CHAPTER III

Methodology and Quantitative Studies of Ethical Hacking: Evidence-Based Decision and Policy-Making

This chapter features additional research by Kevin Kim, Adrian Agius, and Richard Li.

3.1 Report for Public Safety Canada, 2011

As mentioned in chapter 1, some of this book is based on a report that was commissioned in the fall of 2010 by Public Safety Canada. The report—*Ethical Hacking*—was finalized in 2011.[1] The report was not made available to the public at the time but was subject to freedom-of-information requests. Generously, Public Safety Canada has allowed me to retain intellectual property to publish the research. The bulk of the report, therefore, has found its way into this book.

The 2011 report examined five types of ethical hacking: hacktivism, online civil disobedience, penetration/intrusion testing, security activism, and counterattack/hackback. Each category was defined and a series of related aspects were examined using the following sub-headings:

- Selected Case Studies
- Motivation
- Main Targets
- Relation Between Targets and Motivations
- Fundamental Principles of "Hacker Ethics"

- Perceptions of the Illegality of Activity
- Deterrence Effects of Case Law and Convictions
- Relevant Case-Law Convictions
- Observations

The case studies for the report were selected based on a gathering of ethical-hacking incidences globally from 1999 until 2010. An extensive, multidisciplinary literature review (information systems, psychology, fiction, risk management, computer science, law, political science) was conducted and is included in the references at the end of this book. This was a labour-intensive process where we did a comprehensive literature review in three languages—English, French, and Russian. Most of the incidences were discovered due to media coverage of the topic.

The report was written in a few months because I was able to draw on my PhD work in cybercrime, entitled "Botnet Badinage: Regulatory Approaches to Combating Botnets" (PhD diss., University of New South Wales, 2011), where I interviewed people involved in the cyber-security industry (including law enforcement), as well as with some cybercriminals, and attended conferences around the world, including in eastern Europe, the United Kingdom, Canada, Australia, Hong Kong, and the United States. Throughout this process, I was inspired by many of the selfless and brave risks many cyber-security professionals took to help protect and secure networks and infrastructure, and to safeguard users. Their actions were often not done for monetary gain. Their passion for cyber security was undeniable. They worked in silence, with users and organizations, to protect systems all the while unaware of the efforts taken and the self-sacrifices made. And the risks they took were not always proportionate to the benefits gained. Many of these risks involved the uncertainty of legal sanction, whether it be criminal or civil lawsuits. Many of the technical and legal challenges for ethical hacking bare some similarities with hacking activities in general.

At the time, in 2010, there were few interviews or empirical studies on ethical hacking. The studies that existed were purely qualitative. Two of the most significant qualitative studies of ethical hackers were Dr. Suelette Dreyfus and Julian Assange's (2011) book, *Underground*, and Dr. Alexandra Samuel's (2004) PhD thesis on hacktivism and political participation. Dreyfus and Samuel were interviewed for the report.[2]

At the time of writing the report, there were only a few quantitative studies on hacking. The most prominent was a United Nations Interregional Crime and Justice Research Institute (UNICRI) study by Raoul Chiesa, Stefania Ducci, and Silvio Ciappi published in 2009 as *Profiling Hackers: The Science of Criminal Profiling as Applied to the World of Hacking*. The book provided a comprehensive look at hackers globally, using both qualitative and quantitative analysis. However, because hacktivism had not yet become popular there was no differentiation of hacking for political or social cause within the analysis. That is because most forms of ethical hacking, hacktivism, and online civil disobedience did not take flight until after 2011. While the authors were not able to be interviewed for my report, I gratefully borrowed some statistics and other information from *Profiling Hackers*.

While this book borrows heavily from the 2011 report for Public Safety Canada, it will deviate from it in two main ways. First, incidences from 2011 to 2018 are analyzed in this book, which means that new qualitative studies are referenced. Second, this chapter uses three new methodologies to provide a more comprehensive quantitative examination of ethical-hacking incidences around the globe.

In the first instance, we used the global database GDELT Analysis Service to explore incidences between 2013 and 2014. This methodology allowed us to examine incidences reported in a hundred different languages.

Second, we used standard SQL to customize the search queries using Google's BigQuery for the years 2015, 2016, and 2017, and then used the visualization features from Tableau software. This also allowed us to search through a hundred different languages of online media and blog reports of ethical hacking.

Lastly, we wrote python scripts to analyze data from cyber-jihad and hacking forums on the Dark Net to look for hacking incidences (contemplated, planned, or executed) that might have elements of ethical hacking. These data sets run from 2012 to 2016 for most of the hacking forums, while the cyber-jihad forums run from 2000 to 2012.

These methodologies are examined below and in greater detail in sections 3.3 to 3.5. The important findings from the different methodologies are summarized in section 3.2.

3.2 Summary of Findings

My original methodology for the report only allowed the capture of incidences that were reported online through manual searching in the media in English, French, and Russian. Adding GDLET and then SQL using BigQuery did not change the source of incident retrieval as these only pull information from reported online media and blog sites. These methodologies did, however, allow us to discover incidences reported in over a hundred languages, allowing for a picture of how ethical hacking was emerging globally.

The above methodologies are limited as they only allowed us to look at incidences that had been reported online by media and in blogs. This of course left a large gap in accuracy in determining how prevalent ethical hacking was becoming around the world. The opportunity arose where we were able to use publicly available Dark-Net data sets from the AZSecure-data.org involves an online portal that provides access to data collected on the Dark Net. The majority of the forum datasets have been collected by the University of Arizona's Artificial Intelligence Lab. While these data sets were only available in English, they revealed the magnitude and growth of ethical hacking in ways that traditional media analysis could not. Some of the important findings are summarized in table 1.

