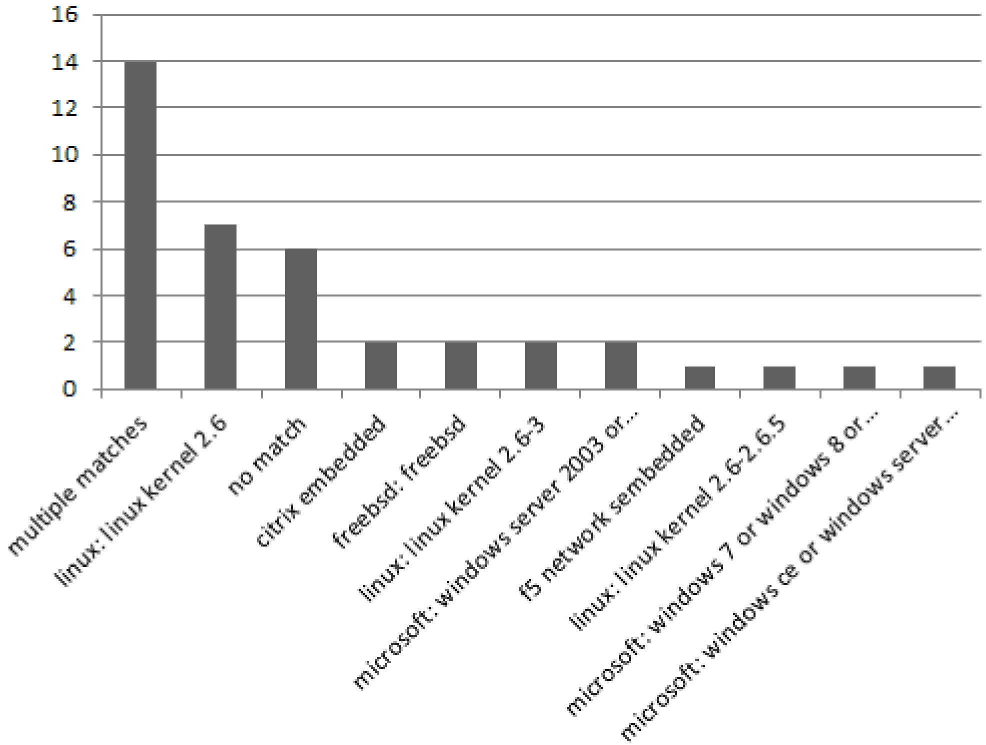

Figure 19. Count of IP Addresses by Identity

Important Ports. The lack of fingerprint information is surprising considering how well the IP addresses were well-represented in both port data sources. Over 80% of the addresses had at least one open port. There were a total of 6,646 unique port numbers open, with the most frequently occurring being ports 80 with 2156 addresses and 443 with 2116 addresses. As these are the ports for HTTP and HTTPS respectively, this indicates the devices are configured as web servers. Of the open ports, 80.8% of them occurred only once, and Figure 20 shows the open ports that occur for 16 or more addresses.

