ANALYZING GLOBAL CYBER ATTACK CORRELATES THROUGH AN OPEN DATABASE

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Brady Aiello

June 2018

COMMITTEE MEMBERSHIP

TITLE:	Analyzing Global Cyber Attack Correlates Through an Open Database
AUTHOR:	Brady Aiello
DATE SUBMITTED:	June 2018
COMMITTEE CHAIR:	Bruce Debruhl, Ph.D. Associate Professor of Computer Science
COMMITTEE MEMBER:	Foaad Khosmood, Ph.D. Associate Professor of Computer Science
COMMITTEE MEMBER:	Franz Kurfess, Ph.D. Professor of Computer Science

ABSTRACT

Analyzing Global Cyber Attack Correlates Through an Open Database

Brady Aiello

As humanity becomes more reliant on digital storage and communication for every aspect of life, cyber attacks pose a growing threat. However, cyber attacks are generally understood as individual incidents reported in technological circles, sometimes tied to a particular vulnerability. They are not generally understood through the macroscopic lens of statistical analysis spanning years over several countries and sectors, leaving researchers largely ignorant of the larger trends and correlates between attacks. This is large part due to the lack of a coherent and open database of prominent attacks. Most data about cyber attacks has been captured using a repository of common vulnerabilities and exposures (CVE's), and "honey pots", unsecured internet-connected devices which record attacks as they occur against them. These approaches help in the process of identifying vulnerabilities, but they do not capture the real world impact these attacks achieve. Therefore, in this thesis I create a database of 4,000 cyber attacks using a semi-open data source, and perform analytical queries on it to gather insights into how cyber attack volume varies among countries and sectors, and the correlates of cyber attack victims. From here, it is also possible to relate socio-economic data such as GDP and World Happiness Index to cyber attack volume. The end result is an open database of cyber attacks that allows researchers to understand the larger underlying forces which propel cyber attacks.

ACKNOWLEDGMENTS

I would like to thank Cal Poly for the opportunity to study computer science, and all of their generous scholarships which made this work possible. I would also like to thank Cisco for funding our research of security in the Internet of Things. Thank you to my committee. Professors Khosmood and Kurfess gave me helpful feedback that has made this work much stronger. Thank you to my advisor, Professor DeBruhl, who has helped me find direction and focus on what is important in this thesis. Thank you to Paolo Passeri, whose data is the backbone of this research. Without your meticulously assembled open data, none of this work is possible. Thank you to my parents who encouraged me to pursue my dreams. And lastly, thank you to my wife Erin and my son, Khaver for bearing with the difficult demands of our higher education for so many years. You are the best family anyone could hope to have.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

According to a 2017 study by Cybersecurity Ventures:

> Cybercrime is the greatest threat to every company in the world, and one of the biggest problems with mankind. The impact on society is reflected in the numbers.
>
> Last year, Cybersecurity Ventures predicted that cybercrime will cost the world $6 trillion annually by 2021, up from $3 trillion in 2015. This represents the greatest transfer of economic wealth in history, risks the incentives for innovation and investment, and will be more profitable than the global trade of all major illegal drugs combined [57].

According to a study by Hiscox Insurance on cyber attacks in 2017 across the US, Great Britain, Spain, Germany, and the Netherlands, involving 4,103 organizations in private and public sectors, the average annual cost due to cyber attacks to a single business is $229,000, and %73 of organizations were not prepared for a cyber attack [47]. For organizations with more than 1,000 employees, the average annual cost of total cyber incidents was $356,000 in Spain, and $1.05M in the US [47]. The largest firms in the US lost $25M annually to cyber crime, while the largest firms in the Great Britain and Germany lost $20M [47]. Businesses with fewer than 100 employees lost between $24,000 on average in Spain to $63,000 on average in Germany [47]. The greatest cost for a single incident ranged from $800,000 in Spain to $5M in Germany, and the US in the middle at $2M [47]. Cyber crime presents a substantial and growing threat to both the private and public sector.

Cyber security incidents also cost individual consumers. The WannaCry attack

of May 2017 affected 300,000 machines, and cost $4B [47]. The attack, carried out by North Korea, disabled the machines of users all around the world, promising to free them for a ransom, an instance known as "ransomware" [34]. The attacks hit hospitals in Great Britain especially hard, impeding medical work, and risking lives [34]. In September 2017, Equifax experienced a data breach exposing the information of 145 million Americans, and some citizens of Great Britain and Canada, exposing their Social Security numbers, dates of birth, driver's license numbers, driver's license dates and states, home addresses, and credit card numbers [29].

As computers become more ingrained in human life, cyber attacks even have the capacity to kill. In 2017, a pacemaker model implanted in 465,000 people was recalled over concerns about holes in its security that could let a hacker drain the battery or alter the heart's rhythm arbitrarily [45]. In July 2017, a security researcher demonstrated how a popular internet-connected car wash system could be commandeered from anywhere in the world to attack anyone inside the car wash [65]. Full administrator privileges were granted by entering the default password "12345" [65].

Car washes are one example of large equipment connected to the internet. However, in industrial settings, there are many types of internet-connected physical equipment which may be exposed to cyber criminals. The rise of industrial attacks puts the safety of an entire population at economic and physical risk. In December 2015, the Ivano-Frankivsk region of Western Ukraine experienced a loss of power after cyber attackers remotely took control of the cursors of employees at 3 power stations, changed their passwords, shut off power to every region they managed, disabled backup power, reset employee passwords, leaving 230,000 residents without power [68]. Between 2005 - 2010, a worm was discovered on computers and industrial programmable logic controllers (PLC's) which manage Iran's nuclear program [31]. The join effort between the US and Israel degraded Iran's nuclear centrifuges was highly sophisticated, only targeted machines related to the nuclear program, faked sensor data on the machines,

destroyed 1/5 of its nuclear centrifuges by spinning them out of control, and went undetected for 5 years [31] [67]. Politically motivated attacks against public infrastructure pose an imminent economic and physical threat to all people.

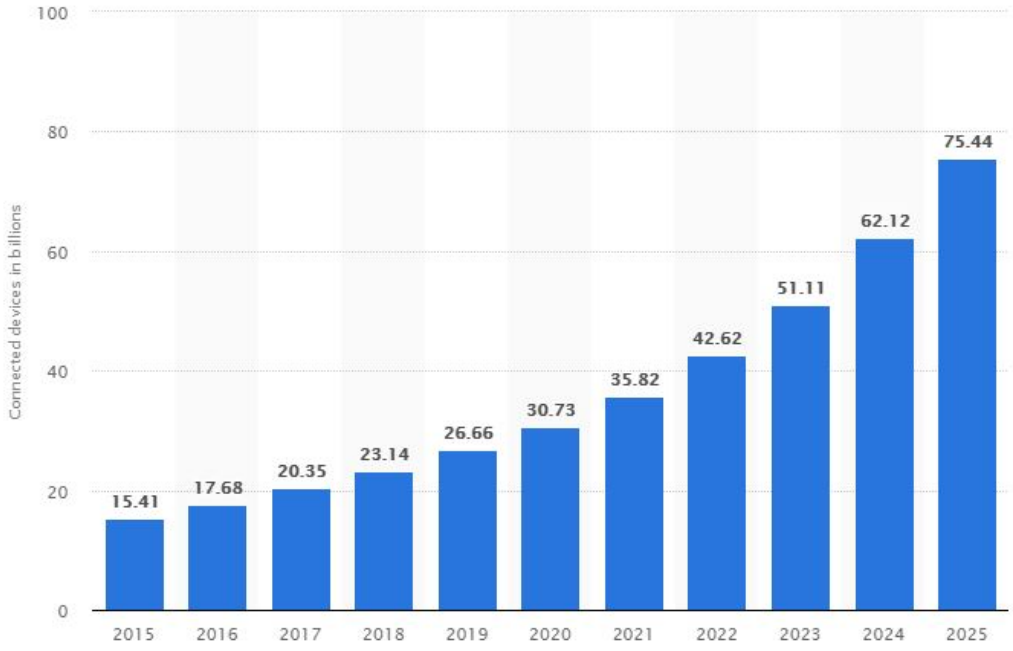


Figure 1.1: Projection of Worldwide Total M2M devices, created by Statista.

Increased internet connections and traffic will expose more machines to cyber attacks. Cisco explains 4 reasons that network usage are likely to become significantly larger in the near future: an increase in internet users, more machine-to-machine (M2M) devices, faster internet speeds, and increased video viewing. From 2016 - 2021 users are projected to increase from 3.3B - 4.6B, M2M devices will increase 17.1B - 27.1B, average broadband will increase from 27.5 Mbps - 53.0 Mbps, and video viewing will increase from 73% - 82% of total traffic [21]. All of these advances expose more people and more machines to cyber threats. The increase in IoT (M2M) devices, as estimated by Statista [19], and shown in figure 1.1, is particularly concerning, because these are often unsecure. According to a 2017 study by the Ponemon Institute, 80%

of IoT applications are not tested for vulnerabilities [41]. This is understandable, as an influx of countless cheap IoT devices streaming from many manufacturers means that there is not much funding available for security testing, and that there is a litany of development platforms of varying security.

It is clear that Cyber crime is a growing and vastly underestimated threat. As the world puts its businesses, personal lives, personally identifiable information, and credit card and banking information online, cyber crime is more likely to affect everyone, not only businesses and governments. Given the growing prevalence of cyber attacks, it behooves all internet users to understand the large historical trends of previous cyber attacks, in the hope of predicting future cyber attack behavior.

Unfortunately, there is a dearth of open and organized knowledge about cyber attacks. The current data sources are discussed at length in chapter 2, Background. In this thesis, I assemble a database of cyber attacks 2014-2017 without paying for any data access, by using a semi-open data source. I then make various types of queries against the database to determine how cyber attack volume changes over time, what socioeconomic factors exacerbate cyber attacks, and what other unknown facets of cyber attacks can be discovered. The main results of these questions are as follows. From 2014-2017, global cyber attack volume peaks August 2016, and the United States reports the most cyber attacks received. Before August 2016, attacks against the private sector and governments dominated; after August 2016, attacks against individuals started to increase, and now are the most targeted victims. Political and economic cycles play an important role in cyber attack volume. Cyber attacks sometimes peak 3 months prior to a national election, while others peak directly following an election, as a means of protest. Wealthier countries receive more attacks, and in the case of Great Britain, cyber attacks usually peak in November each year. One interesting discovery is that, following the peak of cyber attacks in August 2016, by October 2016, cyber attacks were less than half of the peak, and by December 2016

they approached levels similar to the maximum. Strangely, it seems that after a huge push for cyber attacks, hackers take a break for 2 months, subsequently returning to business as usual.

Chapter 2

BACKGROUND

In this chapter I discuss previous works similar to this thesis, what they teach, and how my thesis stands apart from rest. Most previous work on global trends in cyber attacks does not suffice, as it falls into one of the following categories: it discusses only a subset of cyber attacks, is not statistically rigorous, or does not use an open data source. No previous work has generated a cyber attack database for public use as mine does.

In Debeck's, 2011 MS cybersecurity thesis, "The Correlates of Cyber Warfare: A Database for the Modern Era", he discusses the general ignorance and disorganization of data pertaining to cyber attacks, and proposes a worldwide network of routers to track attacks as they occur [37]. This hypothetical system would provide perfect knowledge of the country of origin of any attack, tracking the attack as it passes through various routers. Debeck's motivation for such a tremendous undertaking is the hope of correlating cyber attacks with political, cultural, and socioeconomic conditions in a way analogous to "The Correlates of War" project which does the same for natural war [1]. This work's emphasis on a general void of reliable and open cyber attack data, which holds true 7 years after its release, was the original impetus for this thesis. Unlike Debeck, in this thesis, the cyber attack database I create is not a hypothetical one, but actually created from cyber attack data.

Ghandi's 2011 IEEE article, "Dimensions of Cyber-Attacks: Culture, Social, Economic, and Political" identifies a few dozen cyber attacks that are directly linked to socioeconomic events [42]. By scouring news articles, and performing in depth research on the attacks and their motivations, Ghandi makes the case that socioeconomic events are a substantial motivator of cyber attacks, and that their crucial

role in cyber attacks requires further research. However, Ghandi's does not statistically correlate any socioeconomic factors with cyber attack volume; it is more of a narrative that connects socioeconomic and political headlines with cyber security breaches. Ghandi states that, "the movements of Russian troops in Georgia were correlated with cyber-attacks on the Georgian communications infrastructure and defacement of government websites." This is believable given the evidence presented, but there is no mathematically derived correlation, which is generally a Pearson-R correlation. Again, Ghandi notes, "[a lack of cyber attacks] seems to reinforce the notion that attacks are strongly positively correlated to political and cultural conflicts", without deriving the actual correlation. In general, the work is believable, but of small scope, lacking in statistical rigor, and mostly anecdotal. In contrast, this thesis rigorously quantifies cyber attack volume for various countries, target sectors, types of attacks, and the various correlations between types of cyber attacks, as well with socioeconomic data.