Future studies should run analytics on Dark Nets in languages other than English where data is available. Equally if not more important is the performance of data analytics for ethical-hacking discussions on social media, and in the hacker communication platform called Internet Relay Chat (IRC).

I have begun to run more data analytics on other Dark-Net forums, as well as in the IRC. For an updated look at ethical hacking, and to use interactive ethical-hacking maps, please visit www.ethicalhackinglaw.org/statistics or you can link to this through www.alanacybersecurity.com.

Table 1. Summary of Findings

Method	Coverage	Estimated Number of Incidences	Notable Differences and Observations
Online media sample (1999 until 2017)	1999–2012	137	English, French, and Russian media and blog postings.
GDLET	2013–2015	10,000	100+ languages, which revealed that ethical hacking was prevalent globally. Could only search with pre-selected terms.
SQL BIGQUERY	2015–2017	50,000	100+ languages which revealed that ethical hacking is prevalent globally. Could search with a variety of terms chosen by the researchers. There was an absence of incidences reported in Vietnam, Malaysia, Mexico and Brazil which may be due to heavy government influence of privatised media, censorship and fear of physical attack of journalists and bloggers. Countries with known heavy state censorship such as China, Iran and Saud Arabia reported many incidences of ethical hacking though these instances were consistently related to patriotic hacking.
Cyber-jihad Dark Net forums	2000–2012	43,000	We only looked a cyber-jihad forums in the English language. False positives were difficult to ascertain.

Method	Coverage	Estimated Number of Incidences	Notable Differences and Observations
Hacking Dark-Net forums	2012–2016	922,000	We could not fully clean the data due to size restrictions, and are therefore unable to isolate return searches that referred to the same incident. Our rates of false positive are unknown. We are likewise unable to provide positive predictive value. We analyzed one string of communications where there were four participants each responding approximately three times. If we performed this sample on the 922,000 the rate of single incidences would be closer to 77,000 separate incidences.

3.3 GDELT Analysis Service—Event Data (with Kevin Kim)

Our previous methodology only allowed us to view incidences reported in English-language media and blogs. With the GDELT research we could see a more global picture of ethical hacking, though the research revealed that there were many countries where no incidences of ethical hacking occurred. Possible reasons why, along with a more nuanced exploration of this research, is found below.

The GDELT Analysis Service is a free cloud-based database that allows you to explore, visualize, and export global event data. No technical expertise is required to use this service. This service does not allow the user to search with free text. The user must choose a theme or combination of themes. GDELT is an open database with approximately 1.5 billion geo-references within its data sets. The references are from media, other open data sets, and blogs across one hundred languages. The references run from January 1, 1979, to the present.

For our purposes, we selected the Events Data Heatmap to perform and visualize our research.

The next step required us to select a search theme or combination of themes (see fig. 2). Ethical hacking, hacktivism, and online civil disobedience were not identified as possible themes. The closest search parameters we could find were "cyber attack" in combination with "civil liberties."

We limited our searches to 2013 and 2015, receiving the data as a heat map with exported CSV (comma-separated values) file in Excel, where the geo-referencing details as well as links to the website or blog reporting the incident—see figure 3.

Each incident reveals the link to the media or blog source. The most prevalent online reporting occurred in the areas coloured red, de-escalating to orange, yellow, green, and blue.

Curiously, there are many countries where no incidences of ethical hacking occur. This indicates that there are some possible inconsistencies. We then compared the heat maps and timelines with *Freedom on the Net* reports from the non-governmental Freedom House. *Freedom on the Net* is the most widely utilized resource worldwide

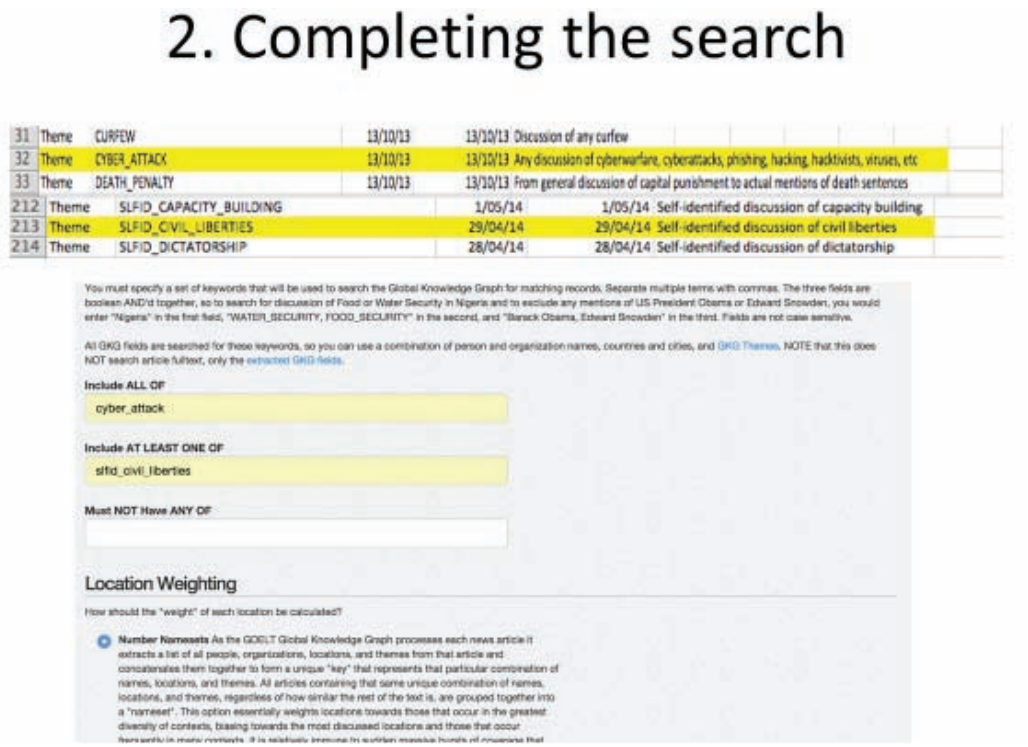

Figure 2. Search Themes.