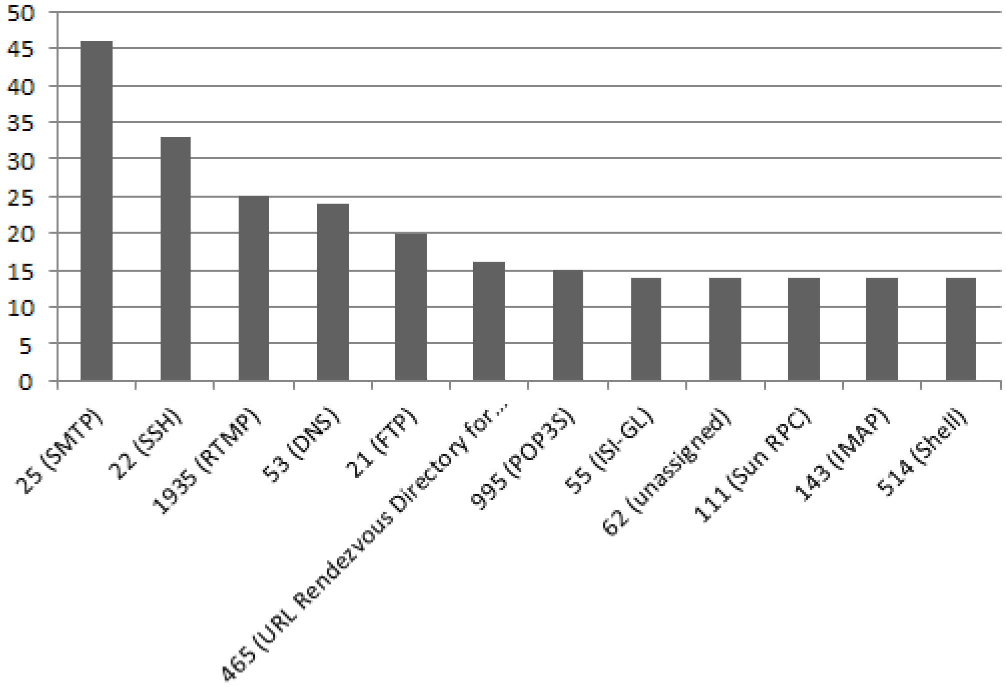

Figure 20. Count of IP Addresses by Open Ports

Most of these more frequent ports are not uncommon. For instance, the occurrence of 25, 995, and 143 indicate the presence of e-mail servers in the data set. Of more interest for this data set, is the wide range of open ports on these devices that occur only for one or two IP addresses. Since the presence of open ports indicates that there is some service listening for incoming connections, it would appear that something is opening seemingly indiscriminate ports on the devices. As organizations like Amazon tend to deploy devices with a standard configuration to aid in maintenance, troubleshooting, and resilience, this phenomenon would merit further in-depth analysis.

Blocked connection ports occurred for 29.4% of IP addresses. There were 10,275 unique blocked ports reported in the Internet Storm Center data source from January through August of 2013. The most frequently occurring were ports 1234 (VCL media player), 53 (DNS), and 80

(HTTP). Almost all the ports, 93.4%, only occurred once for this data set. This may indicate that the IP addresses did some form of network scanning or were searching for some of those seemingly indiscriminate ports that were identified as open from the IC 2012 data source. Again, this is something that would merit further analysis. Figure 21 shows the ports that occurred for four or more addresses.

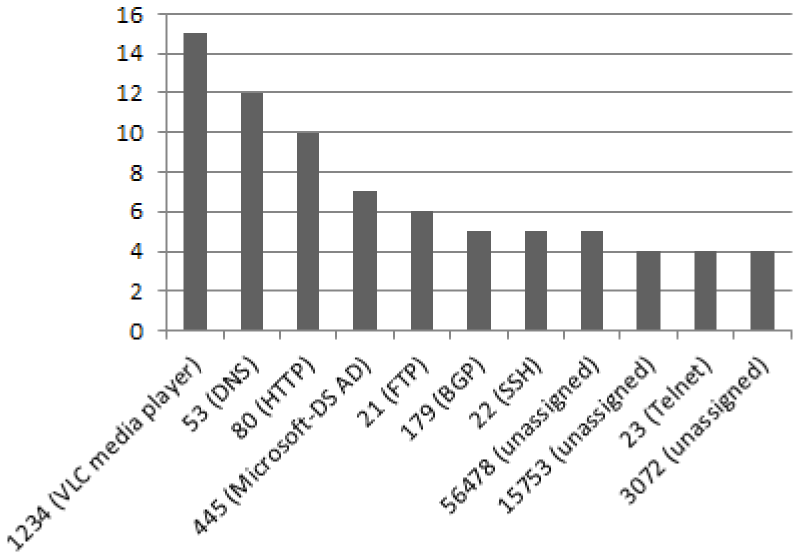

Figure 21. Count of IP Addresses by Blocked Connection Port

From the information on the open and blocked ports, it appears that *Mabeza Infected* is not particular about ports. The occurrences of ports 80 and 443 make sense since most of the IP addresses are attributed to Amazon, but the wide range of other ports and low number of occurrences for any individual port number within that range does not point to any specific port number being important for the malware. If the malware required specific ports, it would be expected to see a large number of those particular ports open on infected devices.

4.2.3 *Mabeza Infected* Overview

The January through August 2013 time period used to obtain IP addresses for this case study do not correspond to the Internet Census 2012 time period, but the other data sets had versions corresponding to that date range. Based on the information from the other data sets, the devices appear to be related to organizations and connections where it is reasonable to assume the devices do not change often, other than software patching. Also, this malware was known to be in existence in 2012, during the time period of the census. Both of these details suggest that the information provided in the Internet Census 2012 data set is still relevant to this threat analysis. The overview of this threat is summarized in Table 4.

