Identifying Digital Threats in a Hacker Web Forum

Mitch Macdonald, Richard Frank*, Joseph Mei, and Bryan Monk
School of Criminology, Simon Fraser University
British Columbia, Canada
{ mcmacdon, rfrank, jmmei, bmm8 } @sfu.ca

***Abstract* — Information threatening the security of critical infrastructures is exchanged over the Internet through communication platforms, such as web forums. This information can be used by malicious hackers to attack their computer networks and data systems. Much of the literature on the hacking of critical infrastructure has focused on developing typologies of cyber-attacks, but has not examined the communication activities of the actors involved. To address this gap in the literature, the language of hackers was analyzed to identify potential threats against critical infrastructures using automated analysis tools. First, discussion posts were collected from a selected web forum using a customized web-crawler. Posts were analyzed using a parts of speech tagger, which helped determine a list of keywords used to query the data. Next, a sentiment analysis tool scored these keywords, which were then analyzed to determine the effectiveness of this method.**

Keywords—hacking; critical infrastructure; natural language processing; sentiment analysis

I. Introduction

Cyber-attacks targeting the computer and control systems of critical infrastructures are increasing in both scope and prevalence. There is evidence of a dramatic global increase in attacks targeting power grids, financial institutions, and transportation networks [17]. For example, approximately one quarter of organizations operating critical infrastructure report malware attacks [21]. Due to the increased connectivity of the internet and lack of capable guardianship in the form of computer security, malware is easily spread over the internet. Virtual infrastructures are frequently targeted in these attacks to gain access to massive databases of information. In 2014 alone, three of the ten largest data breaches in history occurred and more than 822 million records were exposed worldwide, exceeding 2013's statistics [22]. The proceeds of cyber-attacks surpass $US 1 trillion per year, affecting millions of individuals, corporations, and governments worldwide [32].

Due to the increased threat of digital attacks, the communal networks of cyber criminals must be monitored. More specifically, their communication activity must be analyzed to determine potential threats. Chapter II provides background on critical infrastructure, before implicating hacker communities as threats to these systems. The characteristics of hacker communities are then discussed and the motivations for the current study are contextualized; however, analyzing virtual communities, such as web-based discussion forums (or simply, web forums), poses serious methodological challenges due to the large volume of posts made to these websites. These issues will be addressed in Chapter III. First, we first used a customized web-crawler specifically designed to collect data from web forums. Second, we used two analytic tools to analyze the language of forum members: 1) a parts of speech (POS) tagger extracted nouns from the data and 16 were selected as keywords; and 2) keywords potentially referencing critical infrastructure were cross-referenced with those referring to hacking tools and hacker jargon, each of which were then assigned sentiment scores. Results are presented in Chapter IV, which indicate the most negative sentiment is directed toward the financial and government sectors. Finally, in Chapter V, the implications of the findings are discussed and contextualized within the current literature. The limitations associated with this study and directions for future search are also discussed. Chapter VI concludes the study.

II. Background

A. Critical Infrastructure

Critical infrastructures are social and economic entities that contribute to the overall functioning and security of a nation [27]. Whether physical or virtual structures, critical infrastructure includes the systems, facilities, technologies, networks, assets, and services that support the health, safety, security, and economic well-being of both individuals and communities, as well as the functioning of government [26, 31]. These include, but are not limited to, government, healthcare, energy, information and communication technology, and transportation systems. Furthermore, critical infrastructures are often interconnected and interdependent within and across provincial, state, or territorial boundaries, as well as international borders [23]. These interconnected systems are massive, geographically dispersed, and are controlled by many different players—mostly within the private sector [23]; however, critical infrastructures are increasingly being connected to the Internet, making them vulnerable to cyber-attacks that cause physical damage, disrupt vital services, or threaten the economy [1, 24].

As critical infrastructures become increasingly digitalized, cyber-attacks targeting these systems are predicted to continually increase. The most industrialized countries might also be the most vulnerable to these attacks, because they are the most dependent on digital networks [2]. Diverse groups of actors, including nation-states, terrorist groups, and criminal organizations, methodically target and plan attacks against critical infrastructures [25]. These cyber-attacks target the computer networks, industrial control systems, or supervisory control and data acquisition systems that operate and monitor their physical systems. These attacks are carried out to shut down these networks, delete files, manipulate equipment, or steal sensitive information [5]; however, the expansion of the Internet has shifted global power dynamics such that other non-state actors, including single or small groups of hackers, are now capable of launching cyber-attacks against critical infrastructures [5, 28]. For this reason, it becomes pertinent to analyze hacker communities.