Sharma's 2013 work, "A Social Dimensional Cyber Threat Model With Formal Concept Analysis and Fact-Proposition Inference", the authors describe how to use the socioeconomic conditions surrounding cyber attacks to server the construction of a decision tree that can characterize and predict cyber attacks using fact proposition inference [63]. In the fact proposition inference model used in this work, facts are inferred from propositions via a Bayesian belief network, where propositions are analogous to antecedents, and facts are analogous to consequents. Though Sharma focuses on the ontology of socially-motivated attacks, he claims tangible results from unnamed cyber attack data sources. For instance, the consequence "Information and Data Loss", the victims "Individuals/Civilians", the means, "Penetration Attempt", and the technological aspects, "SQL and Code Injection" all represent the highest beliefs in their category in the belief network. Unfortunately, the web app created for this paper is not available, the belief network is never shown in its entirety, and

the cyber attacks data source is never revealed. All of these factors make the results difficult to use or rely upon. In contrast, in this thesis I reveal my data source, and open source my work for others to reproduce.

In her 2014 paper, "Global Mapping of Cyber Attacks" Carley does much of what I do in this thesis [33]. She quantifies types of attacks, identifies which countries receive the most cyber attacks, and correlates volume of cyber attacks received with various socioeconomic data. Carley's analysis is truly an impressive work, and it relies upon her previous work of simply creating a database of cyber attacks [56]. Unfortunately, Carley's work relies upon closed data. The attack signatures are freely available, but the number of Symantec machines who have experienced an attack, and the IP addresses of those machines, which makes identifying the nation of any particular machine affected possible, are all closed source. This was only possible because Carley and her fellow researchers were granted special access to Symantec's World Intelligence Network Environment (WINE) Intrusion Prevention System's (IPS) telemetry data. This real world data sampled from over 10 million machines around the world no doubt allows for the highest quality of data, but it is infeasible for most researchers. Carley also does not analyze temporal changes in attack data at all, and only analyzes attacks from November 2009 - September 2011. In contrast, I only use open data, I analyze temporal and seasonal trends in different types of cyber attacks and their victims, and I analyze cyber attacks from January 2014 - February 2018, a longer and more recent history.

Aviles' 2015 master's thesis in cybersecurity, "How US Political and Socio-economic Trends Promotes Hacktivist Activity" tries to understand hacktivism, among other things, by analyzing cyber attack reports from Hackmageddon [25] [60]. Unfortunately, the work is very anecdotal, and only examines cyber activity in December 2014 and January 2015. The only charts and results obtained are mere reproductions from Hackmageddon [60]. This work is not statistically rigorous, but a qualitative

approach to seeing cyber attacks through a socioeconomic lens. I also use Aviles' data source. However, in this thesis I perform rigorous statistical analysis.

Kumar's 2016 research paper, "DDoS Cyber-Attacks Network: Who's Attacking Whom?", analyzes global trends in DDoS attack behavior [51]. DDoS, or Distributed Denial of Service, is a type of attack in which a machine is overwhelmed by the network requests made by thousands of machines in different geographical locations, causing the victim machine to become unavailable. Using open DDoS data from Digital Attack Map [6], Kumar discovers that, from May 2013 - March 2016, the top 10 victims of DDoS in descending order are the US, China, Peru, France, Canada, Poland, Great Britain, Brazil, Germany, and South Korea. The top 10 DDoS aggressors in descending order are China, the US, the Netherlands, Germany, South Korea, Brazil, Great Britain, Russia, France, and Turkey. Kumar notes that the country of the attacker is only known for about $\frac{1}{3}$ of the attacks. It is similar to what I achieve in examining international relationships with respect to cyber attacks. However, as this work only examines DDoS attacks, one of many types of attacks, it is more limited in scope, than my own. I examine all types of cyber attacks, and I analyze many other aspects of cyber attacks than the aggressor country and victim country.

Solano's 2017 IEEE conference paper, "Socio-economic Factors in Cybercrime" progresses Ghandi's work with much more statistical rigor, correlating cyber attack incidents with 32 socioeconomic factors, including GDP PPA, unemployment, political stability, freedom of press, happiness, access to broadband internet, population, and life expectancy [64]. The cyber attack incidents are derived mainly from reports from `www.hackmageddon.com` [60], and socioeconomic sources are derived from the World Bank [15], the International Labour Organization [10], Freedom House [8], Polity IV [54], Reporters Without Borders [14], Transparency International [13], Credendo Group [4], and The Economist Intelligence Unit [3]. Solano's work is one of the more complete studies on the socioeconomic correlates of cyber crime to date,

but all results that are not extremely strong correlations are discarded. Solano believes that only 3 correlations, all greater than 0.80 and with p-values all less than 0.05, are important enough to discuss. Discussion of interpretation of Pearson correlations and p-values is in section 3.7. Solano notes 2 other statistically significant results with correlation greater than 0.60 disparagingly. This is a great disservice, as Pearson correlations greater than 0.50 are considered strongly correlated [36]. It is therefore very likely that Solano obtained many important results which will never see the light of day. Solano finds that in Syria, the correlation between political risk and the number of attacks received is ($r = 0.864$, $p = 0.012$), in Mexico, the correlation between lending interest rates and attacks received is ($r = 0.840$, $p = 0.036$), and that in Ghana, the correlation between GDP and attacks received is ($r = 0.821$, $p = 0.045$). Additionally, the correlation between the perception of corruption of Russia and attacks on Australia to be ($r = 0.642$, $p = 0.025$), and the correlation between the perception of corruption of Lithuania and Australia to be ($r = 0.639$), with a "similar" p-value. The authors of this paper are dissatisfied with the results, concluding, it "has not been able to find definite correlations between security incidents and the socio-economic variable of the countries involved", citing the under reporting of cyber attack incidents as a likely source of error [64]. By contrast, in this thesis, I reveal all of my results, including many in which there is no statistically significant correlation between cyber attack features. This is scientifically honest, and allows other researchers to replicate the results.

Chapter 3

THE CYBER ATTACK DATABASE IMPLEMENTATION

To organize cyber attack data in a way that is structured, persistent, and easily accessible from a single endpoint, an RDBMS is selected. This SQL solution allows researchers around the world to apply data science to cyber attack trends. All other options are inappropriate. Keeping data in spreadsheets or CSV's potentially splits data, does not allow complex queries, and does not enforce any rules for field values. Storage in document-oriented NOSQL databases such as Dynamo, Mongo, or Redis would not take advantage of any structure inherent in common data fields. A graph database such as Neo4j would allow for more complex connections made between attacks, but development time would be much longer, and because graph databases like Neo4j require learning a niche query language, the end result would not be accessible to the vast majority of cybersecurity researchers.

The implementation of a cyber attacks database is contingent on the available data. Finding reliable structured data on attacks can be very difficult, but we show how it can be done. Then the data is conditioned to fill in missing cells, and make text values more symmetric. Next, the data is transformed to a SQL-friendly version. Lastly, the SQL database is created and populated with the cleaned data.

3.1 Open Data on Cyber Attacks

Currently there are no viable open data sources of cyber attacks [37]. This makes sense because of the following. Private corporations study cyber attack data for sale, so they aren't motivated to maintain an open database of attacks. Government agencies also have their own private data stores on security incidents, but governments secu-

rity agencies also want to conceal what they know, and could end up exposing their own state-sponsored attacks. So they are also motivated to keep data secret. Most individuals do not have the time or resources to maintain a database of cyber-attacks, which is constantly growing. Hence, most researchers don't have access to data regarding the trends in cyber attacks. Past research has required meticulously collected articles from security-oriented news sites, such as in Gandhi's "Dimensions of Cyber-Attacks"[42]. Other research, such as Kumar's "Cyber Attacks DDoS Network"[51], has relied on `www.ddosmap.com` as a data source for understanding DDoS attacks. Still other research, such as Solano's "Socio-economic factors in cybercrime" [64] and Avile's "How US Political and Socio-economic Trends Promotes Hacktivist Activity" [25], has used the data from `www.hackmageddon.com` to analyze larger cyber attack trends. This last option covers many different types of attacks (not only DDoS), is structured, contains enough metadata to make detailed analysis, is semi-open, and saves researchers the trouble of handpicking cyber attack incidents. Therefore, it is the sole cyber attack data source used in this thesis.

3.2 Hackmageddon

The site "Hackmageddon" is the labor of security professional Paolo Passeri [59]. Passeri manually collects the cyber attacks data from following many security news sites, and uploads graphs and charts of them which he uses in a security blog. Users of Hackmageddon can submit an attack, which Passeri will review, and may include in the data at his discretion. To gain access to the original data used on the security blog, this author needed to contact Passeri personally. Thankfully, Passeri freely opens the data to anyone who asks.

The data is in the form of Excel Spreadsheets, a file for each half of a month, with data from 2011 to present (March 2018). Excel is Passeri's primary way of storing and

analyzing attacks. The fields are as follows: ID, Date, Author, Target, Description, Attack, Target Class, Attack Class, Country, Link, Tags.

1. **ID**: An integer unique to that half-month file
2. **Date**: The date the attack was reported
3. **Attack**: The type of attack (e.g. DDoS)
4. **Target Class**: The sector targeted.
5. **Attack Class**: Cyber War, Cyber Crime, Cyber Espionage, or Hacktivism
6. **Country**: ISO alpha-2 country code (e.g. US)
7. **Link**: A URL for a new article describing the attack
8. **Tags**: Important keywords

Often Author and Target are missing, and for older spreadsheets the Author field is actually a picture of the country, or hacker organization. For example, if Anonymous is behind an attack, a .png of a Guy Fawkes mask is often placed in the Author field.When the Author field is known, the picture must be replaced with the Author in plain text. Additionally, the Author and Target fields are sometimes empty. If the Author or Target is made known in the Description field, it must be filled. Otherwise, the field should have "NULL" as its value. Fields that are left empty, or filled with a "?" must all be filled with "NULL" as their values. Because a single spreadsheet holds only half of that month's cyber attacks, they must also be combined into a larger spreadsheet to make insights into the higher level attack trends. Filling in missing data and making field entries consistent requires considerable work, and inspecting every entry for an entire year, approximately 1000 entries, takes about a day. Examples of "Attack" include "DDoS" and "Account Hijacking". Passeri has

Field	Example
Id	16
Date	17-10-2016
Author	Guccifer2.0
Target	Democratic National Committee(DNC)
Description	Guccifer 2.0 is back and leaks new fresh documents relating to the US political system (documents allegedly showing email conversations between DNC employees and Hillary Clinton's presidential campaign staff discussing Donald Trump's position on his tax returns).
Attack	Unknown
Target Class	Org:Political Party
Attack Class	CC
Country	US
Link	http://www.ibtimes.co.uk/hacker-guccifer-2-0-leaks-files-claiming-dnc-researched-donald-trumps-taxes-1587073
Tags	Guccifer 2.0, Democratic National Committee, DNC, Hillary Clinton, Donald Trump

Table 3.1: The fields of the "attacks" spreadsheets used by Hackmageddon, and a sample row.

defined 4 abbreviations for Attack Class: CC (Cyber Crime), CE (Cyber Espionage), CW (Cyber War), and H (Hackitivism). Sometimes this field has the full name of the Attack Class, and sometimes many attack classes are listed. The "Country" field refers to the victim country. This is easily determined in instances of cyber war, or

when a large group of consumer that mostly or solely live in a certain country are targeted, though some attacks target a few or many countries. In these cases, the "Country" either lists the ISO alpha-2 country codes of 2-3 countries delimited by spaces or commas, or may simply read, "> 1".

3.3 Making the Spreadsheets SQL-Friendly

Using the raw attack data contained in separate spreadsheets proves unwieldy to deep data analysis. This problem is not assuaged even after they compiled into larger records. So, they are entered into a MySQL database. There are some issues that must be rectified first. The "Id" field is only unique across one spreadsheet (half a month of attacks), and must be unique across the entire database. The "Date" field is a reserved word in SQL so it is changed to "AttackDate". The entries are in the European format, "d/m/yyyy", instead of the standard SQL format, "yyyy-mm-dd". The "Target Class" and "Attack Class" fields contain spaces. So these all need to be changed to "TargetClass" and "AttackClass". "Country" is also ambiguous, as a user may interpret this as the author's country or the victims' country. To remove the ambiguity, this is changed to "TargetCountry". Another consideration is that though the "TargetCountry" field may reference another table of country data, sometimes Passeri's attack entries have more than one country per entry. Therefore, the "TargetCountry" field cannot be queried by a simple "GROUP BY" query, and cannot be a foreign key to a countries database. However, this is not a major problem, because only a few countries experience the majority of attacks, and general SQL commands are sufficient for seeing major attack trends and correlations. These insights can be used for more targeted SQL queries that get all attacks where "TargetCountry" contains "US", such as "US, UK". The last step is to convert the Excel spreadsheet to a comma-separated value (CSV) file in UTF-8 format. After these changes are made,

the fields in Passeri's spreadsheets map directly to a MySQL schema, which shall now be described.