Table 4. Mabeza Infected Overview

Questions Asked	Case Study 2 Findings
Who are the organizations that are compromised?	~ 90% Amazon
Where are the organizations that are compromised?	>85% in the US, primarily Texas, Washington, and California
How are compromised devices connected to the internet?	Using dedicated connections over leased lines
What ports are important to the threat?	80, mail ports, and having lots of pseudo-random, often all on one device

Based on the information from January through August of 2013, the *Mabeza Infected* malware primarily has exploited IP addresses belonging to Amazon.com. While there are exploitations related to IP addresses of other organizations, their numbers are miniscule in comparison to the number associated with Amazon. Most of the IP addresses are geographically located throughout the United States, with addresses for Amazon occurring in eight states, including Texas, Washington, California, and New York. Amazon also had some exploited IP addresses resolve to Canada, France, Germany, and the Netherlands. Where connection information is known, the connections are mostly over fixed, leased lines, which, as would be expected, is the only type that occurs for Amazon. Two of the occurring ASNs are comprised entirely of mobile wireless connections over a mobile gateway—one belonging to the Tunisia Backbone and one belonging to an ASN with a description of IAM, which a quick whois lookup from the RIPE website reveals to be “Itissalat Al-MAGHRIB, MAROC TELECOM”.

While there is not much information available concerning the identities of the devices associated with the known exploited IP addresses, there is a lot of port information. Both open ports on the IP addresses and the blocked ports where the IP addresses attempt to connect indicated that no ports in particular are important to this malware group, unless the malware requires ports 80 or 443. The information may indicate that the malware does some form of port scanning or is using some algorithm to choose a port to open for some purpose.

From this information, network administrators at Amazon would learn they need to check out the wide range of unexpected listening ports and the traffic sent through them, on devices located in California, Washington, and Texas. They would also see that they need to check webservers across the U.S. for compromise. They may also decide that better monitoring of listening ports is warranted. Analysts outside Amazon may be more interested in the concentration of compromised devices to a few organizations or in finding out if there is significance within the malware itself to having exploited devices with many open ports. Other analysts may find the information useful in creating reports or policies related to generic malware or threat issues or as a data set for more theoretical threat research.

5.0 CONCLUSION

In this thesis I have discussed a method that provides an alternative to traditional log, network flow, and other forensic analysis to learn about a threat. The method takes data sets from multiple sources that answer a set of questions to tell who are targeted, where exploitations are occurring, and what are some of the characteristics of the exploited devices. Unlike the data sets required for traditional analysis, these data sets are external to exploited organizations. Consequently this analysis method is available to a wider group of researchers—academic, industry, and government.

The general method for building the proposed analysis process is to choose sources, create tools, initialize analysis, use the tools for data extraction and reporting, and analyze the output. Choosing sources and creating tools is a one time job that results in a process that can be used to evaluate many threats. It is important to make choices that are accurate, cover a broad portion of the internet, and can be used in an automated process. Once the process is developed, retrieving by initializing the analysis with a set of exploited IP addresses or domain names, invoking the tools, and transforming the output into a high-level overview of a threat is easy and relatively quick (on the order of hours, not days).

After explaining a general methodology to create a process, I described an example implementation using the method. This included a description of each of the sources and why they were chosen. The tools and automation were explained, without getting into the code level details, which would change based on the sources used and which version of a source was chosen for inclusion.

The case studies presented in Chapter 4.0 illustrate how the methodology can be utilized to provide valuable information. In the Advanced Persistent Threat 1 study, it was shown that the method resulted in information that indicated the threat was targeting specific infrastructure based on location, what sort of organization owned IP addresses, and showed how results can indicate that exploits find certain device characteristics and ports useful. Furthermore, this study illustrated how the results compared with, and complemented, results from a more in-depth, time-consuming analysis method carried out by Mandiant.

The *Mabeza Infected* case study showed that the method resulted in information that indicated the threat was mainly exploiting devices from a specific organization. It showed that even when one individual component of the data sources did not provide much information, the overview was still informative.