IPR2025-01398
CrowdStrike EX1026 Page 1

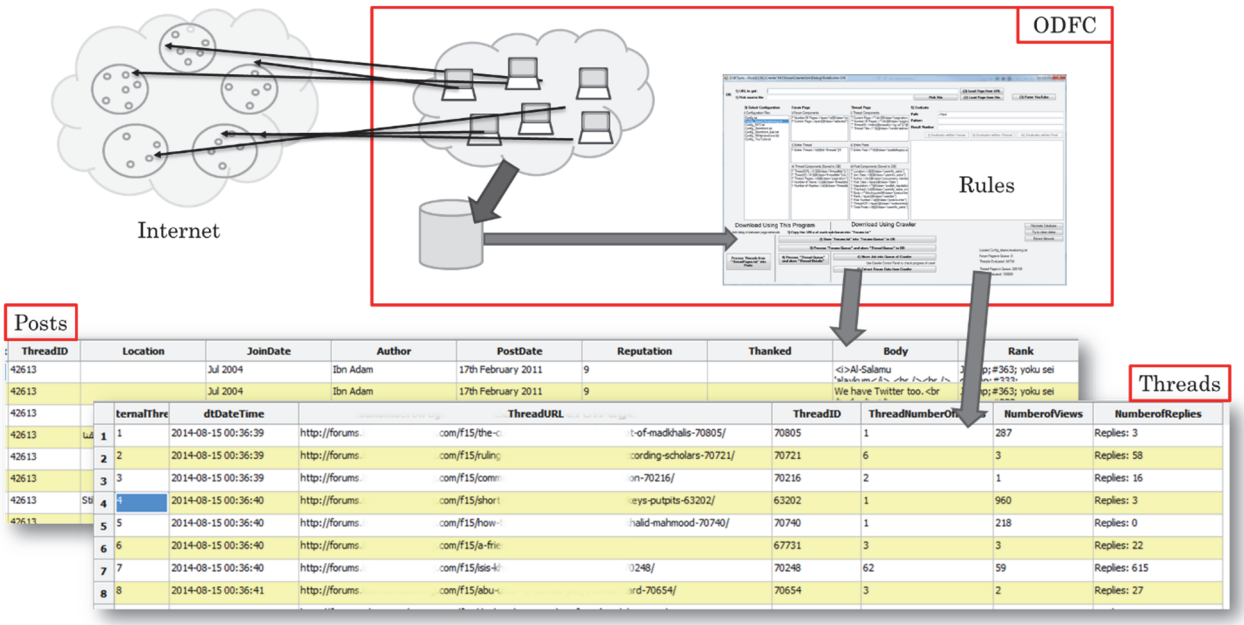

Figure 1 - A summary of the entire process

B. *Hacker Communities*

Hacker communities are loosely constructed social networks that facilitate the exchange and distribution of information [4, 6]. Hacker communities are dark networks, so members explicitly protect their anonymity, which serves to counter detection from law enforcement [29]. One way to protect anonymity is by communicating over the internet. For this reason, web forums are a popular method of communication amongst hacker communities [6]. Approximately 140 hacker web forums have been identified on the public Internet that cater to English-speakers, although this figure is likely a gross underestimation of the prevalence of these online communities [30]. Many of these web forums are freely accessible, which means that membership is not required to access content. Simply put, anyone who has access to the public Internet has the ability to view the information exchanged through open web forums. As noted by Eric Filol, this concept is important, because "the proliferation of open data, not cyber warfare, poses the larger threat to [a country's] critical infrastructures" [8]. It is then relevant to not only examine the extent to which critical infrastructures are discussed in hacker web forums, but also the language and sentiment used by forum members when referring to critical infrastructures to better understand the threats these communities pose [3].

C. *The Current Study*

Much of the current literature has focused on describing or developing typologies of cyber-attacks, but has not directly examined the individuals implicated in these crimes. Through a dual approach of POS tagging and sentiment analysis, this study makes two primary contributions: 1) to the cyber security detection methods and analyses used to identify open source data (exchanged freely over the internet) and potential threats to critical infrastructure; and 2) to determine the extent to which and how critical infrastructures are expressed in hacker web forums. The end goal of this research is to better understand the context around the use of selected keywords; therefore, we believe these methods are complementary to one another. Following keyword selection, the sentiment of posts cross-referencing 'critical infrastructure' with 'hacking tools' and 'hacker jargon' were measured. These posts were then analyzed to determine the context in which they were used. In other words, we used sentiment analysis as a mechanism to identify potential threats cross-referencing critical infrastructure and the tools or strategies used to attack them. Ultimately, our goal is to develop automated tools that combine these methods to streamline data analysis and facilitate real-time monitoring to identify potential threats as they emerge.

III. Methods

A. *Data collection*

Data was collected using software called the Open Discussion Forum Crawler (ODFC). ODFC is a customized web-crawler designed to capture the content posted to web forums; the software captures information from a user-selected web forum by downloading its webpages, parsing the page apart with the use of forum-specific 'rules' to capture all the useful information present on the forum, which are then stored in a corresponding database. For an overview, see Figure 1. The database is designed to resemble the structure inherent to web forums and is navigated in the same way: each forum has many sub-forums, which in turn have many threads of discussion, each with at least one post.