3.4 Creating the Cyber Attack Database

After all addressing all the necessary changes to the schema, the creation of the table is as follows:

```
CREATE TABLE attacks(
    Id INT(11) PRIMARY KEY AUTO_INCREMENT,
    AttackDate DATE,
    Author VARCHAR(96),
    Target VARCHAR(256),
    Description VARCHAR(512),
    Attack VARCHAR(64),
    TargetClass VARCHAR(72),
    AttackClass VARCHAR(16),
    TargetCountry VARCHAR(32),
    Link VARCHAR(256),
    Tags VARCHAR(512)
);
```

The table 3.2 shows the schema and a sample row.

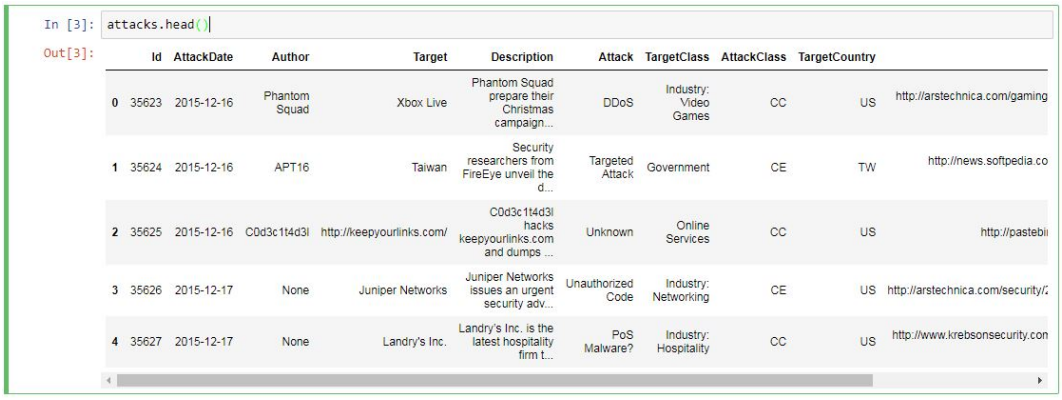
3.5 Inserting The Data

After creating the database, it is populated with the sanitized data from the spreadsheets. The dates are reformatted to MySQL syntax form, then converted to UTF-8 encoded CSV's. The CSV's are combined into a single file representing a single year of

Field	Type	Size	Example
Id	INT(11)	4	38820
AttackDate	DATE	3	2016-10-17
Author	VARCHAR	96	Guccifer2.0
Target	VARCHAR	256	Democratic National Committee(DNC)
Description	VARCHAR	512	Guccifer 2.0 is back and leaks new fresh documents relating to the US political system (documents allegedly showing email conversations between DNC employees and Hillary Clinton's presidential campaign staff discussing Donald Trump's position on his tax returns).
Attack	VARCHAR	64	Unknown
TargetClass	VARCHAR	72	Org:Political Party
AttackClass	VARCHAR	16	CC
TargetCountry	VARCHAR	32	US
Link	VARCHAR	256	http://www.ibtimes.co.uk/hacker-guccifer-2-0-leaks-files-claiming-dnc-researched-donald-trumps-taxes-1587073
Tags	VARCHAR	512	Guccifer 2.0, Democratic National Committee, DNC, Hillary Clinton, Donald Trump

Table 3.2: The schema of the "attacks" table in the Cyber Attacks database, and a sample row. The "Id" field, an auto-incremented integer, is unusually large solely from adding and removing many entries to the table in development.

attacks (again, about $1000/yr$). Instead of reading the CSV's directly into a MySQL database, it is read into a "dataframe" object, a data type from the "pandas" python package, for inspection. This lets us double-check that all fields and rows look as they should, before adding them to the table. Pandas also does a nice job of handling null values without an explicit `NULL` in the cell, as well as escaping quotes and other sequences that are meant to be part of the field entry. After they are inspected, pandas dataframe objects have a nice `to_sql()` method we use for inserting data into the database an entire dataframe at a time.



```
In [3]: attacks.head()
```

Out[3]:

	Id	AttackDate	Author	Target	Description	Attack	TargetClass	AttackClass	TargetCountry	
0	35623	2015-12-16	Phantom Squad	Xbox Live	Phantom Squad prepare their Christmas campaign...	DDoS	Industry: Video Games	CC	US	http://arstechnica.com/gaming
1	35624	2015-12-16	APT16	Taiwan	Security researchers from FireEye unveil the d...	Targeted Attack	Government	CE	TW	http://news.softpedia.co
2	35625	2015-12-16	C0d3c1t4d3l	http://keepyourlinks.com/	C0d3c1t4d3l hacks keepyourlinks.com and dumps ...	Unknown	Online Services	CC	US	http://pastebi
3	35626	2015-12-17	None	Juniper Networks	Juniper Networks issues an urgent security adv...	Unauthorized Code	Industry: Networking	CE	US	http://arstechnica.com/security/
4	35627	2015-12-17	None	Landry's Inc.	Landry's Inc. is the latest hospitality firm t...	PoS Malware?	Industry: Hospitality	CC	US	http://www.krebsonsecurity.con

Figure 3.1: Inspecting the attacks as a pandas dataframe in a Jupyter notebook.

3.6 Querying The Data

```
SELECT TargetCountry, COUNT(*) as num_attacks
FROM attacks
GROUP BY TargetCountry
ORDER BY num_attacks DESC;
```

The most interesting parts of the data are the volume of attacks per month, and how attack volume is correlated in various ways. We can get a quick look at attack volume by simple queries, such as the following one quantifying attack volume by country.

This produces table 3.3. Often, there are discrepancies in how the TargetCountry or Sector is recorded, so a more flexible query can gather any attack in which 'US' appears, which may be in a field such as "US UK". So actual volume numbers are more accurately calculated using a query of the following form:

```
SELECT COUNT(*) as num_attacks
FROM attacks
WHERE TargetCountry LIKE "%US%";
```

Doing this repeatedly for the top 12 countries in attack volume produces a more accurate estimate of attack volume. If make "countries" table, we have an easier way to get more reliable estimates for all countries of the form:

```
SELECT countries.CountryCodeTwo, countries.CountryName,
COUNT(attacks.TargetCountry) AS num_attacks FROM
countries LEFT JOIN attacks ON
attacks.TargetCountry LIKE
CONCAT("%", countries.CountryCodeTwo, "%")
GROUP BY countries.CountryCodeTwo
ORDER BY num_attacks DESC;
```

After getting an idea of the countries and sectors that are targeted the most, we can focus on detailed queries to gain insight into cyber attack trends.

3.7 Analysis Overview

The analysis is performed on all attacks recorded by Hackmageddon over the 4 year period from January 1, 2014 to January 1, 2018. Attack volume by country and sector are visualized for the countries and sectors that receive the most attacks. The correlation between attack volume for different sectors and different countries are analyzed by calculating the Pearson r value and p-value. Recall that statistically

significant p-values are less than 0.05 and highly significant values are less than 0.01 [40]. Also, recall the following about Pearson Correlation, in table 3.4. This shall become useful when analyzing the significance of our results.

TargetCountry	num_attacks
US	1693
> 1	599
GB	254
IN	124
NULL	110
CA	96
AU	77
RU	71
IL	65
FR	58
KR	57
UA	55
JP	54
CN	47
IT	45
DE	43
PK	40
BR	37
TR	37
NL	27
SA	25
PH	22

Table 3.3: The top 20 countries by attack volume 2014-2017 inclusive, as queried on the attacks database using a more targeted query for > 1 and `NULL` values, and the remainder of countries using the aforementioned `LEFT JOIN` technique.

Pearson Correlation	Correlation Interpetted As
$0.00 - 0.10$	very weak
$0.10 - 0.30$	weak
$0.30 - 0.50$	moderate
$0.50 - 1.00$	strong

Table 3.4: The interpretation of Pearson Correlation values [36].

Chapter 4

CAPTURING SOCIO-ECONOMIC DATA

As discussed, previous works have shown a correlation between socio-economic factors and the volume of cyber-attacks a country receives [64] [42] [25] [37]. Here, we aggregate many socio-economic factors from open sources to study the degree to which various socio-economic factors influence a country's volume of attack received. Most of factors are taken from Heliwell's *World Happiness Report 2015* [49]. The data included on this web page in the form of Excel spreadsheets and found by clicking, "Chapter 2 Online Data Expanded with Trust and Governance" contain data on many socio-economic factors, which shall be detailed in the next section. However, because GDP was not included in this dataset, only a GDP per capita, whose precision had been destroyed by dividing by a very large number (population), additional GDP data is found through *World Bank* [26]. Two more fields "HappinessRank and "HappinessScore" are found through a dataset that Sustainable Development Solutions Network hosts on Kaggle [58]. SDSN is an organization which actively supports the World Happiness Report [9]. We shall now delve into the contents of these datasets.

4.1 The World Happiness Report

The first World Happiness Report was released in April 2012, and has released one report annually ever since, with its most recent report being released March 2018 [7]. Here we run 2 analyses: first of which only uses the 2014 data from the 2015 report, and the second of which draws from the aforementioned 2018 report, using only the data pertaining to years 2014-2017 (inclusive). All data collected in the World Happiness Report is taken from the World Gallup Poll [9].

Like the Hackmageddon data, before using the World Happiness Report data, it all needs to be wrangled [55]. In the first analysis, we are relating the socio-economic data from 2014 to the total attack volume over the total 2014-2017 period. This type of analysis assumes that the socio-economic factors relevant to a particular country don't vary significantly enough to account for the disparity in overall cyber attack volume. Conversely, in the second analysis, which compares total cyber attack volume by year to socio-economic data for that year, assays to relate changes in both year-to-year. The assumption is that using data from every year could produce more detailed results, but relating them to the only first year suffices for seeing large trends in how total attacks received in this 4 year period.

Next, because there are not any columns for ISO alpha-2 or ISO alpha-3 country codes, only full country names, they must be added. Without them, joining to the `attacks` table is error-prone. Then, two-word column names are combined into single-word names, and all missing values are filled with "NULL". Many columns are mostly empty as well, so these areas where data was too difficult to collect are discarded. Finally, we are left with the following fields.

1. CountryName
2. CountryCodeTwo
3. CountryCodeThree
4. Population
5. Region
6. HappinessRank
7. HappinessScore
8. LowerConfInt_Hap

9. UpperConfInt_Hap
10. StudyYear
11. LogGDPperCapita
12. SocialSupport
13. HealthyLifeExpectatBirth
14. FreedomLifeChoices
15. Generosity
16. PerceptionOfCorruption
17. PositiveAffect
18. NegativeAffect
19. ConfidenceInGov
20. DemocraticQuality
21. DeliveryQuality
22. PeopleCanBeTrusted

The fields "HappinessRank", "HappinessScore", "LowerConfInt_Hap", and "UpperConfInt_Hap", do not appear in the original data from the World Happiness Report, though "HappinessScore" is simply "LifeLadder" in the original study as outlined in the WHR appendix [22]. Instead, the auxiliary columns are pulled from the Kaggle dataset [58] which derived from the original WHR dataset and its annual summary by a community of data scientists. Additionally, the happiness fields only pertain to the first inquiry, and in this first overall inquiry, only data from 2015 is used,

because there aren't any datasets with happiness scores from 2014 or prior. Because socio-economic data does not vary much between years, but varies greatly between countries, 2014, the earliest year in the cyber attacks database is chosen. The implication is that socioeconomic conditions at the beginning of the study serve as an input which affects the output, cyber attack volume, over the subsequent years. To demonstrate the effects of this slight variability of socioeconomic data, consider how little the correlation between annual GDP and the total cyber attacks experienced 2014-2017 varies in table 4.1 There are not socio-economic data for 2014 for every country, in

Socio-Economic Factor	r	p-value
GDP2016	0.186	0.023
GDP2015	0.181	0.026
GDP2014	0.179	0.027
GDP2013	0.176	0.029
GDP2012	0.176	0.03
GDP2010	0.175	0.03
GDP2011	0.172	0.033

Table 4.1: Annual GDP Correlates of Total Cyber Attack Volume Received 2014-2017 in ascending order by p-value.

which case, the most recent data is used. 112 countries are analyzed using 2014 data, 29 countries are analyzed using 2013 data, 8 are examined using 2012 data, 5 using 2011 data, 1 using 2007, and 1 using 2006 data. Most of the countries lacking data from 2014 are countries which also don't receive many cyber attacks. Though most fields are self-explanatory, a few are elusive. The "...ConfInt_Hap" fields are the lower and upper confidence interval of the happiness score,"HealthyLifeExpectatBirth" was originally "Healthy Life Expectancy at Birth", "FreedomLifeChoices" in its original form was "Freedom to Make Life Choices". The field "DeliveryQuality" is a weighted

average of several fields. The Statistical Appendix for Chapter 2 of the 2015 World Happiness Report explains its significances as follows:

> Variables in the expanded data set: Democratic and delivery quality measures of governance are based on Worldwide Governance Indicators (WGI) project (Kaufmann, Kraay and Mastruzzi). The original data have six dimensions: Voice and Accountability, Political Stability and Absence of Violence, Government Effectiveness, Regulatory Quality, Rule of Law, Control of Corruption. The indicators are on a scale roughly with mean zero and a standard deviation of 1. We reduce the number of dimensions to two using the simple average of the first two measures as an indicator of democratic quality, and the simple average of the other four measures as an indicator of delivery quality, following Helliwell and Huang (2008) [17].