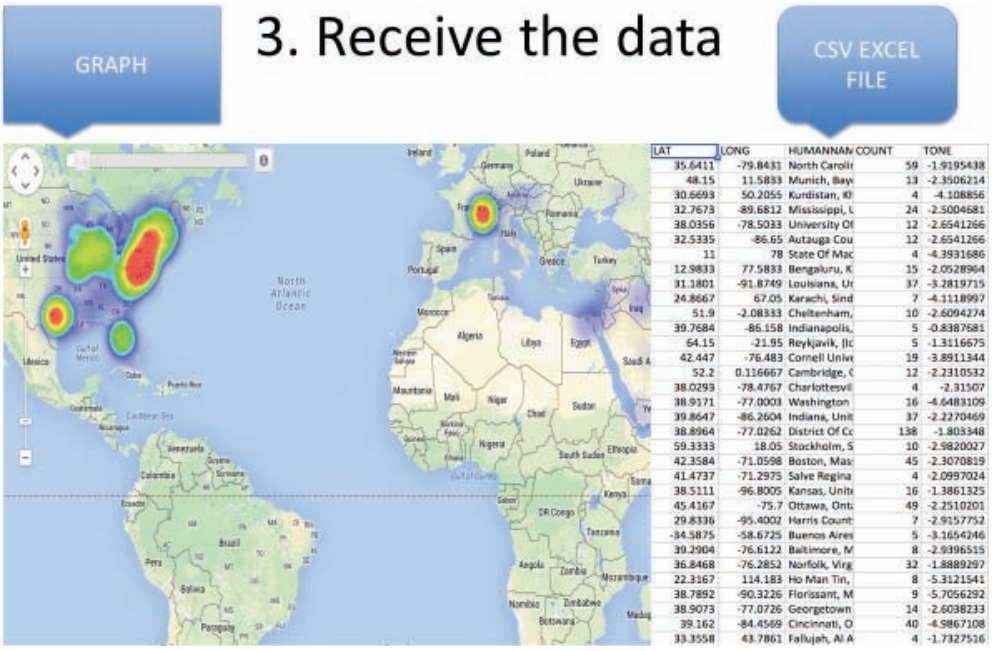

Figure 3. CSV Data Retrieval and Heat Map.

for activists, government officials, journalists, businesses, and international organizations seeking to understand the emerging threats and opportunities in the internet-freedom landscape globally, as well as regards policies and developments in individual countries. We focused on countries that were "cold" in the heat map and timeline. When we consulted the Freedom House reports looking for incidences of ethical hacking we could then speculate why these results were not appearing in our analytics. We looked at the general framework around journalism protections, censorship, and other aspects in Vietnam, Malaysia, Mexico, and Brazil that might affect why media and bloggers were not reporting incidences of ethical hacking.

According to the *Freedom on the Net* report for Vietnam in 2013 and 2014, there was extreme crackdown on freedom of expression, with several high-profile Internet writers and bloggers arrested and prosecuted. There were and are strong censorship laws coupled with strategic arrests and prosecutions for dissidence and political opposition. The same can be said for Malaysia, where there is also heavy media censorship and frequent arrests of journalists. Journalists in Brazil are regularly attacked by corrupt law enforcement and criminal organizations. News media is privately owned but relies

heavily on state advertising, which is said to lead to government manipulation of media. The same may be said of Mexico, where journalists and bloggers have been routinely attacked. The government also heavily subsidizes and advertises on the country's biggest media outlets, Televisia and TV Azteca.

Curiously, other countries with strict media censorship, such as China, still reported many incidences of ethical hacking (over 1,000 incidences appear for China from 2013 to 2014). Similarly, in Iran and Saudi Arabia, where there is also strict media control, there were still many incidences of ethical hacking that appeared in our data (Iran had 511 incidences and Saudi Arabia had 110). This discrepancy could be explained by the fact that these countries have excellent engineering and computer-science sectors, with many skilled and savvy computer users who would know how to use proxies and host blogs on sites that are not easily taken down. This discrepancy can also be explained by the deep levels of nationalism and patriotism within these countries. Studies of hacktivism in China[3] found strong correlations between hacktivism and patriotism, especially within the "red hacker" Honker Union, a Chinese group whose own code of conduct includes "Love your country. Strictly forbid attacks against any legitimate institutes within the country…. Uniformly defend the country and respond to defiant acts by foreign countries."[4] While patriotic hacking may not be condoned by the Chinese government, it also isn't censored in Chinese media or blogs.

3.4 Google's BigQuery (with Richard Li)

While the GDELT databases allowed us to capture a more global snapshot of ethical hacking, the process of running the analytics was very slow, requiring systems to run days to deliver basic analytics. Google's BigQuery allowed us to process higher volumes of data. We used the same terminology as we did in GDELT, and the same amount of languages were automatically translated. We captured nearly double the amount of incidences using BigQuery as explored below.