Due to the thematic grouping of sub-forums and the chronological listings of threads, each webpage within a web forum is very structured. This structure is shown in Figure 2, which illustrates multiple sub-forums displayed one after the other. Each sub-forum has a name, followed by a description of the topic(s) of discussion, the number of replies, and views. Figure 3 is similar to Figure 2, but depicts the structure of posts,

Figure 2 - Forum content

Figure 3 - Thread content

```
<HTML>
    <DIV>
      <A href="index.html">Home</A>
    </DIV>
</HTML>
```

a) Ideal scenario

```
<HTML>
    <DIV>
      <A href="index.html">Home</A>
</HTML>
```

b) Real-Life result

Figure 4 - Ideal vs real-life HTML

rather than sub-forums. Posts exhibit the same implicit structure, as each post repetitively lists the *Author*, *body*, and *date* (for example). This structure is consistent from post-to-post and across threads. ODFC downloads each webpage (see Chapter III.A.1) and then performs extensive clean-up (Chapter III.A.2). Users then input rules, so that the software extracts relevant data (Chapter III.A.3), which can then be formatted through the use of 'patterns' (Chapter III.A.4).

1) Data Capture

The data-capture process starts with ODFC being given a list of sub-forums to retrieve (or a list of all sub-forums to download when the goal is to capture the entire forum). The process is as follows: 1) ODFC retrieves the first page of each sub-forum; 2) based on user-specified rules (see Chapter III.A.3), it calculates the URLs for all the pages within a sub-forum; 3) for each URL (i.e. each page of the sub-forum), ODFC retrieves the content of the webpage; 4) rules are applied to the webpage to determine relevant thread information—the URL of each thread and the number of webpages it is spread through; 5) ODFC downloads all thread-pages sequentially and rules are applied to it to retrieve targeted data in each post (*author*, *date*, *body* of the post, etc.), which is then stored in a database. Once the entire web forum is downloaded into the database, the data can be extracted in various ways, such as through keywords, specific authors, or even a social network.

2) Cleaning up the webpage

Before data can be extracted from a webpage, it has to be identified. To find relevant information, ODFC uses rules to retrieve various components from within the webpage. A rule is a {*path*, *pattern*} combination, where the *path* follows the *XPaths* standard for querying trees and *pattern* uses the standard *Xquery*. These techniques require the object being queried to be in a tree-like structure; however, HTML is not required to follow a tree-structure, as many website authors do not properly close tags. For example, in Figure 4, (b) has an unclosed <DIV> tag, while (a) shows the correct version. Unclosed tags confuse the data retrieval process. Thus, retrieved webpages must first be scanned for any unclosed tags, which are then closed through the addition of HTML closing statements (such as the extra </DIV> in Figure 4b). The resulting HTML then follows a standard tree-structure, which can be translated into a XML file for further querying using our rules. This allows us to specifically pinpoint any detail in the webpage and store it in the database. Due to differences in structure, rules are customized for each web forum.

3) Finding information within a webpage using paths

A path is defined as a sequence of branches from the root, using the format of the XPaths standard[1], and results in zero or more nodes in a tree. By applying a *rule* to a cleaned-up HTML webpage, ODFC is able to retrieve the exact information indicated in the rules (such as the *Author* of a thread). To determine the author of the second post in Figure 5, it must first be recognized there are three posts, all of which are attributed to *Authors*. To identify *Authors*, a series of branches (*html*, *body*, *table*, *tr*, and finally *td*) must be traversed, resulting in a path "/html/body/table/tr/td". In Figure 4, this process results in three distinct *Authors* (one for each post). To focus on the second author, the *rule* must be modified to traverse down specifically only on the second <*tr*> tag in <*table*>. The *rule* thus becomes "/html/body/table/tr[2]/td", which results in ODFC targeting "*Author 2*".

4) Interpreting data through patterns

The application of a rule usually results in the proper data collection, but not necessarily in the proper format. That is, rules can be used to pinpoint where a data element is, but cannot point to a piece of data within the larger data. For this reason, *patterns* are applied to eliminate unwanted text, perform counts, or perform mathematical operations on the results of the *path*. *Patterns* use the standard language of *Xquery*. Referring to Figure 5, to determine how many pages there are in this thread,

[1] http://www.w3.org/TR/xpath/

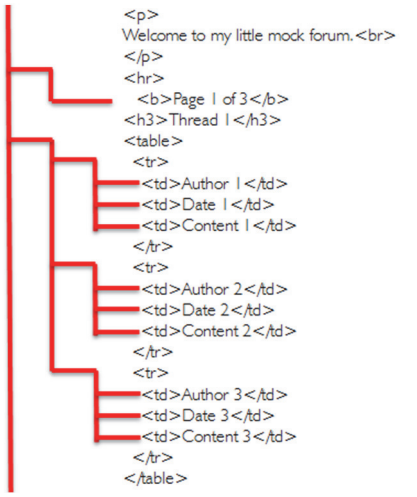

Figure 5 - HTML content as a tree.

Hacking Tools	# of References
botnet	1254
inject	1016
malware	671
virus	774
Trojan	501

Hacker Jargon	# of References
exploit	866
breach	341
backdoor	95
DDoS	918
SQL	1257
vulnerability	524

Critical Infrastructure	# of References
airline	312
bank	15 200
government	446
hospital	197
mobile	1302

Table 1 – Nouns extracted from web forum data used as keywords for sentiment analysis.

the first step is to recognize this information is shown in the text ("Page 1 of 3") following the <hr><b> tags. The complete *path* to this information is "/html/body/hr/b", or simply "//hr/b". This results in identifying "Page 1 of 3", but does not exactly answer the question of how many pages there are. To get to the "3", a *pattern* is created specific to that *path*. For example, the *pattern* "Page 1 of (?<RESULT1>)" would result in <RESULT1> = "3" being returned. This phrasing allows us to query multiple data-elements using a single *pattern*.