Detailed information about all fields is also available in the same appendix for the WHR [17].

Chapter 5

RESULTS

Risk is often defined in the utility function:

$$R = \sum_{k=1}^{N} P_{r_k} U(X_k). \tag{5.1}$$

where P_{r_k} is the probability of an event occurring, in the kth state, and $U(X_k)$ is the utility (loss) function of resource X in state k [52]. Put simply, risk in any particular state is the probability of adverse event occurring multiplied by the adversity of the event. In the case of cyber attacks, the utility function often requires information such as money lost, loss of public trust, number of devices affected, logins stolen, etc. These are sometimes difficult to come by, and are not included in the Hackmageddon data. However, the probability that an attack will occur against a particular victim country or sector could be estimated by measuring the volume of cyber attacks against a certain country or sector. In this section I discuss the results of analyzing cyber attack volume. Grouping attack victims by target country and target class (sector) allows us to identify the most targeted countries and sectors at any granularity of per diem or greater. This information is valuable because countries and sectors that are targeted at high frequency can generally expect to be targeted at high frequency in the future. In other words, a kid who is frequently bullied is at greater risk of future bullying than a kid who is not. Therefore, by measuring the volume of cyber attacks that victim countries and sectors receive, I am actually measuring the expected value of cyber attack volume for a country or sector, which as previously shown, is part of the risk equation. We measure the probability by taking advatage of the fields "TargetCountry" and "TargetSector", using them to make queries of the form,

```
SELECT dates.AttackYear, dates.AttackMonth,
    ym.TargetClass, ym.num_attacks
FROM dates LEFT JOIN
    (SELECT YEAR(AttackDate) AS AttackYear,
        MONTH(AttackDate) AS AttackMonth,
        TargetClass, COUNT(*) AS num_attacks
    FROM attacks
    WHERE TargetClass LIKE "%Gov%"
    GROUP BY AttackYear, AttackMonth, TargetClass) ym
ON (ym.AttackYear = dates.AttackYear
AND ym.AttackMonth = dates.AttackMonth) ;
```

This chapter is divided into By Country, By Sector, and Socio-economic Factors. In the section By Country 5.1 I give an overall understanding to how attack volume varies temporally and across target countries, and expound on which country pairs' attack volume is most closely correlated. Because these two studies don't share much overlap, I divide them into two distinct sections. The first answers the first question of this thesis, "How does cyber attack volume vary with time?" for various countries. The second answers the third question, "What else can we learn about trends in cyber attacks?" I perform the data analysis on temporal data, and visualize it via simple graphs displaying the volume of cyber attacks, as well as Pearson R correlations. In the Sectors section I perform the same analysis on the most targeted sectors. Here there is considerable overlap between the most targeted sectors and the most correlated sector pairs (by attack volume received), so they are addressed in a single section. I examine cyber attack volume against certain countries and sectors all on a month-by-month basis. This is a large enough window that attack volume against minority targets and sectors is nontrivial, but granular enough to answer the first

question of this thesis. The section "Socio-economic Factors" only covers correlates, because the census data for these factors are only taken once a year, whereas cyber attack traffic is analyzed at the smallest level on a month-by-month basis. This lays the groundwork for understanding the likelihood of attack, one of the components of risk assessment. The overarching implications of these results are covered in the Discussion section.

5.1 By Country

Countries that receive the most attacks are often not correlated with one another, but are of the most importance. Because their is not much overlap between these two groups, it is useful to break the study into two sections:

1. Trends 5.1.1: Study the cyber attack volume per month of the most heavily targeted countries in a subsection of their own
2. Correlates 5.1.2: Study the country pairs whose attack volumes are most closely correlated

5.1.1 Trends

In this section I discuss the results of analyses on temporal data, and their clustering by month, on the most targeted countries. Clustering by month is easy to do in SQL, and also is very easily understood by people. Clustering smaller, on a week-by-week basis is more difficult to visualize over the 4 year period, and loses some meaning. Clustering on a larger basis, such as by year, obscures how cyber attack volume varies wildly throughout the year. In figures 5.1 - 5.16 I show the cyber attack volume per country. By analyzing and visualizing this data we can better understand how the volume of cyber attacks changes over time for each country. Linear trends, or cyclical

patterns may appear, which are easier to understand visually.

In figure 5.1 I show the total attack volume between January 2014 and February 2018. Note that figures 5.1 and 5.2 represent a high attack volume relative to the change in volume, so they are not shown with a y-axis starting at zero. In figure 5.1 the graph is jagged, with a global maximum of 116 attacks in August 2016 followed by a global minimum of 50 attacks two months later. This jaggedness means that cyber attack volume is changing dramatically in irregular ways. It may indicate dramatic changes in socioeconomic conditions. In figure 5.1, I plot the average of total attacks over the past six months in order to produce a smoother line that helps interpret larger trends. We see in figure 5.1 that rolling average total attacks generally increases from the start of the rolling average plot in June 2014 until the global maximum in August 2016, dipping until July 2017, and ramping back up until at least January 2018.

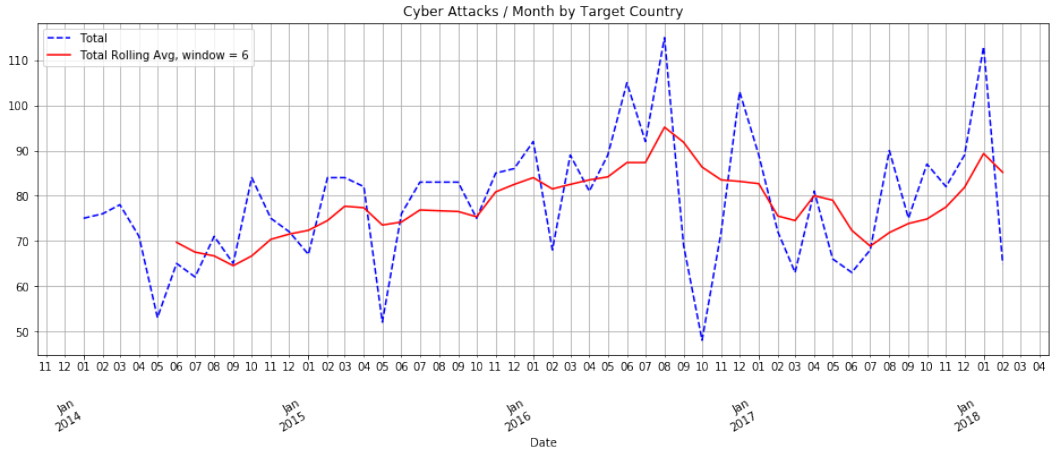


Figure 5.1: The timeline of total cyber attack volume, with a 6 month rolling average.

The graph in figure 5.2 is strikingly similar to the previous. 5.1. This is due in large part to the fact that the US receives more attacks than any other country recorded in this database. As previously shown in 3.3 the US receives more than 6.5 times the cyber attacks as the next most targeted country, Great Britain. As I illustrate in figure 5.3, the US dwarfs the next 4 most targeted countries by an order

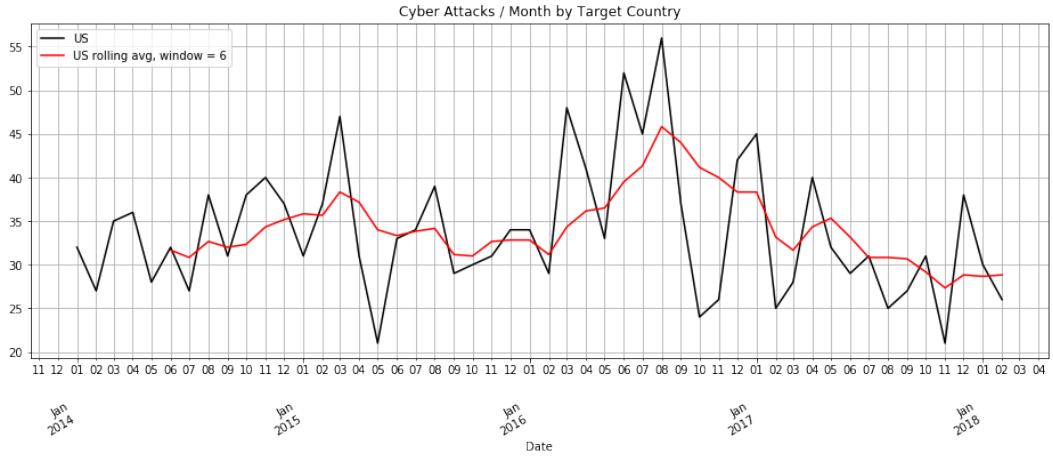


Figure 5.2: The timeline of US attack volume for the US with a 6 month rolling average.

of magnitude.

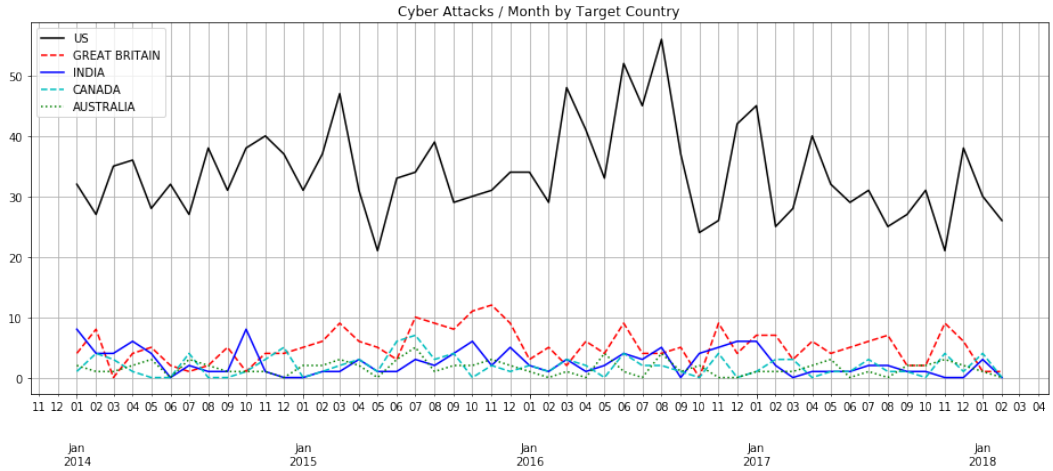


Figure 5.3: Attacks by Country Timeline — Timeline of cyber attack volume by country for top 5 victims.

Great Britain and India are 2nd and 3rd most attacked countries, respectively. Their graph in figure 5.4 confirm visually what is discovered through their Pearson correlation: they aren't correlated in any significant way($r = 0.248$, $p = 0.082$). All Pearson correlation values are on a volume-per-month basis, and can be found in table 5.1 . Great Britain has a global maximum of 12 attacks in November of 2015. Great

Britain has global minima at March 2014 and October 2016. India's attacks peak at global maxima 8 in Jan 2014 and Oct 2014, with global minima of zero attacks recorded at June 2014, Dec 2014, Jan 2015, Sept 2016, March 2017, Nov 2017, Dec 2017, and Feb 2018.

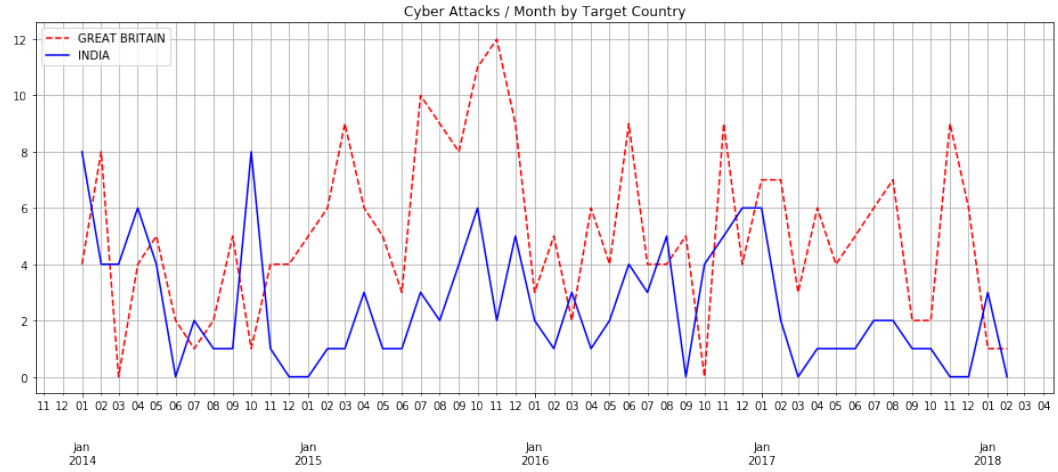


Figure 5.4: Timeline of cyber attack volume by country for Great Britain and India

By volume of cyber attacks received, Canada and Australia occupy spots 4 and 5 respectively. They are also not correlated in any significant way ($r = 0.156$, $p = 0.279$). Canada peaks in July 2015, with 12 months of 50 receiving zero reported attacks. Australia peaks in May 2015 and August 2015, with 10 months receiving zero reported cyber attacks.