Below are incidences that were captured using a slightly different methodology and visualization. BigQuery is a data-analytics service that allows users to enter their own queries (i.e., not limited to set themes), export the data, and conduct analytics on the data using standard SQL. SQL allows complex search queries returning

near-real-time results, as opposed to GDELT, which only allowed searches with pre-determined themes.

The data in GDELT is open and free, but performing search queries using Google's BigQuery is not free. Due to limited funding, we were only able to perform analytics over 2015, 2016, and 2017.

We used the same terms—"cyber attacks" and "civil liberties"—as we had previously done. While we have results for 2015, 2016, and 2017, only the image for 2017 is displayed in figure 4. We were then able to use Tableau software to visualize the incidences.[5]

There were over 50,000 incidences of ethical hacking in 2017 alone. As we will see below, once data other than media and blogs were used, the incidences climb exponentially. Of course, the volume of incidences only shines light on the prevalence of ethical hacking. Understanding motivation, likely targets, the cause and effects of such occurrences can only be found through different data-mining techniques, and through qualitative research.

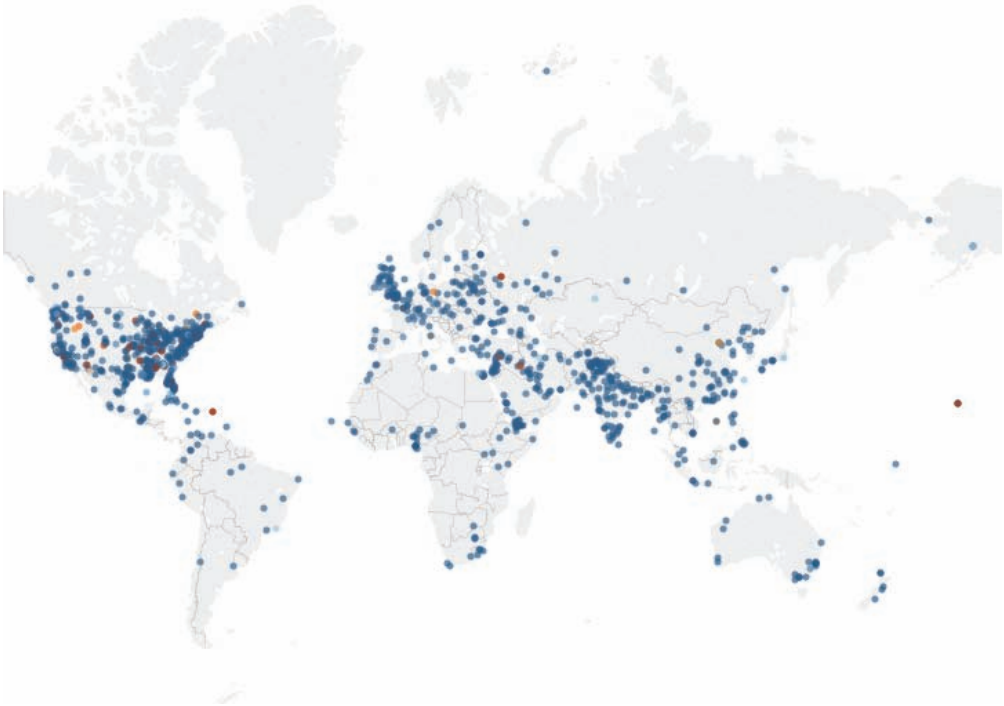

Figure 4. Ethical Hacking 2017.
Data retrieved and analyzed on May 4, 2017.

3.5 Dark-Net Analysis of Malware and Cyber-Jihad Forums

The data sets being used for this research are broadly categorized as the dark Web or Dark Net. Recall that the surface Web is the layer of the Internet that most people use on a daily basis—it involves using the World Wide Web protocol; it is also indexed, which means that you can use search engines such as Google to find content. The Deep Web is the next layer of the Internet where most data traffic occurs. It is not indexed by search engines; therefore, is unsearchable for the layperson via Google or Bing. One must go to a specific website in order to perform such a search. For example, I cannot access my medical-claim history in Australia through Google. I must first go to the Medibank site, enter my username and password, and then conduct a search. The dark Web or Dark Net is a subset of the Deep Web that can only be accessed by using encrypted services such as a TOR or VPN or both. It is the portion of the Internet where the darkest and most illegal activities occur. It is also not indexed. For our purposes, the dark Web as cited here refers to forum activity captured for the purposes of analyzing malware and the Jihadi social-media movement.

We ran analytics on many Dark-Net forums, which were categorized into two types: cyber-jihad and hacking/malware forums. Each of these are explained in greater detail below.

3.5.1 Cyber-Jihad Forums (with Adrian Agius)

We analyzed data that had been previously collected by the Data Infrastructure and Building Blocks (DIBB) program. The DIBB project is a collaboration between the University of Arizona, Drexel University, University of Virginia, University of Texas at Dallas, and the University of Utah. It is partly funded by the National Science Foundation, an independent US government agency supporting research in non-medical fields of science and engineering. The sets of data collected by the DIBB are forums, threads, and posts scattered across both the public-facing internet and the dark Web. DIBB provides open-source data (i.e., open-source intelligence information) for the intelligence and security informatics community.

For the purposes of streamlining the data processing and cleansing required for analysis, only forums that contained predominantly English-language postings were considered. English categorizations were provided by the DIBB, which greatly assisted in pre-determining

sets to be downloaded. Each of the data sets used for analysis was stored on a server operated by the DIBB. Each forum was allocated a text file and was directly downloaded into a local environment for processing.