There are other possible applications for this methodology beyond simple threat analysis. Since analysts and researchers can use this process on any set of IP addresses, another interesting application of this method is sampling a network population. Using good sampling techniques to create a set of addresses that represent the address population, an overview can be created of the whole network population. The results may be interesting on their own right, but also can be informative when analyzing a threat. Comparing the results of an analysis of a specific threat to the network population as a whole can provide an idea of just how anomalous is the profile of the threat. Similar profiles do not necessarily indicate a problem with the analysis, nor does it mean that the threat profile is not useful. Rather profile similarity or disparity is more an indication of how a threat works. The more similar the profiles, the less specific the targets of that threat and the more closely the threat works within usual network behavior.

Whether used as an initial threat analysis, to complement existing information, or to profile a sample of a network, there are some considerations and limitations analysts should keep in mind in regard to the information available and how results are interpreted.

The information an analyst has available, from the sources chosen for the extraction and summarizing process to the initial information of device exploitations that are related, will determine the robustness and accuracy of the method's output. Inaccurate source or a set of exploitations that are not actually related will produce results that are incorrect and may do more harm than good. Also, small sets of exploited devices may not provide any information, especially for sets that are external to the analyst's organization. Organizations should be able to provide information on all the devices they own, but when dealing with external devices, it is likely the available data sources used in analysis will have some gaps in coverage.

When it comes to interpreting results, interpretations from internal and external sets of information have different domains. The results from analysis of internal data sets cannot be projected into a "global" view of a threat. The data may not be representative of how the threat works in general, but rather is an indication of how the threat exploits in respect to the local organization. On the other hand, while the results of analysis from external data sets are more likely to reflect a global view of a threat, analysts still must keep in mind that the information they analyzed was likely only a portion of the devices a threat actually exploited. For small sets of exploited devices, this may result in results similar to only having information from the analyst's local organization, especially if the information comes from devices belonging to a small number of organizations or individuals.

5.1 LIMITATIONS

When analyzing the output of any process created through the methodology described in this thesis, it is important to recognize there are some caveats for interpreting the information. Some of the limitations can be alleviated by analyst skill, but others are a product of the requirements. Limitations that cannot be lessened by an analyst's skill include issues with time, and the need for the prior knowledge of initial indicators that are related to the threat of interest. Limitations that can be alleviated by an analyst's experience include identifying whether the indicators belong to devices that are really exploited, and whether the information for the devices in the various data sources actually belong to those devices or if it belongs to edge devices of an internal network.

Timeframe of Exploit. The timeframe when the threat of interest exploited the devices associated with the indicators is a required piece of background information to ensure an accurate overview of the threat. If all the devices were exploited after the time that corresponds to all the overview data sources, analysts could say at most, what the devices looked like before exploit. If all the devices were exploited before the time that corresponds to all the data sources used to obtain the overview, there is no guarantee that the IP addresses have not been re-associated with new devices or that the devices themselves have not been cleaned up since exploitation. Information in the data sources should be within a timespan that is close to the time period when a device was known to be exploited.

Reliable Indicators. Another piece of required background information for this process to work is a set of indicators that are related to the threat of interest. The methodology described above cannot categorize indicators into related sets. Consequently, if a set of indicators are not already known to be related, the output of the process would not be a useful overview. Section

5.2 presents the possibility of future work in using the output of the process to do categorization. It also suggests using the methodology to build a process for generating other research data sets that do not necessarily require related indicators as a starting point for analysis.

Analysis Skill. Skilled analysts are able to use their past experiences to identify sources for indicators that are likely to represent actual exploits and not false positives or deliberate misinformation. Likewise, they may recognize anomalies in the data that indicates a device is a proxy or firewall mediating internet access for an internal network and not the actual device represented by an IP address. Ultimately, the validity of any overview generated by the process will reflect the analysts ability to choosing good data sources and to interpret the relevant data from those sources.

5.2 FUTURE WORK

This thesis focused on a methodology to generate a description of threats. It was not meant to be a guide into using the results. Two straightforward uses of the descriptions are to explain threats and their infrastructure targets to non-technical people and to guide further in-depth analysis. Beyond these uses, there are several possibilities that merit further research.