The cyber attacks reported for the 6 and 7 spots, Russia and Israel, also uncorrelated ($r = 0.086$, $p = 0.552$), are increasingly sparse as shown in figure 5.6. Russia has 13 months with zero attacks recorded, and Israel has 21 such months. There doesn't seem to be much here. With few attacks, there doesn't seem to be much we can learn. In the next subsection 5.1.2, we will see that that is not necessarily the case.

In figure 5.7 I show the average cyber attacks per month for the US. The maximum

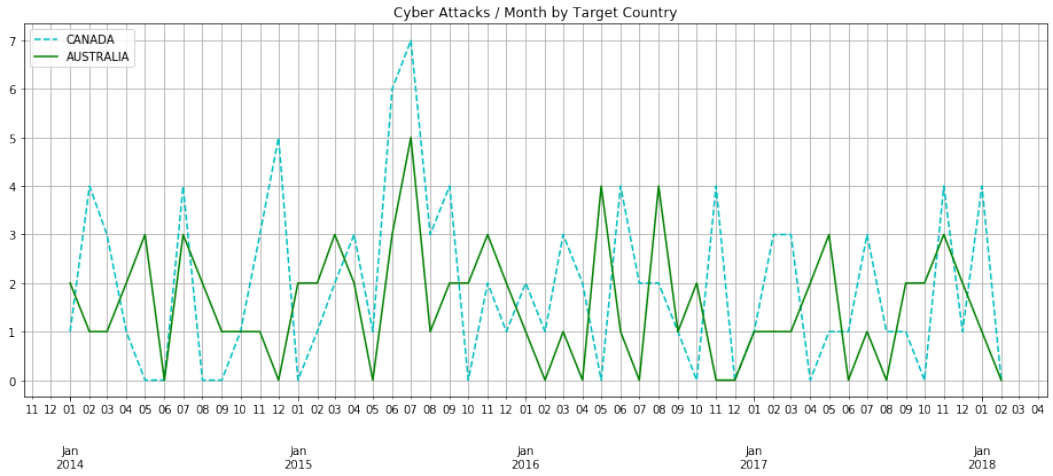


Figure 5.5: Timeline of cyber attack volume by country for Canada and Australia

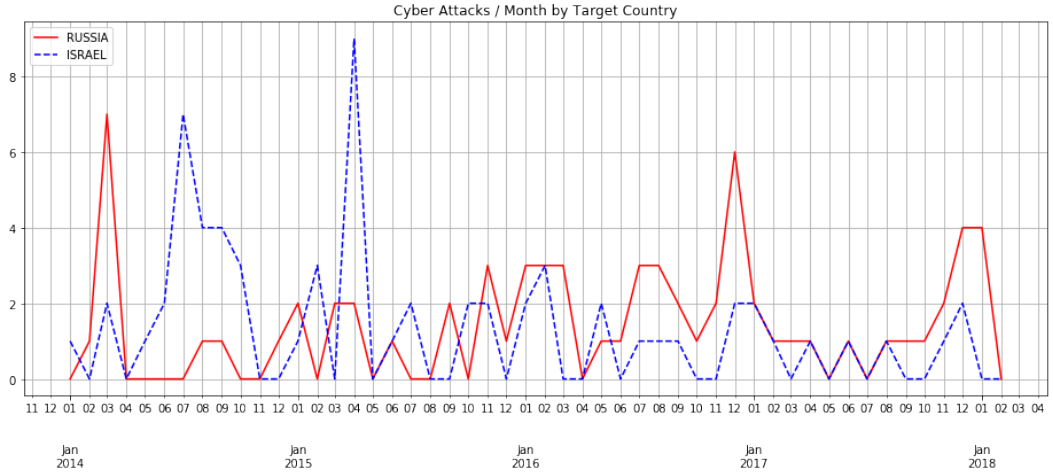


Figure 5.6: Timeline of cyber attack volume by country for Russia and Israel

average attacks per month also lands on the global maximum over the four year period, in August. A quick glance back at figure 5.2 shows that August was not the maximum for years 2014, 2015, or 2017. Though August was the third highest month in 2014, the second highest month in 2015, it was the second lowest in 2017. The massive spike in 2016 therefore played a significant role in making August the highest month on average for the US.

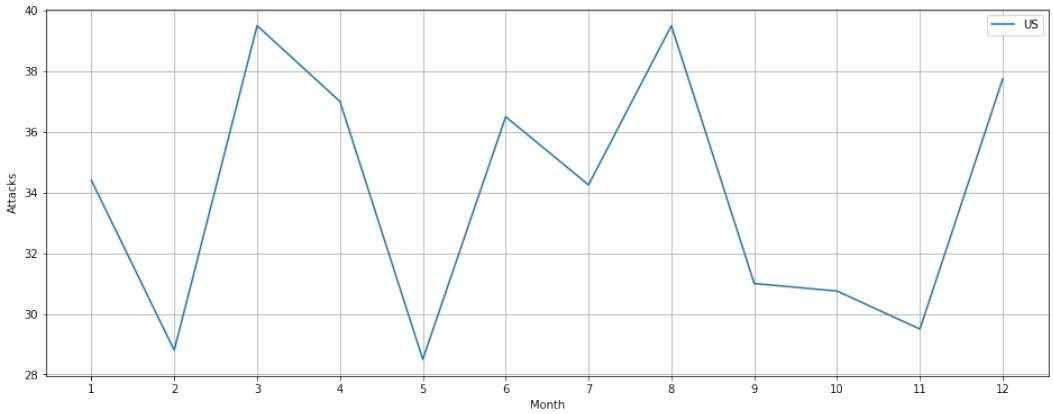


Figure 5.7: Average Cyber Attacks / Month for the US.

In figure 5.8 it can be seen that Great Britain's cyber average cyber attacks peak in November. Glancing back at 5.4 it is shown that in 3 out of 4 years, Britain's cyber attack volume peaks in November.

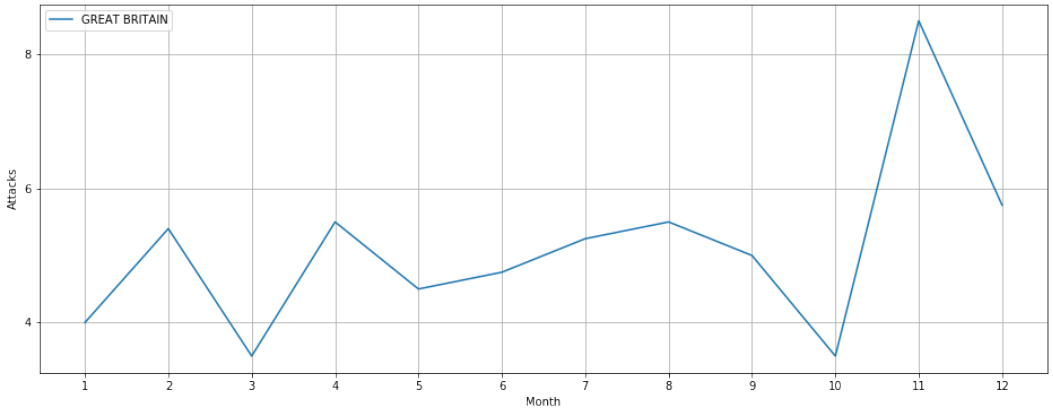


Figure 5.8: Average Cyber Attacks / Month for Great Britain.

India's attack volume hits a global maximum on average in October, as seen in fig 5.9. This behavior is expected, as India's cyber attack volume peaked in both October in 2014 and 2015, and October is the fourth most targeted month for India in 2016 and the fifth most targeted month for India in 2017.

Pakistan's by month average is shown in 5.9. Because Pakistan is the fifteenth most attacked nation, at only a sum of 40 attacks over the 4 year period, small

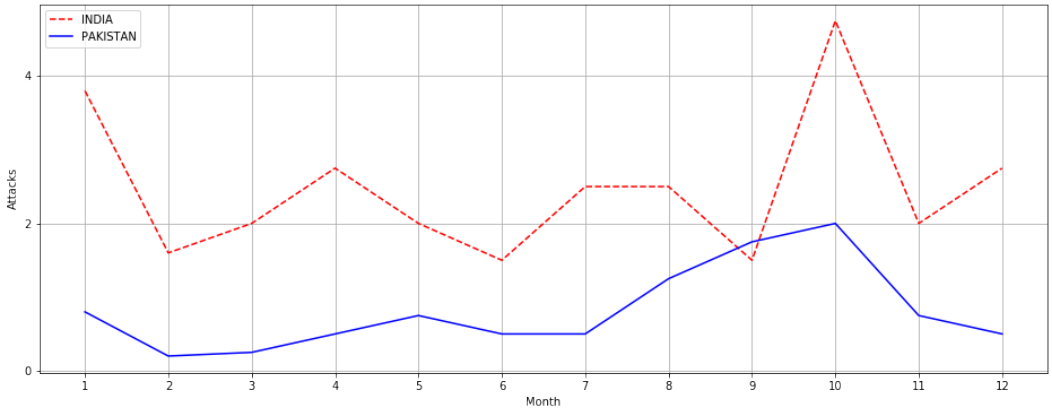


Figure 5.9: Average Cyber Attacks / Month for India and Pakistan.

bursts in cyber attacks drastically affect the average. Though the average peaks in October, October was the peak in only 2 years, as shown in 5.14. In 2014, it peaks in October with 6 attack recorded, more than 3 times the per month average of 1.74 attacks. Additionally, in 2016, the peak is 1 attack recorded, where October shares the maximum with

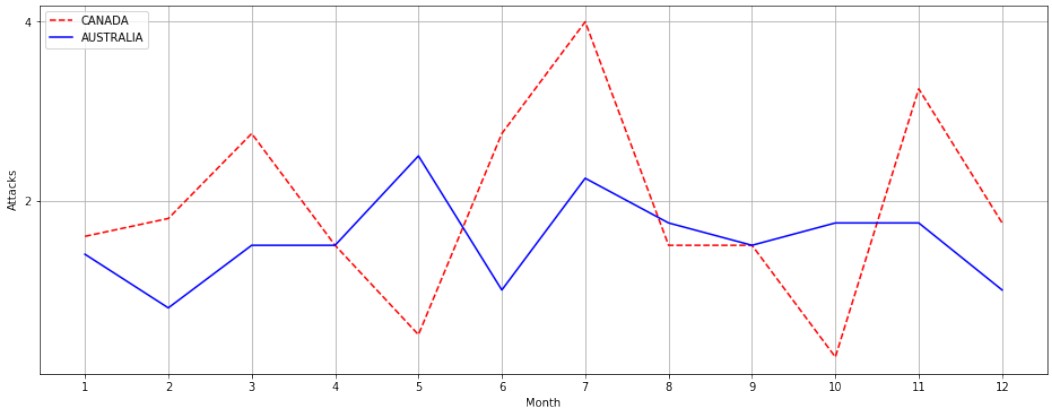


Figure 5.10: Average Cyber Attacks / Month for Canada and Australia.

5.1.2 Correlates

Table 5.1 shows the correlation between attacks received volume between the top twenty attack target countries. The correlation table indicates which countries are

likely to have similar changes in attack volume during the same month. The reason may be retaliatory, i.e. country 'A' and country 'B' may see attacks received increase at the same time because they are attacking each other. The correlation may also have a common cause, in the form of political, social, economic instability, or a common aggressor. This data is of course, not all the data that can be produced with this new cyber attack database, but produces 190 useful Pearson R correlations. A 20×20 correlation matrix render 400 entries, but 20 values along the diagonal are all $r = 1.00, p = 0.00$, because any item is 100% correlated with itself with absolute significance, and the entries on one side of the diagonal are duplicated on the other side of the diagonal.The r value is the Pearson correlation, where the absolute value determines how closely they are correlated. This is previous discussed in 3.7. Positive r values indicate that in general an increase in attacks in one country translates to an increase in attacks on another country, and negative r values indicate that an increase in attacks on one country translates to a decrease in attacks against another country, and vice versa. As discussed in 3.7, p-values are always positive, and the smaller the more significant, where a value less than 0.05 is statistically significant. The table is sorted in ascending order by p-value, which indicates statistical significance. However, this is not the only thing we care about. For instance, if two countries both have zero cyber attacks reported for the entire 4 years, then they would have a perfect correlation, with absolute significance. The most correlated countries, with the greatest significance, are Turkey and the the Philippines in the top row of 5.1, with a total attack volume over the 4-year period of 36 and 22 attacks, respectively. Here we may want to focus on countries that have a large to moderate volume of cyber attacks, and see how they correlate. The US and China are both in the top 5 target countries, and we can see that they are moderately correlated in attack volume ($r = 0.309$, $p = 0.029$).