The python script could not be easily run over the existing data sets without being further cleaned. According to data scientists Rahm and Hai Do:

> *Data cleaning*, also called *data cleansing* or *scrubbing*, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.[6]

Cleaning data is a laborious process. Dark-Web forum data is in an unstructured format, which means that one must clean and categorize the data to render it useful for analytics. In our case, the forum data was already partially cleaned by DIBB, but required further cleaning and categorization for our purposes.

Storing all the data collected into a single repository would make it easier to run any functions or queries required to gain insight into the data set as a whole. Thus, the first python script written to process the data was one which appended each set of forum data into a single file. Each file appended was tagged with its filename, as well as a broad categorization as either a "cyberterror"- or "geoweb"-themed forum. This will provide contextual relevance for later use.

Once collected and tagged appropriately, the consolidated file was then ingested to detect anomalies and remove any problematic lines of data. This included lines where special characters, including Arabic and other special keyboard characters, were present. As a blanket rule, any line of data that contained such characters were

deleted from the set of data. (Ideally, in the future we will run data analytics on Chinese-, Russian-, and Arabic-language forums.) We were able to delete entries that were duplicated in order to reduce false positives. Therefore, if the same pseudonym discussed an online ethical-hacking incident, this single incidence would be counted as one, as opposed to being counted as thirty separate incidences due to the pseudonym referencing the same incident thirty times.

The following structure prevailed.

MessageID—ID of forum post
ThreadID—ID of parent thread
ThreadName—Name of parent thread
MemberID—ID of posting member
MemberName—Name of posting member
Message—Content of forum post in question
Pyear—Year of post
Pmonth—Month of post
Pday—Day of post
Pdate—Aggregated field of above three times
ThreadFirstMessageID—ID of first message in thread
Forumname—Generated field, recording name, of originating forum
Classification—Generated field, describing type of message content

To give context to programs analyzing the content stored in the aggregated CSV file, the content contained in each forum post needed to be tokenized. Tokenization is the process of individually segmenting each word within a larger corpus. Tokenization is a fundamental part of natural language processing (NLP), which allows for computers to interpret language to perform various operations to generate insight about what is being said. Given the size of the data set at hand and the inability of a single analyst to traverse each entry, NLP provides for an effective way to analyze this data set.

Tokenizing the current data set required yet another script. Using the Natural Language Toolkit (commonly known as NLTK) developed out of Stanford University, tokenization is made possible. Given that the content of each forum post is what we are required to tokenize, the process of tokenization will only be applied to forum messages. At the conclusion of tokenization, a final field, "tokens,"

was added to the above-mentioned structure, storing the tokens associated with each message in a structure that preserves its relationship with the original table.

At the conclusion of processing, the working data set was approximately 7.43 gigabytes in size. However, for this analysis we were able to leverage search terms in order to condense to be more specific in our findings. The search terms we used to narrow the collected set were:

Ethical Hacking
Hacktivism
Anonymous
Cyber
DDoS
Lulzsec
Chaos Computer Club
Online
Hacking

The presence of these terms needed to be felt across either the ThreadName or MessageID fields in each post. These searchers resulted in a total data-set size of approximately 256 megabytes, which represents the subset of data used in the analysis below.

The following Dark-Net forums were cleaned and analyzed:

afghanForum
afghanForums
allsomaliforum
ansarl.txt
banadir24
Gawaher.txt
IslamicAwakening.txt
IslamicNetwork.txt
Itdarashag
Karbush
Myiwc.txt
Pastunforums
Somaliaonline
solamliUK
TurtoIslam.txt
Ummah.txt

Figure 5 looks at the frequency of instances in cyber-jihad forums where there are elements of ethical hacking from 2000 to 2012.

There were approximately 43,000 hits on terms related to "ethical hacking" from 2000 to 2012 in the cyber-jihad forums. Many of the conversations, however, had elements of ambivalence where the intended use of hacking remained unclear.

Figure 6 provides an example of content found in the various forums.

This analysis may indicate that cyber-jihad forums are more akin to traditional hacking forums in that they are more oriented around providing general advice and tutorials rather than traditional jihad forums, which focus more around discussing terrorist events and discussions around religious and political issues. As will be seen below, the analysis of hacking forums retrieved very different results from the cyber-jihad groups.

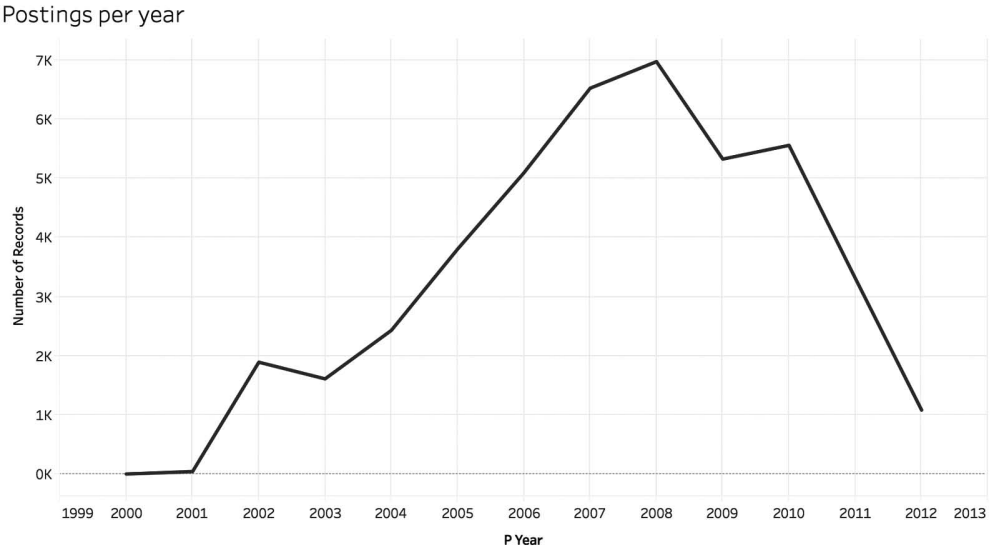

Figure 5. Ethical Hacking in Cyber-Jihad Forums.