One possibility for use is to determine if different variants of malware are incorrectly being attributed to the same group of individuals. Sometimes threats are not one individual person or group, but are different actors using the same malware. When researchers see variations in malware or nuances in targets, it may indicate that several different groups of people are behind the threat. The information that results from applying the method in this study to generate descriptions may provide indicators that imply there are multiple groups of people behind what is currently perceived to be one threat.

Another possibility for use is to classify infrastructure into components. For instance, command and control servers have different requirements than the machines they are controlling. It may be possible to use the result with machine learning techniques to categorize the different devices identified as infrastructure for a threat at a component level.

One other possibility for the results is to identify other devices exploited by the threat. Beyond identifying devices on a particular network that fit the description, it may be possible to, in a sense, reverse the process and generate a list of devices across the internet that fit the description. This could be done by searching through the various data sets for the entries with a given characteristic, then intersecting the results across all data sets. If this methodology could be shown to be accurate in detecting other devices exploited by the same group, it would be helpful for researchers who need to contact exploited individuals and organizations. It may also provide assistance for researchers determining the origination of a threat.

The methodology presented here of using disparate data sets as building blocks to overview a threat can be expanded for other uses. For instance, data sets containing real world information are often unavailable for academic researchers. Finding more sources that all share a common attribute, such as IP address or domain name, would allow them to create their own data sets.

APPENDIX A

ETHICS AND THE INTERNET CENSUS 2012

The Internet Census 2012 (IC 2012) was conducted by an anonymous person who built a botnet to do an nmap scan of the IPv4 address space using unprotected devices. This raises ethical issue with the data set. Should researchers use the data as it was obtained illegitimately? Should researchers use the data as it contains information pertaining to other persons' belongings without their permission? If so, what are the researchers' obligations for the data?

To begin, it is helpful to understand what the internet census data (ICD) is and is not. First, what the ICD is. The ICD is a collection of network related information that is tied to IPv4 addresses. Most of this information was obtained by an nmap scan. The information from the nmap scan includes data that tells if an IP address was responsive, if the IP address had open ports, what information was returned when different ports were queried (called service probes), and the fingerprint for the IP address (which can be used to determine possible operating systems or what type of device is tied to the IP address). Not all addresses have all the information—only those that were responsive could provide any of the other information, and even responsive IP addresses did not all provide further information to the scan. Furthermore, some of the information is not related to the IP address queried, but rather a firewall or proxy that is filtering traffic for the queried IP address. In addition to the scan data, there are also traceroute paths and reverse DNS results in the ICD. The traceroute data shows path information between two IP addresses. The reverse DNS results came from queries against several of the largest domain name servers and tell domains that were associated with the IP addresses at the time of the query.

Second, what the ICD is not. The ICD is not personally identifiable information. Though there are some names, email addresses, and phone numbers in the reverse DNS results and service probe responses, there are not birthdates, place of birth, addresses, social security numbers, or other personal information. The email addresses and phone numbers appear to be from error messages that are directing the recipient to contact a help desk. The names may also occur in the service probe responses, but are mostly found as components of domain names in the reverse DNS results. Ultimately, IP addresses can be tied to organizations using the ICD in conjunction with public data (for instance, whois lookups, IP address geo-location) and in some cases can be used to state that a person used an IP address at one point in time with a specific device. However, when used alone the ICD cannot be used to tell anything else about an individual.

Furthermore, the ICD is not is current information. The scans occurred from March through December of 2012, meaning when the ICD was posted the newest information was already several months old. The internet is constantly changing—individual IP addresses are assigned, unassigned, and reassigned, devices are added, removed, and upgraded, firewall and proxy rules are modified. In short, internet related information becomes outdated very quickly. Information that is even several months old does not reflect the current state of affairs. The historic information is useful to see how the internet or individual IP addresses looked at the point in time of a scan, but cannot be used to describe the internet as a whole or an individual device connected to an IP address today.

Going back to the questions, should researchers use the ICD? Creating at botnet on devices you do not own and without permission is unethical, even if the device is sitting in the open alone. The question for researchers with the ICD, though, is should we be using the data from a botnet, not should we create one. This has multiple aspects from the different possible uses of the ICD.

Researchers that examine botnet structure, evolution, and impact, both immediate and future, have to look at the information related to the botnet and what it collects to be able to determine the necessary and appropriate response and to learn how to prevent similar botnets in the future. Only using information with the actual owner’s permission is often not practical or even possible. In some cases, the “owner” is unknown, may be the botnet operator, is in a foreign location, or unwilling to admit they were compromised.