In the figure 5.11, Turkey and the Philippines have very few attacks, but their

attack volume is correlated. Out of all countries, it is in fact the most correlated ($r = 0.467$, $p = 0.001$). Though Turkey has a global maximum of 4 attacks in January 2014, its next highest attack month is November 2014. Conversely,the Philippines has a global maximum in November 2014, and its second highest attack month is one month after Turkey's. So, though their 1^{st} and 2^{nd} maxima do not perfectly align, the collectively occupy a similar temporal space, and behave similarly over the 4 year period. It is therefore highly unlikely that their correlation is merely due to their both receiving a low attack volume. Explaining the reasons for many of these correlations is difficult, but shall be discussed in the Discussion section 6.

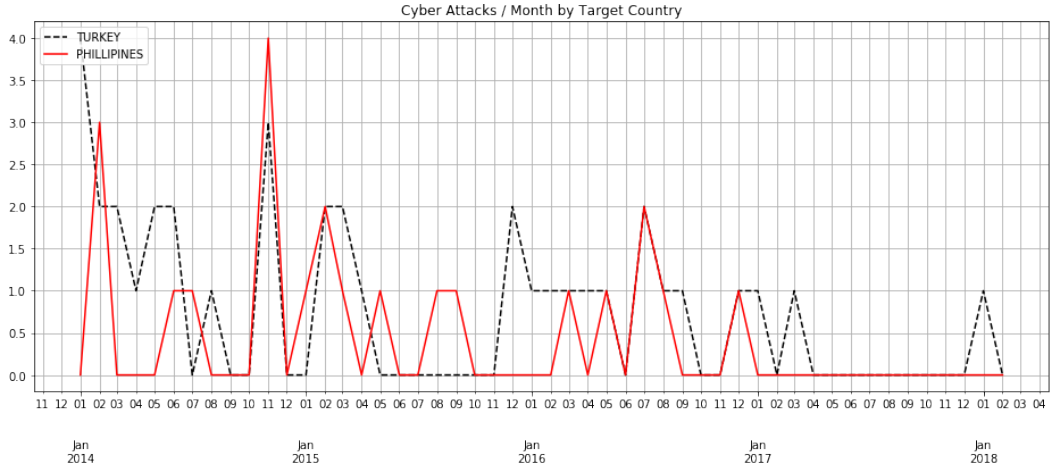


Figure 5.11: Timeline of cyber attack volume by country for Turkey and the Phillipines.

Another one of the strange results of this research is in figure 5.12. It is surprising to see that Canada and Pakistan are weakly inversely correlated ($r = -0.294$,$p = 0.038$) as shown in table 5.1. For much of the graph in figure 5.12 Canada's attacks will increase while Pakistan's will decrease, and vice versa. Possible reasons for this, and other correlations will be discussed in the Discussion section 6.

I show that Japan and China's reported attack volume in figure 5.13. As shown in table 5.1, China and Japan are moderately correlated ($r = 0.339$, $p = 0.016$). In the graph we can qualitatively notice that the volume of attacks received by Japan

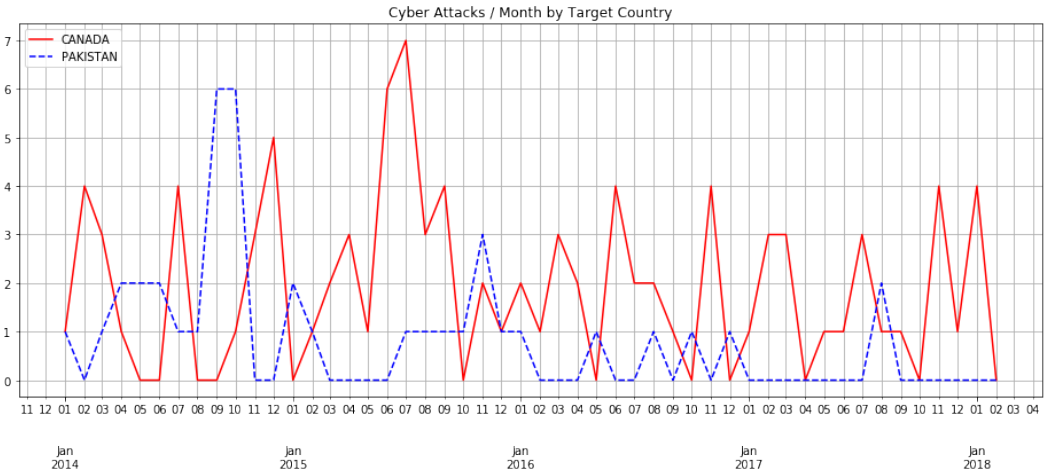


Figure 5.12: Timeline of cyber attack volume by country for Canada and Pakistan.

generally increases and decreases with those of China.The unique maximum of attacks received by China (5 attacks) occurs at the same month, January 2016, as one of Japan's 3 maxima (3 attacks) over the entire 4 year period.

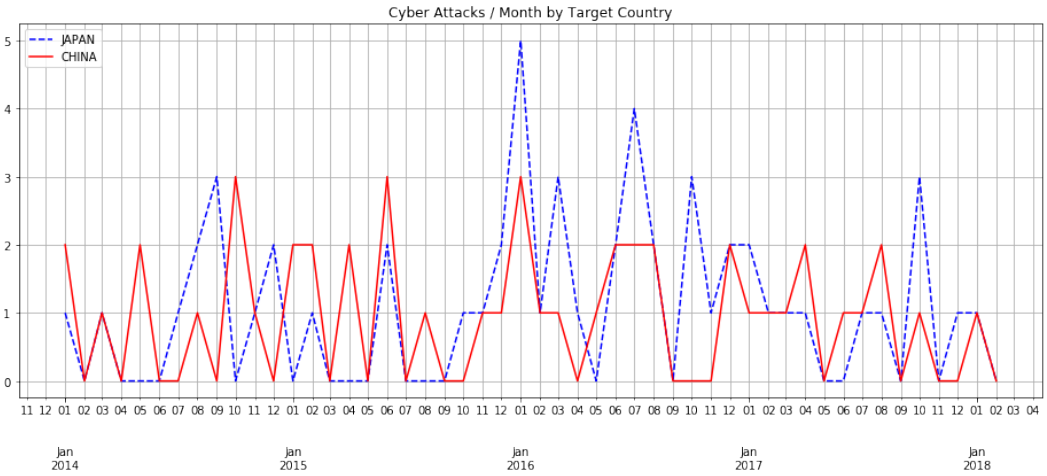


Figure 5.13: Timeline of cyber attack volume by country for Japan and China.

As with the previous graph, in figure 5.14, we see that Pakistan and India both share one of their 2 global maxima in the same month, October 2014. Similarly, Pakistan's 2015 maximum occurs 1 month after India's, and in 2017, Pakistan goes

from 7 months of zero reported attacks to receiving 2 attacks in August, the same month that India has its second greatest attack attack volume. 2016 is somewhat less interesting, because Pakistan only has 5 months of one attack and 7 months of zero attacks. However, in 2016 both India and Pakistan share a maximum in December. Running the Pearson correlation, the attacks received by Pakistan and by India are at least weakly correlated, ($r = 0.289$, $p = 0.042$), as one might expect from the previous visual inspection.

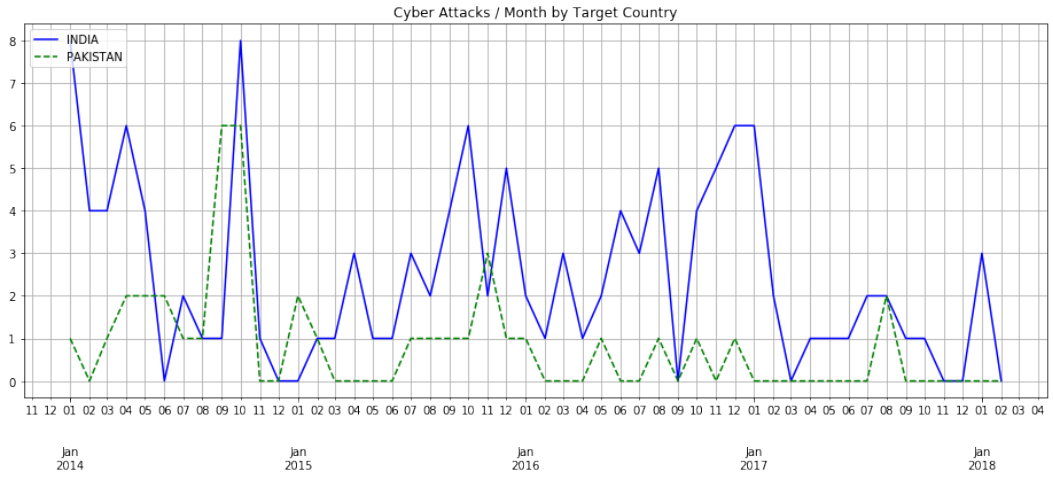


Figure 5.14: Timeline of cyber attack volume by country for India and Pakistan.

I show Italy and Germany in figure 5.15 are also moderately inversely correlated ($r = -0.356$, $p = 0.011$). Graphically we can notice that Italy reaches its global maxima of 3 attacks per month Nov 2014 and Feb 2016, while Germany has zero recorded attacks. Similarly, as German reaches its maximum of 3 attacks per month in August 2016, Italy has zero recorded attacks. Similar behavior is easily seen for several nth-most voluminous months, thus giving some intuitive understanding to their moderate inverse correlation. In general, there are surprising and often inexplicable correlations

The US and China can figure 5.16. Their correlation is more difficult to visualize because of the order of magnitude difference in attacks reported. They neither share any global maxima, nor do they share any maxima for the year. However, the US

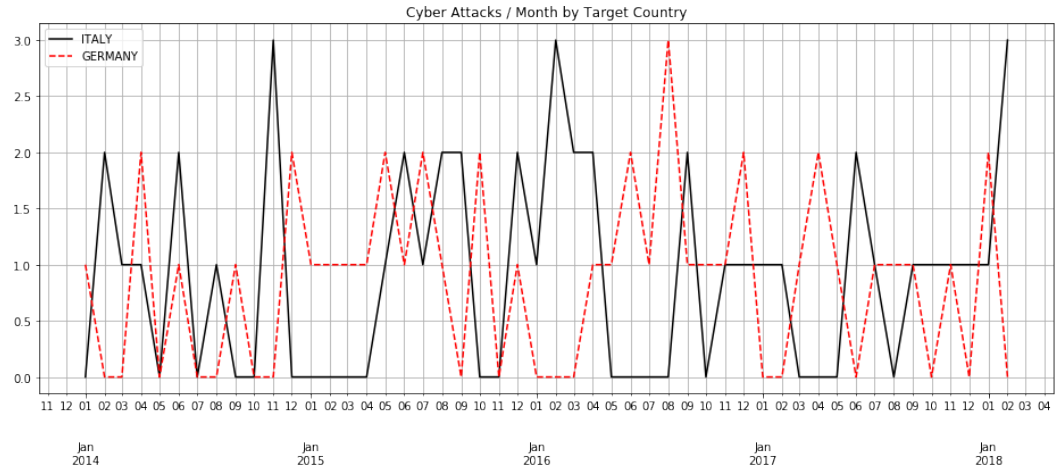


Figure 5.15: Timeline of cyber attack volume by country for Italy and Germany.

and China do share global minima in May 2015 and Nov 2017 (US = 21, CN = 0). They are moderately correlated ($r = 0.309$,$p = 0.024$)

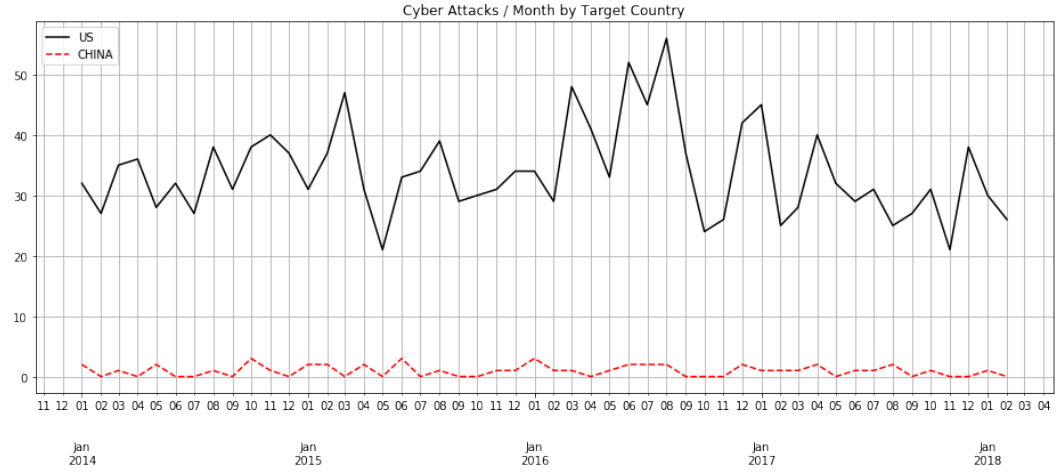


Figure 5.16: Timeline of cyber attack volume by country for the US and China.