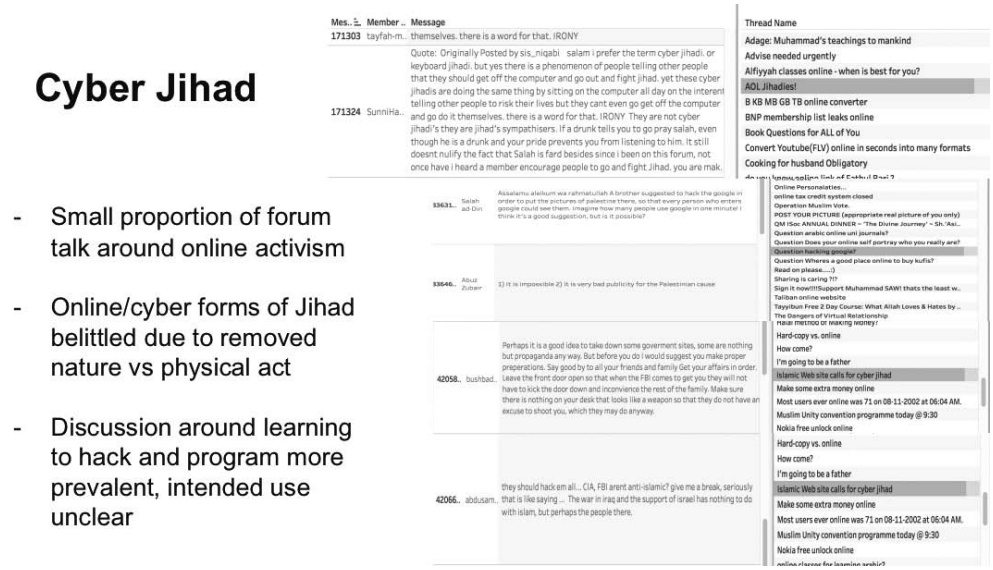

Figure 6. Example of Content Found in Forums.

3.5.2 Hacking Forums (with Richard Li)

For the Dark-Net analysis of hacking forums, we used dark-market data sets (forums) that had previously been scraped and, in most instances, cleansed and indexed. The AZSecure data sets were also from the DIBB and were further cleaned as per the same methodology that was used in the cyber-jihad forums. The forums analyzed were:

Dark-Net Market Archives
(https://www.gwern.net/DNM%20archives)

- Grams archive and select dark-market forums used

AZSecure Other Forums
(http://www.azsecure-data.org/other-forums.html)

- Only the English-language forum HackHound used

AZSecure Other Data
(http://www.azsecure-data.org/other-data.html)

- Only network traffic data and websites data used

We did not run analytics on real-time dark-market forum chatter as this was beyond our analytical skills, would require hundreds of hours cleaning and indexing the data, and, most importantly,

would require a secure facility/server to process the information, as many of the forums contain live malware.

Some of the forums required us to clean/scrub the data. Other forums were indexed and cleaned. In these instances, we changed the clean-up script to accept multiple file inputs, and added skipping blank lines and initial whites pace—see below.

```
import pandas as pd
import sys

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', -1)

reload(sys)
sys.setdefaultencoding('latin-1')

for line in sys.stdin:
        csvfile = pd.read_csv(line.rstrip(), error_bad_lines=False,
skipinitialspace=True, skip_blank_lines=True)
        csvfile.to_csv("datasetoutput.csv," mode='a', index=False)
```

We then put CSV files into the same directory with script and ran using `ls -p *.csv | python test.py`.

We changed the clean-up script to be recursive, to find and read all CSV files in current directory, then output as a single CSV file (adds date field based on directory name due to schema)—see below.

```
import pandas as pd
import os
import sys

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', -1)

reload(sys)
sys.setdefaultencoding('latin-1')

path = '.'
count = 0

for dirpath, dirnames, files in os.walk(path):
        for f in files:
                if f.endswith('.csv'):
                        #f is csv file name e.g., Valhalla.csv
                        #os.path.basename(dirpath) is directory name e.g.,
                        2016-04-17

                        filepath = os.path.join(dirpath, f)

                        df = pd.read_csv(filepath, error_bad_lines=False,
                        skipinitialspace=True, skip_blank_lines=True,
                        quoting=3)
                        df['date'] = os.path.basename(dirpath)

                        if (count == 0):
                                count += 1
                                df.to_csv("datasetoutput.csv," mode='a',
                                index=False)
                        else:
                                df.to_csv("datasetoutput.csv," mode='a',
                                header=False, index=False)
```

We only wanted data on forum posts to search for keywords—that is, scrapped forum threads.

We initially patterned extracted forum archives using the command:

```
tar -zxf [forumname].tar.xz --include='[pattern]'
```

(the pattern differed depending on the forum)

The resulting scraped html of forum threads was then parsed by a python script into a CSV file. The python script only outputs data where either the title or post content matched the search terms. To speed up analysis, the actual content of a post is not written to the CSV file after being searched for matching terms.

We open the resulting CSV file (e.g., "datasetoutput.csv") with Tableau, which allowed us to visualize our analytics. We then created a calculated field for content to search; for example, thread title and post content. Afterward, we created a filter for that field with a condition that searches for specified keywords within content.