B. *Web forum selection*

The web forum analyzed for this study was downloaded by ODFC in its entirety. This web forum was selected for three reasons. First, it had large numbers of existing members, number of posts, and discussion threads. These characteristics increased the likelihood this web forum contained relevant and analyzable data. Second, the topics of its sub-forums were relevant to malicious hacking, which broadly corresponds with the focus of this study. Finally, the discussion between forum members is freely or publicly accessible on the Internet, which means these individuals do not have any reasonable expectation of privacy. Adherence to this criterion satisfies ethics concerns associated with this study; however, to address issues concerning privacy and the potential uncovering of personally identifiable information, the domain of this web forum will not be identified, nor will the forum member's user names.

C. *Keyword Extraction*

A parts of speech (POS) tagger was used to analyze the body of the posts and identify linguistic features within the data. POS tagging is a computational technique that dissects text into nouns, verbs, adjectives, as well as other components and applies labels to them. POS tagging has been used across various disciplines to analyze the use of language. One area of focus involves using POS tagging for domain specific searches [11]. For example, Ferraro et al. found that natural language processing (NLP), with a dictionary specific to the clinical domain, was better able to identify specific text features in their samples [12]. While POS tagging is an effective tool in its own right, it is also useful in conjunction with other analysis tools. For example, Dalrymple found using POS tagging as a parsing method increased the efficiency of other established linguistic analysis methods [13]. In other words, it is a method for understanding text without the manual, labour-intensive efforts traditionally associated with large, text-based data. It has also been used for a myriad of other purposes, including supplementing other data analysis techniques, such as sentiment analysis [10].

To automate the analysis of the posts, OpenNLP was integrated into ODFC. OpenNLP is a POS tagger which dissects and categorizes supplied text, splitting sentences into their base components and then labelling each of those components based on their part of speech (nouns, adjectives, adverbs) [9]. The POS tagger then creates an enumerated list, detailing the amount of times a particular word occurs within the text. This list can be used to understand the relative frequencies of words within the text. For this study, we focused only on the nouns. The POS tagger extracted a combined list of 11,432 nouns; however, not all were unique. Due to the design of the NLP software, issues regarding case sensitivity, the overlap between singular nouns and their pluralized form, and abbreviated or slang terminology particular to the hacker subculture all contributed to this greatly inflated number.

The list of nouns were manually revised and 16 applicable nouns were identified that had a relatively high frequency of in-text references and were considered to be potentially associated with malicious hacking and critical infrastructure (see Table 1). These keywords were used to query the database to arrive at a sample of posts, which were then cross-referenced and analyzed using sentiment analysis. The query function uses the root of the keywords, so that "virus" would retrieve posts containing both its singular and plural. Keywords were partitioned into three categories: 1) those that refer to hacking tools; 2) words specific to the hacker jargon; and 3) words potentially associated with critical infrastructure. Influenced by previous research, these keywords were selected based on their association with the hacker subculture and potential reference to critical infrastructure.

D. *Sentiment Analysis*

Sentiment analysis is a computational technique that attempts to analyze freeform text, by assigning *scores* to it after it is fed through a set of parameters or classifications. Sentiment

Analysis has been used primarily as a tool for analyzing user reviews of products and media [14]. Due to its ability to assign *positive* or *negative* scores to comments or reviews, companies can understand the general sentiment of online discussions; although sentiment analysis is also a useful tool for examining social networks and has been used in a variety of online contexts [15]. Within a security context, sentiment analysis has been used to examine the online activities of criminal subcultures, including extremist and hacker groups. For example, Li and Chen used deep learning-based sentiment to identify the key players involved in the sale of malware and stolen data [11].

For the current study, each of the critical infrastructure keywords were independently cross-referenced with each of the hacking tools and hacker jargon keywords; sentiment scores were computed for posts containing both types of keywords (henceforth, keyword pairing(s)). For example, each post that contained the keywords "bank" and "malware" would be assigned a sentiment score. In any instance that a post contained "bank", but not "malware" a score of 0 was assigned and were excluded from the analysis. It is important to note that keywords from the hacking tools and hacker jargon lists may have been used in the same post, so it is possible that posts may have occurred multiple times within the results.

Scores were calculated with the sentiment analysis tool SentiStrength—a dictionary based sentiment analysis program that can apply positive and negative values to text [16]. SentiStrength examines inputted text and then references its internal dictionaries to apply values or scores to it. More specifically, SentiStrength analyzes the words located on either side of keywords within a specified range. A custom 'sentiment wrapper' was used to expedite this process through simple automation of its command line inputs. Using this sentiment wrapper, the context surrounding each post for each keyword was analyzed. Specific words are given a numeric sentiment value based on the intensity of the sentiment. For example, "dumb" has a less negative sentiment value than "hate" [16]. Additionally, factors such as punctuation, capitalization, repetition, and specific emoticons can all affect sentiment values. In the context of this study, the output or score is a single value for each post and keyword combination. Each keyword was aggregated (both averaged and summed) across all posts to determine the sentiment around that keyword for the entire web forum.