Japan and China are also moderately correlated ($r = 0.339$,$p = 0.016$). This can be visualized in figure 5.13. Japan and China both share a maximum of received attacks in January 2016, as well as a second-highest attack month in July 2016, and a third-highest attack month in October 2017. However, they are sometimes very out

of synce, as in Sept 2014 and October 2016, where Japan reports 3 attacks and China reports none, or in Oct 2014 in which China reports 3 attacks and Japan reports none.

5.2 By Sector

Sectors are defined by the "TargetClass" field of the database, originally from Passeri's spreadsheets column name "Target Class" as previously discussed 3.2. There are a total of 629 distinct values of `TargetClass`. They often have prefixes of "Industry" or "Org", which can help organize them better, such as "Industry: Video Games", or "Org: Telecommunication". It often does not seem that Industries and Organizations are clearly delineated. It seems possible, for instance, to call "Industry: Telecommunication", "Org: Telecommunication", and vice versa. There are 109 distinct Org entries, and they account for 309 attacks. There are 306 distinct Target Classes with the "Org" prefix, and 1,078 attacks with an Attack Class containing an "Industry" prefix. Together, that is 1,384 attacks, 36% of the entire database. To this researcher, there seems to be small difference qualitatively between the categorizations of "Industry"-prefixed and "Org" -prefixed target classes: the former is associated with private business, whereas the latter generally refers to political, religious, or nonprofit groups. In monthly attack volume, they are moderately correlated ($r = 0.476$,$p = 0.000$), as shown in 5.3. Graphically, they are also fairly similar, as shown in figure 5.17. Besides the "Org" and "Industry" prefix ambiguity, there is ambiguity and overlap among sectors themselves. For instance, the TargetClass labels "News", "News (Bitcoin)", "Industry: News" and "Industry: News and Publishing" are all different labels that ostensibly reference an attack against the same sector. Additionally, some attacks target multiple sectors, in which the "TargetClass" is recorded simply as "> 1", omitting any information about the particular sectors

Country A	Country B	r	p-value
TR	PH	0.458	0.001
AU	IT	-0.366	0.009
IT	DE	-0.356	0.011
JP	CN	0.339	0.016
US	JP	0.326	0.021
SA	CN	0.324	0.022
IT	PK	-0.318	0.024
US	CN	0.309	0.029
RU	JP	0.305	0.031
IL	PK	0.297	0.037
CA	PK	-0.294	0.038
IN	PK	0.289	0.042
IN	CN	0.268	0.059
KR	NL	0.268	0.059
US	TR	0.265	0.063
IN	TR	0.264	0.064
IL	IT	-0.264	0.064
US	DE	0.251	0.079
US	RU	0.246	0.086
FR	PH	0.244	0.088

Table 5.1: The top 20 country-to-country correlation and p-values of the top 20 targets of cyber attacks, arranged in ascending order by p-value. The full list of cross country correlations is in table 5.1

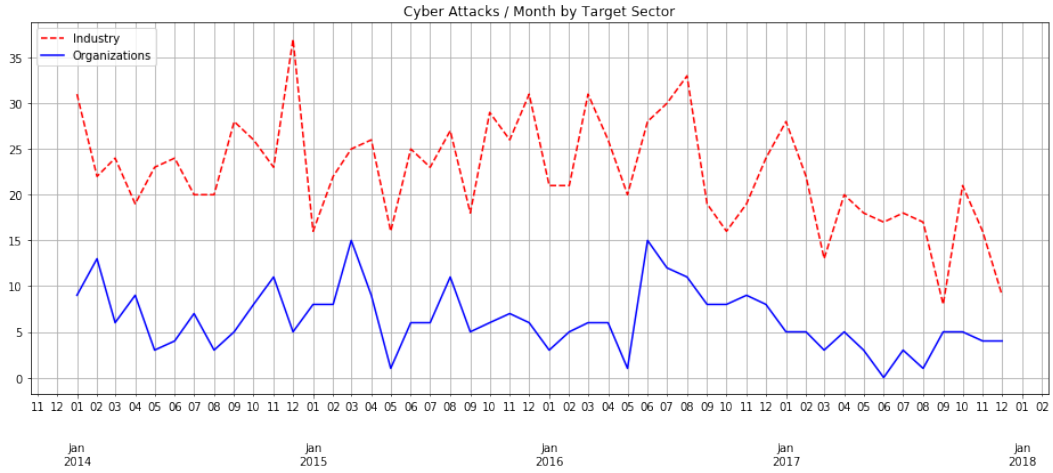


Figure 5.17: Attacks Timeline for Industry and Org Sectors.

affected, which seems unhelpful. These inconsistencies makes data analysis against target sectors very difficult. For this reason, I focus mainly on the sectors in table

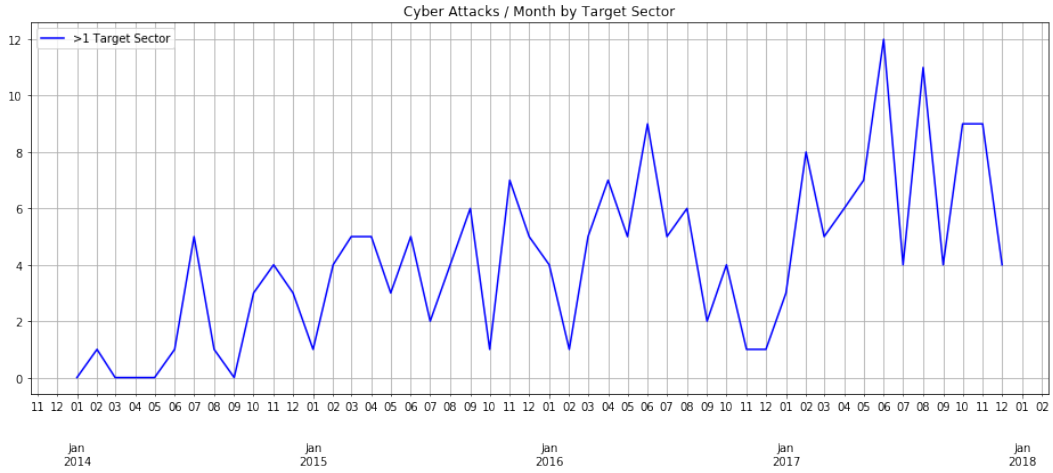


Figure 5.18: Attacks Timeline for > 1 Label

5.2, excluding the label "> 1".

I now explain why "> 1" is not very helpful in this particular analysis. The label "> 1" is the 6^{th} most targeted sector label, but it is not correlated in any statistically significant way to any of the most targeted sectors, nor is it correlated with total attacks. This indicates that attacks of a broad scope are not randomly

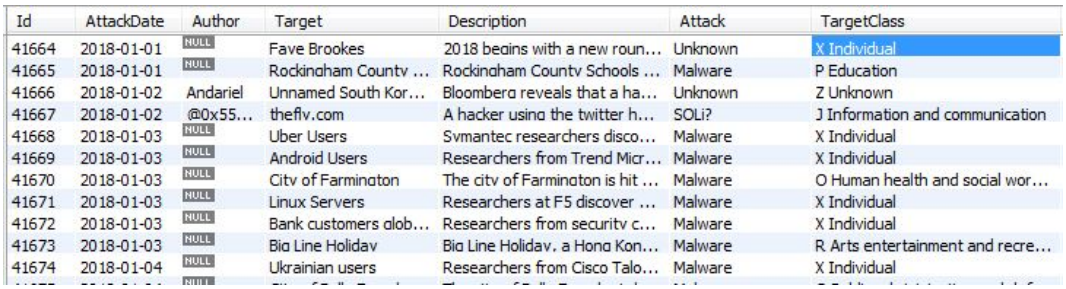
distributed amongst all attacks, but they are also independent of attacks against any other significantly targeted sector. They represent a significant portion of cyber attacks, but do not occur in similar frequencies to other popular target sectors; broad attacks are their own beast entirely. For reference I show their behavior in figure 5.18. They could be analyzed by other means, such as the type of attack, "Attack", such as "DDOS", "Defacement", or by "AttackClass", such as "CC" (Cyber Crime), or "CW" (Cyber Warfare). But broad attacks do not seem to be related to any other important target sectors. This cyber attacks database supports a near infinite combinations of queries for knowledge discovery, but only a fraction of them are presented in this thesis.

Sector	Total Attacks
Various Industries	1078
Government	592
Individuals	440
Org	306
Education	245
> 1	198
Health Care	175
Finance	167
News	94

Table 5.2: The 9 most targeted sectors between January 2014 and February 2018, arranged in descending order by attack volume.

Perhaps these ambiguities will be ameliorated in future reports, as Passeri has decided to make future reports in line with the International Standard Industrial Classification (ISIC), as stated in his Feb 2018 blog post [61]. Indeed, this new classification scheme seems to be the case in all of the cyber attacks Passeri has

recorded for 2018 so far, as shown by querying the attack database I have created with Passeri's data, as shown in figure 5.19. The effort undermines itself because it includes both the ISIC code and its associated label. If this is changed, it will greatly aid future studies, but for now, the vast majority of data collected uses the older ad hoc naming convention. Because the sector labels have changed starting January 2018, I present data only from January 2014 through December 2018.



Id	AttackDate	Author	Target	Description	Attack	TargetClass
41664	2018-01-01	NULL	Fave Brookes	2018 begins with a new roun...	Unknown	X Individual
41665	2018-01-01	NULL	Rockingham County ...	Rockingham County Schools ...	Malware	P Education
41666	2018-01-02	Andariel	Unnamed South Kor...	Bloomberg reveals that a ha...	Unknown	Z Unknown
41667	2018-01-02	@0x55...	theflv.com	A hacker using the twitter h...	SOLi?	J Information and communication
41668	2018-01-03	NULL	Uber Users	Symantec researchers disco...	Malware	X Individual
41669	2018-01-03	NULL	Android Users	Researchers from Trend Micr...	Malware	X Individual
41670	2018-01-03	NULL	City of Farmington	The city of Farmington is hit ...	Malware	O Human health and social wor...
41671	2018-01-03	NULL	Linux Servers	Researchers at F5 discover ...	Malware	X Individual
41672	2018-01-03	NULL	Bank customers glob...	Researchers from security c...	Malware	X Individual
41673	2018-01-03	NULL	Big Line Holiday	Big Line Holiday, a Hong Kon...	Malware	R Arts entertainment and recre...
41674	2018-01-04	NULL	Ukrainian users	Researchers from Cisco Talo...	Malware	X Individual

Figure 5.19: ISIC Compliant Target Classes

Though the attacks have now been put into a database, the ad hoc naming convention previously used for the "TargetClass" column makes querying the data difficult. Many sectors have been well-delineated, and those that have not have been grouped into the "Industries" and "Org" categories as previously described. Though "Industry" and "Org" are only prefixes to the specific sector, such as "Telecom", erring on the side of inclusion helps to make inferences on a grander scale, such as the threat to the private sector in general. To get around this inconsistency, targeted queries similar to figure 5 are used to get accurate counts of attacks against a particular industry.

To understand how attack volume against one sector is related to others, I have created a correlation matrix containing r and p values, as done in the previous section. As I have previously described the creation of the countries correlation table, table 5.1, I remove the diagonal, and the duplicated data on one side of the diagonal. After arranging the correlations in a columnar layout in descending order of p value,

table 5.3 is the result. Because of the complex nature of the motivation behind cyber attacks, one cannot say that the volume of cyber attacks that one sector receives is the cause or effect of the volume of cyber attacks that another sector receives. However, this study allows us to see the shifting priorities of black hat hackers in general.