What is not as apparent in appropriateness is a researcher using information from a botnet for an unrelated study. Should the information be used just because it is available? If the information is personally identifiable, contains trade secrets, is classified, or some other form of legally restricted data, the answer is no, not without some legal right. If the information does not fall within these categories, it may be appropriate to use the information, but with care and respect for the people who were impacted by the collection. Again though, use of the information should be done in a manner that does not cause further problems for the individuals that were victims of the botnet and takes measures to protect their privacy, even though it has already been breached.

APPENDIX B

ALTERNATIVE ASSOCIATIVE METHOD USING SILK

Python scripts were chosen as the method for associating IP addresses to the MaxMind, Neustar, and Internet Census 2012 data as they are simple to understand, modify, and available in a wide range of settings. However, scripts are not the only method for doing the association and they are not the most efficient. Other options are to read the different data sets into database tables or to use the SiLK tool set. This is more efficient than parsing text files as the SiLK tool set stores data in binary format and does filtering and lookups with binary searches [29]. This appendix shows the process for using the SiLK tools to create a mapping file that accomplishes the association. It uses the MaxMind GeoLite Country file as the example.

SiLK, the System for Internet-Level Knowledge, is a network analysis tool created by the CERT Network Situational Awareness Team. Designed to run in a Linux environment, the tool set is available as a free download from http://tools.netsa.cert.org/silk/.

One of the benefits of SiLK is its ability to create files that map information to IP addresses or ports. These files, call pmap files are generated from input files following one of the following formats:

Start_IP_Address End_IP_Address value

Start_Integer_Address End_Integer_Address value

IP_Address_CIDR_Block value

The MaxMind GeoLite Country file has the format:

Start_IP_Address,End_IP_Address,Start_Integer_Address,End_Integer_Address,country _code, country

The following Linux command will transform the file, in this example called GeoIPCountryWhois.csv, into the Start_IP_Address End_IP_Address value format.

```
          awk -F"," '{print $1 " " $2 " " $5}'
GeoIPCountryWhois_08072013.csv | awk -F"\"" '{print $2 " " $4 "
" $6 " "}' > maxmindToPmap.txt
```

The following SiLK command will transform the file from the command above into the actual pmap file.

```
     rwpmapbuild --in=maxmindToPmap.txt --
out=maxMindCountry.pmap
```

The following SiLK command will use the pmap file to generate a pipe delimited text file where the fields are IP address, country code, and count of IP address occurrence (which should always be one). The input file, in this example called inIPs.txt, is a list of IP addresses in dotted notation, one per line.

```
     rwtuc --field=sip inIPs.txt| rwuniq --pmap-
file=mmCountry:maxMindCountry.pmap --field=sip,src-mmCountry --
no-col > maxMindIPCountry.txt
```

BIBLIOGRAPHY

[1] **McMillan, Robert.** Experts bicker over Conficker numbers. *TechWorld.* [Online] April 15, 2009. [Cited: August 16, 2013.] http://news.techworld.com/security/114307/experts-bicker-over-conficker-numbers/.

[2] Russian Business Network. *Wikipedia, The Free Encyclopedia.* [Online] August 16, 2013. [Cited: August 16, 2013.] http://en.wikipedia.org/wiki/Russian_Business_Network.

[3] **National Institute of Standards and Technology U.S. Department of Commerce.** *Managing Information Security Rist Organization, Mission, and Information System View.* Gaithersburg : National Institute of Standards and Technology, 2011.

[4] **Dell SecureWorks.** *Lifecycle of an Advanced Persistent Threat.* s.l. : Dell Secure Works, 2013.

[5] **McAfee.** *Combating Advanced Persistent Threats How to Prevent, Detect, and Remediate APTs.* Santa Clara : McAffee, 2011.

[6] **Bejtlich, Richard.** *CIRT-Level Response to Advanced Persistent Threat.* s.l. : SANS, 2010.

[7] **Shick, Deana and Horneman, Angela.** *Investigating Advanced Persistent Threat 1.* Pittsburgh : Software Engineering Institute, 2013.

[8] **Rieck, Konrad, Holz, Thorsten, Willems, Carsten, Düssel, Patrick and Laskov, Pavel.** *Learning and Classification of Malware Behavior.*Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 108-125, 2008.