Sentiment scores are not limited or restricted to a specific range of numbers. That is, scores are entirely dependent on the negative or positive terms found in the analyzed text. Despite this seemingly infinite scale, extreme scores (both positive and negative) are unlikely. Due to the length of the average post, there are simply not enough words with associated sentiment values to garner extreme scores. While it is possible, the likelihood of a post with a sentiment score greater than 100 or less than -100 is low.

E. Contextual Analysis

The posts extracted by the POS tagger and analyzed with SentiStrength were also manually analyzed. A contextual analysis was used to determine the effectiveness of combining these methods to identify potential threats within this web forum, as well as the 'severity' or extent of these threats.

IV. Results

Table 1 displays the keywords used to query the data collected by ODFC, as well as the number of times they are referenced within the database. There were 5134, 3083, and 17457 posts referencing the keywords in the "hacking tools", "hacker jargon", and "critical infrastructure" categories, respectively, for a gross total of 25 647 posts. Within and across categories, the POS tagger identified that certain keywords were more frequently used than others. For instance, "botnet" and "inject" were the most frequently referenced hacking tools with regards to the data collected by ODFC. Similarly, "SQL", "DDoS", and "exploit" were the most frequently used keywords within the hacker jargon. Finally, the number of posts referencing financial institutions were overwhelmingly overrepresented within the data, as illustrated by the 15 200 posts referencing the keyword "bank", compared to the other keywords relating to critical infrastructure. These figures suggest that certain keywords can potentially be used to distinguish between the level of interest afforded to particular infrastructure systems, as well as the most widely mentioned tools and and jargon associated with attacks.

Table 2 displays a subset of the summary of sentiment scores for each pairing of the cross-referenced keywords. In this way, Table 2 provides a visual depiction of the potential threats (hacking tools and hacker jargon) expressed in this web forum—specific to the critical infrastructure sector. Overall, the summary of sentiment scores indicates that posts mentioning hacker tools or hacker jargon along with "bank" and "government" grossed the largest scores, which were also negative. Rows without scores indicate that those keyword combinations did not occur within the data. For example, "hospital" and "backdoor" were not mentioned together in any post made to the web forum. Furthermore, while the keywords "airline" and "hospital" were seldom cross-referenced, this is likely representative of the larger data, given the relatively few number of times they were independently referenced (Table 1).

In total, there were 688 posts made to this web forum that cross-referenced critical infrastructure keywords with hacking tool and hacker jargon keywords. This figure represents roughly 2.68 percent of the 25 647 gross number of references made to all 16 keywords. Again, some keywords were referenced more than others. From the hacking tools list, "botnet" was cross-referenced in 97 posts, while "Trojan", "malware", "inject", and "virus" were cross-referenced in 91, 86, 59, and 39 posts, respectively. From those words specific to the hacker jargon, "breach" was referenced along with critical infrastructure in 88 posts, whereas "exploit" and "vulnerability" were both cross-referenced with these keywords in 79 posts, and "DDoS", "SQL", and "backdoor" were cross-referenced in 40, 26, and 12 posts, respectively. With regards to critical infrastructure, "bank" was cross-referenced in 459 posts, "government" in 126 posts, "mobile" in 93 posts, and "airline" and "hospital" were cross-referenced in 7 and 3 posts, respectively. More specifically, the three most frequent combinations of keywords all included "bank", which were cross-referenced with "botnet" in 80 posts, "Trojan" in 64 posts, and "breach" in 60 posts. Collectively, these posts represent 204 (29.65%) of the 688 total cross-referenced posts. These figures indicate the relative

Keywords	airline		bank		government		hospital		mobile	
	Avg.	Sum	Avg.	Sum	Avg.	Sum	Avg.	Sum	Avg.	Sum
backdoor	-	-	-1.4	-7	-5	-15	-	-	-2	-8
botnet	-	-	-2.09	-140	-4.33	-13	-	-	-2.4	0
breach	-	-	-6.76	**-399**	**-7.7**	-154	-	-	**-8.57**	-60
DDoS	-	-	**-10.68**	**-299**	-4.86	-34	-	-	-5.75	-23
exploit	-4	-4	-2.15	-112	-4.25	-51	-4.5	-9	-1.78	-16
inject	-5	-5	-2.47	-89	-3.11	-28	-	-	-2.4	-12
malware	-	-	-4.71	**-245**	-3.9	-78	-3	-3	0	-63
SQL	-5	-5	**-12.44**	-112	-2.4	-24	-	-	-3.33	-20
Trojan	-	-	-3.15	**-189**	-0.75	-6	-	-	-2.69	-43
virus	-3	6	-3.6	-90	2.67	8	2	2	-1	-6
vulnerability	-3	6	-6.73	**-222**	**-7.71**	-185	-	-	-5.62	-73
* **Bold** indicates the largest average (white) and summed (gray) sentiment scores within the sample.										

Table 2 – Summary of average and summed sentiment scores for the cross-referenced keywords.

infrequency in which these keyword combinations were used in posts made to this web forum.