- Search terms: Ethical Hacking, Hacktivism, Anonymous, Cyber, DDoS, LulzSec, Chaos Computer Club, Online, Hacking.
- Additional search terms: AntiSec, Anti-Sec, CyberBerkut, Ethical Hacker, Hacktivist, Iranian Cyber Army, Syrian Electronic Army, White Hat.

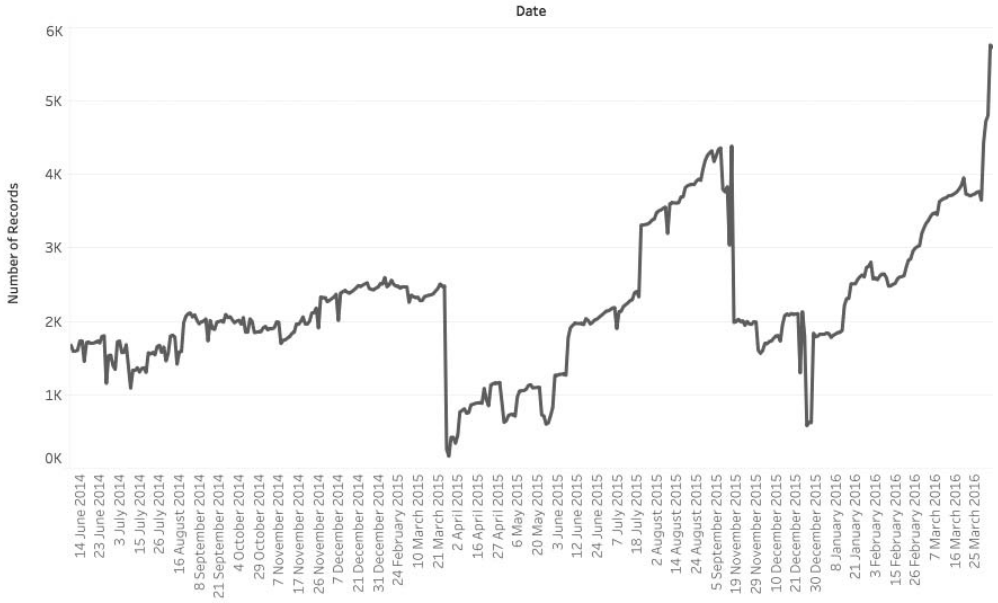

Figure 7. Grams Listings 2014–2016.

Figures 7 and 8 further show the numbers involved with aspects notable to ethical hacking.

The numbers from the Grams data reveal over 922,000 references to ethical hacking. These numbers are further broken down per market below.

Forum data from HackHound retrieved much smaller results than from the Grams data. We used the same-directory CSV-file methodology, and we also had to manually clean up and delete html content that overran a cell. As illustrated in figure 9, we had a total match of 198.

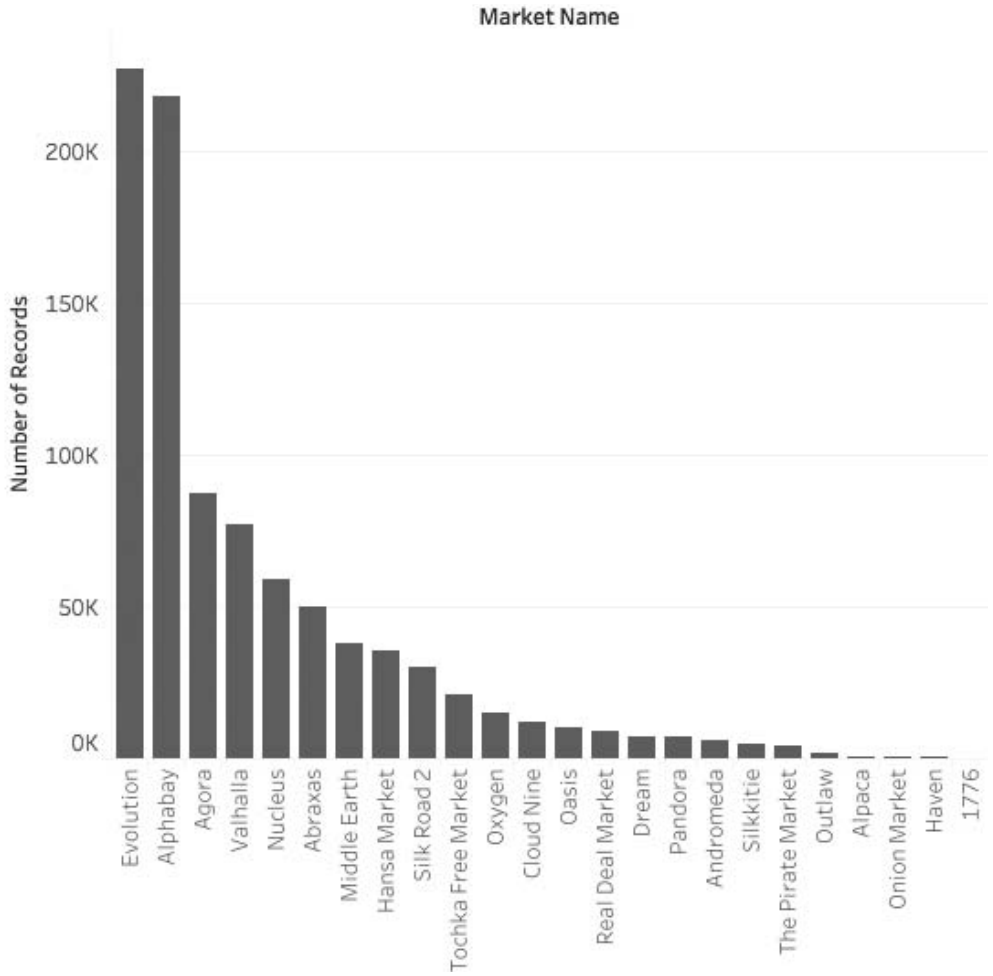

Figure 8. Grams Listings per Market.