[9] **Islam, Rafiqul and Altas, Irfan.** *A Comparative Study of Malware Family Classification.* Information and Communications Security, pp. 488-496, 2012.

[10] **Chen, Zhongqiang, Roussopoulos, Mema, Liang, Zhanyan, Zhang, Yuan, Chen, Zhongrong and Delis, Alex.** *Malware characteristics and threats on the internet ecosystem.* The Journal of Systems and Software, pp. 1650-1672, 2012.

[11] **Zhuang, Weiwei, Ye, Yanfang, Chen, Yong and Li, Tao.** *Ensemble clustering for internet security applications.*Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, pp. 1784-1796, 2012.

[12] **Bailey, Michael, Oberheide, Jon, Jahanian, Farnam, Andersen, Jon, Mao, Z. Morley and Nazario, Jose.** *Automated Classification and Analysis of Internet Malware.* Recent Advances in Intrusion Detection, pp. 178-197, 2007.

[13] **Feily, Maryam, Shahrestani, Alireza and Ramadass, Sureswaran.** *A Survey of Botnet and Botnet Detection.*Emerging Security Information, Systems and Technologies, pp. 268-273, 2009.

[14] **Halevy, Alon, Rajaraman, Anand and Ordille, Joann.** *Data Integration: The Teenage Years.* Seoul : VLDB Endowment, 2006. Proceeding VLDB '06 Proceedings of the 32nd International Conference on Very Large Databases. pp. 9-16.

[15] **Chaudhuri, Surajit and Dayal, Umeshwar.** *An Overview of Data warehousing and OLAP Technology.* 1997, ACM SIGMOD Record, pp. 65-74.

[16] **Yang, Hung-chih, Dasdan, Ali, Hsiao, Ruey-Lung and Parker, D. Stott.** *Map-Reduce-Merge: Simplified Relational Data Processing on Large Clusters.* Beijing : ACM, 2007. Proceedings of the 2007 ACM SIGMOD International Conference on management of Data. pp. 1029-1040.

[17] **Ziegast, Eric.** FloCon 2010 Proceedings: Introductin to SIE. *Cert.* [Online] January 12, 2010. http://www.cert.org/flocon/2010/proceedings.html.

[18] **University of Oregon Advanced Network Technology Center.** University of Oregon Route Views Project. [Online] January 27, 2005. [Cited: August 20, 2013.] http://www.routeviews.org/.

[19] **Réseaux IP Européens Network Coordination Centre.** Routing Information Service. *RIPE NCC.* [Online] 2013. [Cited: August 20, 2013.] http://www.ripe.net/data-tools/stats/ris.

[20] **MaxMind.** GeoLite City Accuracy for Selected Countries. *MaxMind.* [Online] 2013. [Cited: August 20, 2013.] http://www.maxmind.com/en/geolite_city_accuracy.

[21] **Neustar.** IP Intelligence: IP Data. *Neustar.* [Online] 2013. [Cited: August 20, 2013.] http://www.neustar.biz/enterprise/ip-intelligence/ip-intelligence-data.

[22] —. IP Intelligence GeoPoint Data Glossary. April 2013.

[23] **Anon.** Internet Census 2012 Port scanning /0 using insecure embedded devices. [Online] March 17, 2013.

[24] **Marsh, Nicholas.** *Nmap Cookbook: The Fat-free Guide to Network Scanning.* s.l. : CreateSpace, 2010. 1449902529.

[25] **Internet Storm Center.** About the Internet Storm Center. *Internet Storm Center.* [Online] 2013. [Cited: August 20, 2013.] https://isc.sans.edu/about.html.

[26] **Internet Assigned Numbers Authority.** IANA IPv4 Address Space Registry. *IANA.* [Online] 2013. [Cited: November 6, 2013.] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml.

[27] **Huston, Geoff.** Autnums. *Potaroo.net.* [Online] 2013. http://bgp.potaroo.net/cidr/autnums.html.

[28] **Mandiant.** *APT1: Exposing One of China's Cyber Espionage Units.* s.l. : Mandiant, 2013.

[29] **Shimeall, Timothy, Faber, Sidney, DeShon, Markus and Kompanek, Andrew.** *Using SiLK for Network Traffic Analysis Analyst's Handbook.* Pittsburgh : Carnegie Mellon University, 2010.