The sentiment scores for the keyword pairings (displayed in Table 2) generally indicate disparities between the average and summed scores. For instance, only the pairing of "DDoS" and "bank" was amongst the largest scores for both average and summed sentiment. The other largest average sentiment scores amongst the pairings were "SQL" and "bank", "breach" and "government", "vulnerability" and "government", and "breach" and "mobile". In contrast, the summed sentiment scores indicate the other largest sentiment scores were posts that cross-referenced "bank" with "breach", "malware", "Trojan", and "vulnerability". While the summed scores are attributed to the total number of posts that referenced these keywords, they may also suggest the likelihood of potential threats or risk to particular critical infrastructures.

A. Contextual Analysis

A content analysis of selected keyword pairings listed in Table 2 was used to gauge the accuracy of the analysis tools at identifying potential threats to critical infrastructure. The focus was on analyzing the context in which certain keyword pairings were used. Particularly, individual posts that had high sentiment scores were analyzed. For example, "SQL" was cross-referenced with similar frequency with both "bank" and "government"—these pairings were referenced in 9 and 10 posts, respectively—and used in similar contexts:

- "The emails are sourced exclusively from SQL databases only. Very high quality databases. More than 95 percent of them are sourced from government databases".
- "The hacker group Team GhostShell returns with a massive data leak targeting the banking, aerospace, legal, education, government, military, and airline sectors. The data dump contains database details, names, email addresses, mailing lists, addresses, and passwords. The dump also contains the SQL Injection vulnerable links."
- "Here is a website sensitive information on banks, identified databases with bulk financial information, as well as simple mail transfer protocol (STMP) services."

Posts typically contained links to websites hosting information on banking and government institutions, information about recent data breaches from notorious hackers, advertised the sale of data stolen from computer databases, and websites that contained SQL injections and other hacking tools. Posts indicated that these websites, "... have some of the best tools to download. SQL Exploiter Pro v2.15 accesses databases and exploitable websites. It makes hacking websites easy. Pro hackers will love it for how organized it is. So simple, even novice hackers can use it."; however, "vulnerability" was used in varying contexts when cross-referenced with the keywords "bank" and "government". "Vulnerability" and "government" was used when referring to:

- "[Our group] doesn't only offer security for clients, but also hacking services—from owning your [corporate] competitors [or] breaking into [*.gov*] military-grade networks, financial institutions, or just sites you don't like."
- "A key actor in the [hacking] game is to always be ahead of the government. They are ahead of us when it comes to the key factor of technology, so to save yourself, here are some recommendations for guaranteed safety."
- This may help the [hacker group Anonymous] at the end of the month, although I don't know how many of their members are on this forum. I'm going to list *.gov* websites that are vulnerable and explain how they are vulnerable to cross site scripting (XSS)."

In contrast, when used in conjunction with "bank", "vulnerability" was used to discuss the susceptibility of financial institutions in the operation of bank cards and how to fraudulently exploit these vulnerabilities. To a lesser degree custom written programs, such as web injections, were referenced as other methods that could be used to take advantage of these vulnerabilities. More commonly, this pairing was used to refer to the susceptibility of financial institutions against botnet services. As one post described, "It is possible to build large botnets, which [launch] brute force attacks in hours, or even minutes...Practically all banks are vulnerable, some more and some less."

As previously mentioned, "botnet" was almost always used in association with "bank", as indicated by the 80 (82.47%) references made to the pairing. The pairing of "botnet" and "bank" was used in the following contexts:

- "Current botnets available: Zeus (the latest public version with 18 private United States (US) bank [SQL] injections); Spyeye (the latest public version with private European Union and US bank [SQL] injections); and Citadel. I will contact the sellers on your behalf and set it up for you."
- "I have experience with botnet and spamming, looking for work. If you have any custom target, I can work on it, be it bank accounts [or anything else]."
- "Hello, I am one of actual speakers of Anonymous Russia. Today, Friday, February 3, 2012 at 12:00 GMT [we are] starting a lockout of the US economy (banks). This movement is supported by botnet [attacks]. This lockout is designed to punish banks for their involvement in the global crisis, but also to show Americans that no one can violate freedom on the Internet and those who try will be punished."

This suggests a predominant use of these malware-infected networks to attack institutions in the financial sectors. In this context, "botnet" was often used to advertise the sale of various types of malware, all of which are well known to the hacker community, including "Zeus", "Spyeye", and "Citadel". "Botnets" would generally be advertised with "special offers" of SQL web injections that could be used with the corresponding malware. Forum members that operated botnets would also advertise their services to others looking to attack banking websites and steal financial information. While these examples are specific to financial motives within the hacker community, the sentiment analysis tool also identified potential threats motivated by political ideologies. For example, a forum member claiming to represent the hacktivist group Anonymous encouraged other forum members to use or join botnets to support a campaign targeting US banks.

V. Discussion

Taken together, the largest, and most negative, average and summed sentiment scores were in reference to the keyword "bank". This is consistent with the argument that financial motives are pervasive within the contemporary hacker culture [7]; however, overall there is discrepancy between sentiment scores for "bank" that cross-reference the hacking tools and hacker jargon keywords. For example, "breach" had the largest negative summed score, but did not rank within the largest average scores. Similarly, "SQL" had the second largest negative average score, but was not among the largest summed scores. Thus, while the average scores indicate the most negative content, the summed scores indicate the most frequent pairings. This suggests that average scores may better indicate the most serious threats, while the summed scores indicate the most likely targets and sources of threat. For instance, the pairing of "bank" and "DDoS" had both the largest average and second largest summed sentiment scores; financial institutions are popular targets for hackers and DDoS attacks are estimated to be the most frequent method of cyber-attack and account for 23 to 42 percent of all attacks against critical infrastructures [18, 19]. Thus, these findings support speculation that sentiment may be correlated with the actual threat of these attacks.