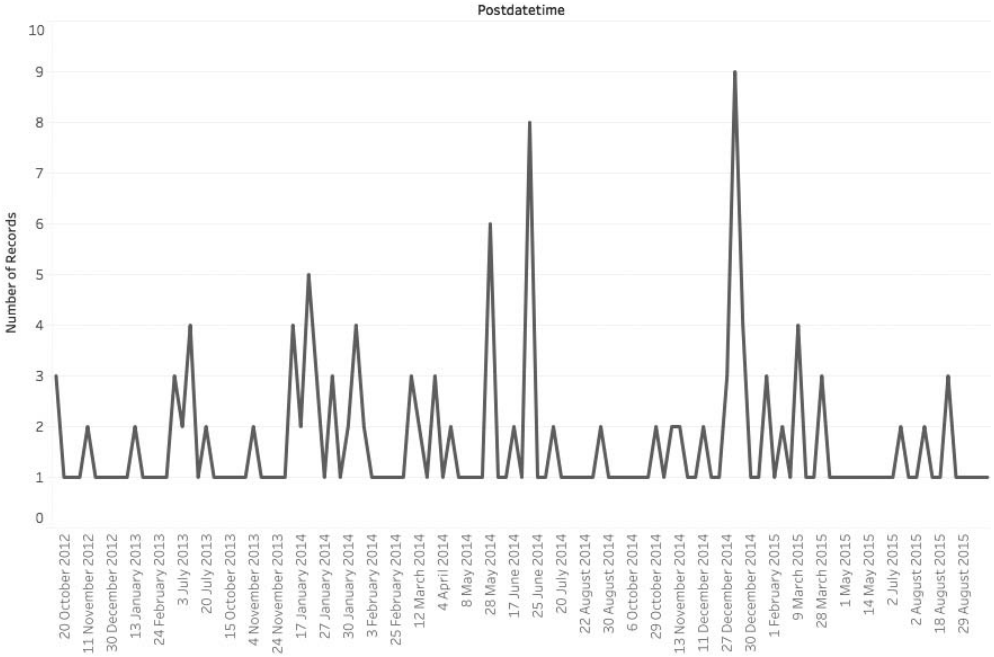

Figure 9. HackHound Forum Incidences 2012–2016.

3.6 Observations

We made use of the Grams archive that had CSV data for dark-market listings over several dark markets (https://www.gwern.net/DNM%20archives#grams). From there we used multiple-directory CSV-file methodology, resulting in total matches of 922,649. These indicate records, and not separate incidences; therefore, we were unable to cleanse the data to the point where we could infer a record like we were able to when using the cyber-jihad forum data. Due to the other forums being smaller in size, we were able to clean that data to make records of single events related to ethical hacking. This is clearly not an ideal methodology. In fact, the methodology has been developed on a rather ad hoc basis from 2011 to 2017. We have not received any research funding for this project; we have merely done what was within our grasp at the time based on the resources and skills available from students interning with the Cyberspace Law and Policy Centre of the University of New South Wales, along with myself. However, the numbers say something rather significant—that ethical hacking is occurring globally and that it is escalating as a means of both political and social protest.

Our traditional research methodology for the 2011 report included approximately 137 different incidences, while our research

using GTELD and SQL BigData query resulted in approximately 50,000 ethical-hacking incidences. With the cyber-jihad forums alone, we retrieved 43,000 records. With the hacking forums, we retrieved close to 922,000 references to ethical-hacking terms. While the 922,000 number does not isolate incidences/records, it is safe to infer that the numbers would be significantly greater than the few hundred hits and the 50,000 hits when the searching was limited to online reported incidences through media and blogs.

Future research would also run analytics on social-media platforms, such as Twitter and Facebook, and popular social-media platforms in other countries, such as in Russia and China (WhatsApp for China, VK for Russia). Indeed, language skills have been a limiting factor to the analytics for dark-market analysis. The GTELD and BigQuery analysis has built-in translation services that provide results in multiple languages, which is a significant advantage. Future research would also be funded and performed with experts in data science.

Methodologies reliant on using manual search queries or data analytics running on media and blogs produce results that can be contextualized. However, they also only pick up a slight portion of the number of ethical-hacking incidences occurring globally, as can be seen with the data analytics run on the dark-market forums. Chapters 4 through 8 look at selected ethical hacking.

Notes

1. A. Maurushat, *Ethical Hacking*. A report for Public Safety Canada (2011).
2. The interview questions appear in the appendix below.
3. Yip and Webber 2011.
4. Honker Union of China.
5. These maps are not interactive in the book. However, the maps are interactive for users on the website http://www.ethicalhackinglaw.org/statistics. You can also link to the ethical-hacking website and information from www.alanacybersecurity.com. On these websites, you can right-click on a blue dot which represents incidences with a link to that geographic location. Right-clicking will provide direct links to incidences reported in media and blogs. A single blue dot could represent hundreds of incidences, or only one. Clicking on the dot, therefore, is important in ascertaining a more accurate picture of the number and source variety of incidences.
6. Rahm and Hai Do 2009.