In comparison with other keyword pairings, those including "government" also had large sentiment scores. For instance, the pairing of "government" and "vulnerability" was amongst the largest average sentiment scores and also had a large summed sentiment score. These sentiments scores correspond with the increase in attacks targeting this sector. For example, reports indicate that the exploitation of vulnerabilities are attributed to 50 percent of all attacks against critical infrastructure systems, including those in the government sector [21]. More generally, "Trojan", "malware", and "virus" were cross-referenced in a gross total of 216 posts or 33.40% of the 688 posts that were analyzed; however, this figure is likely to be slightly inflated, due to the possibility that posts reference both "Trojan" and "virus". Nonetheless, this finding is consistent with statistics, which indicate that malware is responsible for approximately a third of cyber-attacks, such as those targeting the financial and energy sectors [20].

There were three main limitations associated with this research. First, sampling procedures ignored dialogue. In other words, the keywords used to query the data returned only those posts that contained the specified keyword pairings in the body of the text. Thus, the exchange of dialogue between forum members was missed; however, this limitation is inherent to our research design. While this design misses the evolution of dialogue, it was efficient in identifying and extracting posts used for automated and manual analysis. Second, due to SentiStrength's intentionally broad base configuration, many words specific to the hacker jargon and subculture were likely ignored. For example, SentiStrength's dictionary is not built with negative connotations towards words like *infect*, *leak*, and *crack*, which limits its ability to understand the particular context in which these words are used. Future research on virtual hacker communities should take this into account by amending the SentiStrength dictionary with additional terms specific to hacker language; however, while user coded data may prove to be a valuable modification, an underlying issue remains: users may not be able to effectively label or assign a sentiment value to this subculture's lexicon. Third, another issue surrounded the size and nature of the web forum. In regards to the nature of the forum, discussions about financial institutions were overrepresented within the data. That is, posts identifying critical infrastructure were largely in reference to "banks" and other financial intermediaries, which provide the opportunity for monetary gain. This meant that "airlines" and "hospitals", which contain a wealth of personal information, but less financial information, were not as frequently discussed. This could be remedied by expanding the number of forums studied, as others may focus more on attacks towards a variety of critical infrastructures.

Future research using keyword-based queries and detection methods should prioritize critical infrastructures based on perceived threats, risks, and the relative importance of the services, assets, and security that they contribute to the functioning of society. Furthermore, while this web forum was large, it provided opportunity to determine the effectiveness of automated tools to identify potential threats within a large data source. That said, future studies should consider smaller web forums for analysis. This research could compare whether or not smaller communities are more closed and tight-knit, which may potentially breed a more explicit criminal hacker subculture.

VI. Conclusions

Hacking is an ever growing issue in the modern world. With the rise of cloud computing and the tendency to store data outside of a local server, the threat to computer networks and data is ever present. This is especially true for critical infrastructures. Surveys show that these systems have become targeted in attacks and do not generally have up to date or sufficient digital protections [35]. In some cases, hackers post about their attacks before or after they occur. This necessitates the development of analytic tools that can identify the language and targets of hackers, without manually monitoring the large volume of posts made to hacker web forums.

This study sought to produce one of these tools through the use of a customized web-crawler, POS tagging, and sentiment analysis tool. From this analysis specific keywords were pinpointed that resulted in larger sentiment values than others and their posts were analyzed. Despite limitations involving the domain analyzed, it was possible to determine which posts with the co-occurrence of specific keywords contained information regarding different facets of hacking, including discussions of hacking targets, advertising of hacking services, requests for aid in hacking endeavors, and conversations surrounding major hacks perpetuated by well-known hacker groups.

References

1) V. Benjamin and H. Chen. Securing cyberspace: Identifying key actors in hacker communities. Presented at Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on. 2012.
2) T. M. Chen. Stuxnet, the real start of cyber warfare?[editor's note]. Network, IEEE 24(6), pp. 2-3. 2010.
3) D. Décary-Hétu, B. Dupont and F. Fortin. "Policing the hackers by hacking them: Studying online deviants in IRC chat rooms," in Networks and Network Analysis for Defence and Security Anonymous 2014.
4) T. J. Holt. Subcultural evolution? examining the influence of on-and off-line experiences on deviant subcultures. Deviant Behav. 28(2), pp. 171-198. 2007.
5) T. J. Holt and M. Kilger. Know your enemy: The social dynamics of hacking. The Honeynet Project pp. 1-17. 2012.
6) T. J. Holt, D. Strumsky, O. Smirnova and M. Kilger. Examining the social networks of malware writers and hackers. International Journal of Cyber Criminology 6(1), pp. 891-903. 2012.
7) M. Kilger, O. Arkin and J. Stutzman. Profiling. The Honeynet Project (2 Nd Ed.), Know Your Enemy.Addison Wesley Professional 2004.
8) Shaw, G. (2014). Eric Filol on withdrawing from CanSecWest Vancouver: Talk potential blueprint for terrorists. [Online]. Available: http://blogs.vancouversun.com/2014/03/13/eric-filiol-on-withdrawing-from-cansecwest-vancouver-talk-potential-blueprint-for-terrorists/
9) Chambers, L., Tromp, E., Pechenizkiy, M. and Gaber, M. "Mobile sentiment analysis," in: Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, (2012), San Sebastian.
10) Li, W., & Chen, H. "Identifying Top Sellers in Underground Economy Using Deep Learning-based Sentiment Analysis". Presented at Intelligence and Security Informatics (ISI), 2014 IEEE International Conference on. 2014.
11) Rayson, P., "From key words to key semantic domains," in International Journal of Corpus Linguistics, Vol. 13(4), (2008), p. 519-549
12) Ferraro, J., Daumé, H., DuVall, S., Chapman, W., Harkema, H., Haug, P., "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation," in American Medical Information Association, Vol. 20, (2013), pp. 931-939
13) Dalrymple, M., "How much can part-of-speech tagging help parsing?" in Natural language Engineering, Vol. 12(4), (2006), pp. 373-389
14) Dave, K., Lawrence, S., Pennock, D., "Mining the Peanut Gallery: Opinion Extraction and semantic classification of product reviews," in In Proceedings of the 12th international conference on World Wide Web (WWW '03). ACM, New York, NY, USA, 519-528.
15) Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), p. 82-88.
16) Thelwall, M., & Buckley, K. (2013). Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8), 1608–1617.
17) Brocklehurst, K. (2015, Feb 1). Cyberterrorists Attack on Critical Infrastructure Could be Imminent. [Online]. Available: http://www.tripwire.com/state-of-security/security-data-protection/security-controls/cyberterrorists-attack-on-critical-infrastructure-could-be-imminent/
18) Zorz, Z. (2014, Dec 31). 2015 predictions: Cyber attacks aimed at critical infrastructure, Aacks as a Service. [Online]. Available: http://www.net-security.org/secworld.php?id=17712
19) Lorhmann, D. (2015, Apr 17). Hacking Critical Infrastructure is Accelerating. [Online]. Avialable: https://www.techwire.net/hacking-critical-infrastructure-is-accelerating/
20) Paganini, P. (2014, Nov 24). Foreign Hackers Constantly Target US Critical Infrastructure. [Online]. Avialble: http://resources.infosecinstitute.com/foreign-hackers-constantly-target-us-critical-infrastructure/
21) Borilin, S. (2014, Sep 24). So malware attacks against critical infrastructure are inevitable. What's next?. [Online]. Available: https://business.kaspersky.com/so-malware-attacks-against-critical-infrastructure-are-inevitable-whats-next-2/2647
22) Dingman, S., Silcoff, S., & Greenspan, R. (2014, Nov 24). Hacked: The escalating arms race against cybercrime. [Online]. Available: http://www.theglobeandmail.com/report-on-business/hacked-the-escalating-arms-race-against cybercrime/article21305464/?page=all
23) Graham, A. "Canada's Critical Infrastructure: When is Safe Enough Sage Enough?" The Macdonald-Laurier Institute. 2014.
24) Miller, B., & Rowe, D. "A survey of SCADA and critical infrastructure incidents". In Proceedings of the 1st Annual Conf. on Research in Inf. Tech., ACM. 2012, pp. 51-56.
25) Auditor General of Canada. "2012 Fall Report of the Auditor General of Canada: Protecting Canadian Critical Infrastructure Against Cyber Threats". Office of the Auditor General of Canada No. FA1-2012. Ottawa, ON. 2012.
26) Public Safety Canada. "Action Plan for Critical Infrastructure (2014-2017)". Public Safety Canada No. PS4-66/2014E. Ottawa, ON. 2014.
27) Rege-Patwardhan, A. "Cybercrimes against critical infrastructures: a study of online criminal organization and techniques". Criminal Justice Studies, vol. 22, no. 3, pp. 261-271, 2009.
28) Lynn, W. J. "Remarks at the 28th Annual International Workshop on Global Security". June 16, 2011.
29) Holz, T., Engelberth, M., & Freiling, F. "Learning more about the underground economy: A case-study of keyloggers and dropzones" in Spring Berlin Heidelberg. 2009, pp. 1-18.
30) Frank, R., Macdonald, M., & Monk, B. "Identifying Potential Canadian Targets in Online Hacker Forums". Ottawa: Public Safety Canada (ongoing). 2015.
31) United States Department of Homeland Security. (2015). "What is Critical Infrastructure?". [Online]. Available: http://www.dhs.gov/what-critical-infrastructure
32) United Nations. (2011). "Cybersecurity: A global issue demanding a global approach". Available: http://www.un.org/en/development/desa/news/ecosoc/cybersecurity-demands-global-approach.html
33) Nan, C., Eusgeld, I., Kroger, W., "Analyzing vulnerabilities between SCADA system and SUC due to interdependancies," in Reliability Engineering & System Safety, Vol. 113, (2013), pp. 76-93