One insight we can gain from table 5.3 is that the attacks received by individuals and the attacks received against various industries are strongly inversely correlated ($r = -0.595$,$p = 0.000$), and are actually the most strongly correlated sectors, by absolute value. As the private sector in general experiences fewer cyber attadcks, individuals experience more, and vice versa. This can be visualized in figure 5.20. As

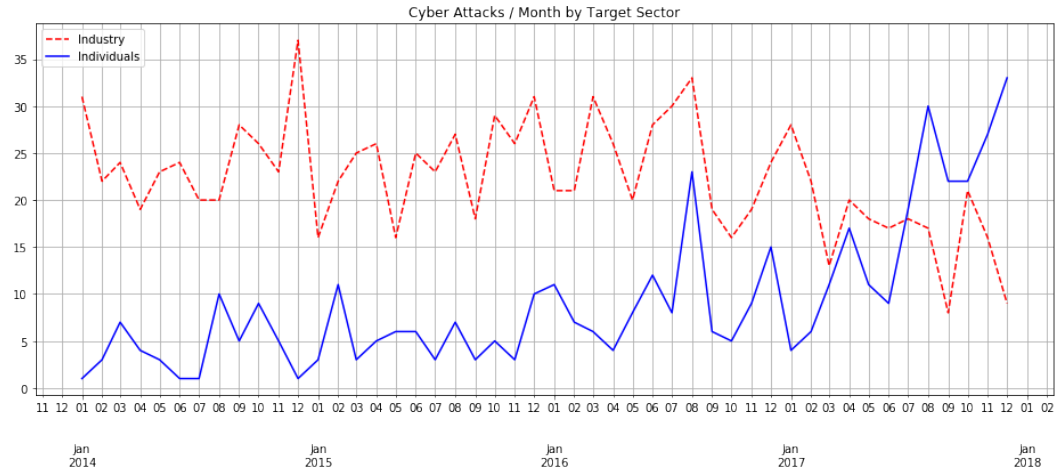


Figure 5.20: Attacks Timeline for Attacks Against Individuals and Industry Sectors

I show in this figure, there is an overall trend of attacks against industry decreasing while attacks against individuals increase. In addition to this long view of the behavior, there are also months in which the cyber attacks against industry and cyber attacks against individuals are at odds. For instance, in December 2014 industry sees a global maximum of attack volume, and individuals are near a global minimum. In May 2015, attack volume against industry reaches a local minimum as attack volume against individuals reaches a local maximum. The converse is true in June 2014. In

August 2018, the month in which overall cyber attacks peaked over the 4 year period, both attacks against industry and attacks against individuals reach a local maximum. Afterwards, they start diverging, and by July 2017, attacks against individuals have overtaken those attacks against industry.

The 2nd strongest correlation, shown in table 5.3 is between individuals and the health care sector. This is also the strongest positive correlation. In figure 5.21 I compare their monthly attack volume. They both share a global maximum in December 2017, local maxima in August 2014, January 2016, August 2016 (at the peak of all cyber attacks), and April 2017. Though they are strongly correlated and have both grown on average, attacks against individuals have grown much more quickly; starting March 2017 through the end of the year, individuals have received at least twice as many cyber attacks as the entire health care industry. This may be because of the similar goals between attacks against health care institutions and individuals, namely personally identifiable information (PII). Out of the 175 cyber attacks against the health care sector, 42 were of the of the attack class ("Attack" column) "Account Hijacking", or 24%. This ratio with respect to attacks against individuals is 35%, whereas in industry, a target sector inversely correlated to individuals, only 10% of all cyber attacks are via account hijacking. Evidently, the type of attack varies greatly between sectors, and health car and individual targets are similarly attacked for user account information.

The Industry and Government sectors are also strongly correlated ($r = 0.548$, $p = 0.000$) in table 5.3. I show in figure 5.22 the monthly attack volume against both sectors. Both sectors seem to be y-shifted versions of one another from January 2014 - July 2014, and then start to exhibit more independence. Overall, both rapidly declined since January 2017, in contrast to the previously discussed attacks against individuals, which have skyrocketed. However, total cyber attacks overall have not behaved this way, as previously shown in figure 5.1. What naturally emerges from

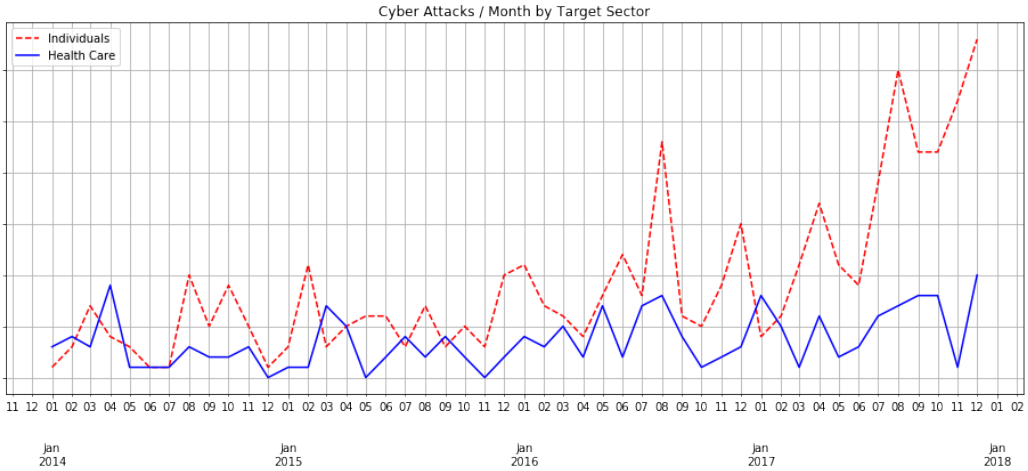


Figure 5.21: Attacks Timeline for Individuals and Health Care sectors

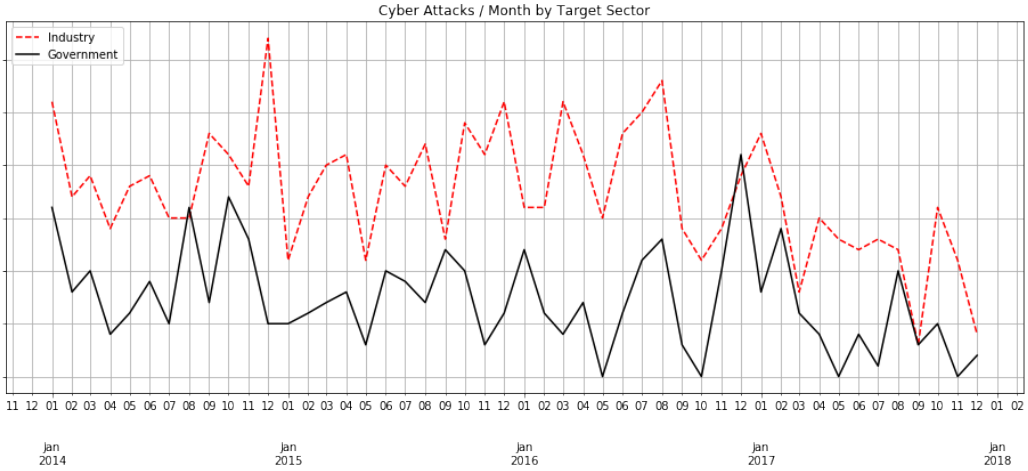


Figure 5.22: Attacks Timeline for Government and Industry Sectors, the top 2 targeted sectors.

these data is the notion of a loose conservation of cyber attacks; as one victim is ignored to some degree, others are targeted more frequently and to similar degree. Of course, this is a loose conservation, because the total volume of cyber attacks also fluctuates, but the targets, be they sectors or countries, fluctuate much more, often in complementary fashion, when there exists at least a moderate inverse correlation of attack volume between targets, as government and industry are in figure 5.23. . In general, it seems that hackers are targeting industries and governments less and

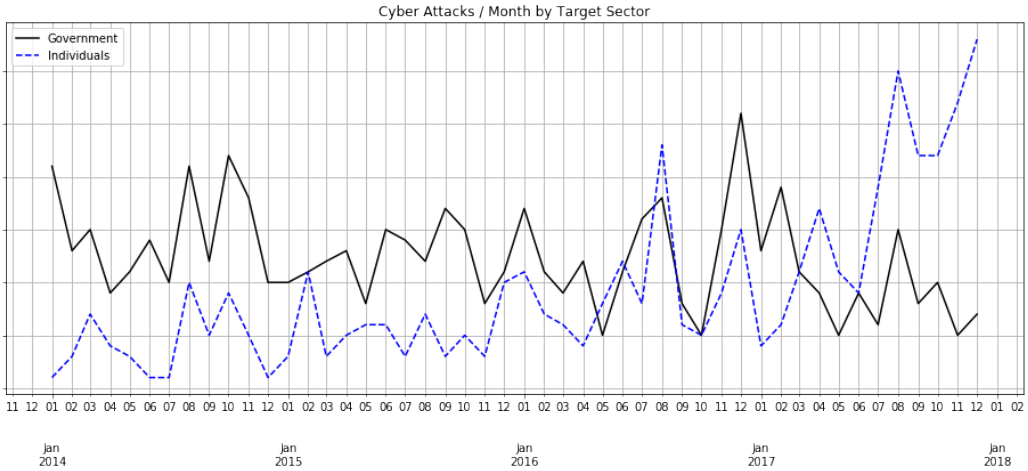


Figure 5.23: Attacks Timeline for Government and Individuals.

targeting individuals more.

Attacks against individuals and the financial sector are also moderately correlated ($r = 0.446$, $p = 0.001$). Their monthly totals are shown in figure 5.24. . Their

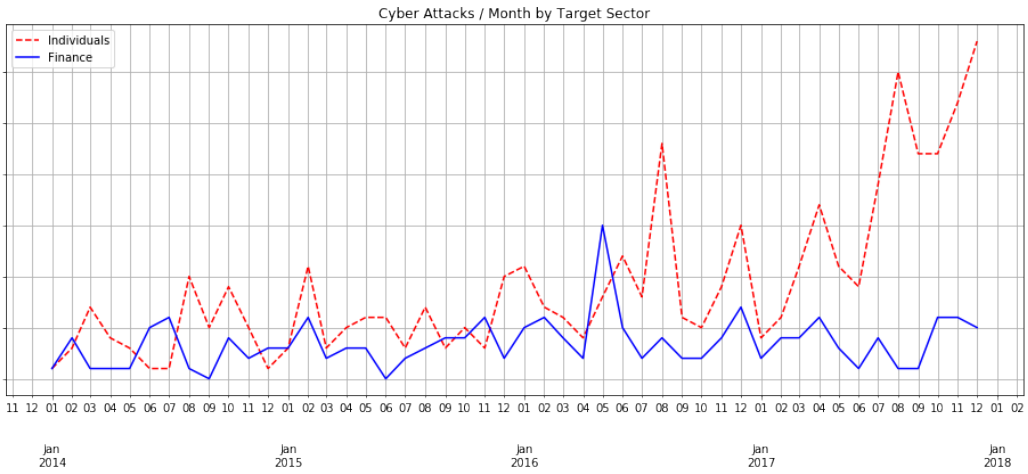


Figure 5.24: Attacks Timeline for Individuals and Finance sectors

correlation is somewhat more difficult to see, as the locations of their local minima and local maxima often do not align. All that can be said is that when individuals are attacked more frequently, the financial market is often attacked more frequently as well. May 2016 represents the highest attack against financial markets over the 4 year

period, in which it experienced 15 attacks, more than 4 times the average monthly attack rate. At least 3 of these attacks were against the SWIFT banking network, whose total losses from attacks 2015-2016 are in the millions [27]. SWIFT came under greater scrutiny after $81 million was stolen from one of its banks in Bangladesh in February 2016 [27].

5.3 Socioeconomic Factors

The results of analyzing the correlation between socioeconomic factors of a country and the total cyber attacks received imply that no one factor is even moderately correlated to cyber attack volume. The origin and meaning of these socioeconomic factors is previously discussed in section 4.1. This corroborates the findings of Solano's "Socio-economic factors in Cybercrime" [64]. Solano's work examines more specific relationships between socioeconomic factors of individual countries and their cyber attack volume. I find that A few socioeconomic measures are weakly correlated with cyber attack volume, as I show in table 5.4. Population ($r = 0.208$,$p = 0.009$) and GDP ($r = 0.179$,$p = 0.027$) are the factors most correlated to cyber attack volume. As GDP and population both represent resources for an entire country, this makes sense. In contrast, other socioeconomic factors which are not correlated to cyber attack volume, such as generosity ($r = 0.133$,$p = 0.125$), or the belief that people can be trusted ($r = 0.134$,$p = 0.335$), represent qualities of a group, but are not considered resources.

Because the US is so anomalous in its extreme volume of cyber attacks, it can be helpful to see how socioeconomic factors correlate with cyber attack volume with the US removed. The result is figure 5.5. Comparing this table to the previous one, the US changes things considerably. By simply removing the US, the correlation between population and cyber attacks received jumps from 0.208 to 0.360, and its

p-value is decimated. Removing the US also takes the correlation between healthy life expectancy from $r = 0.131$, $p = 0.114$ to $r = 0.299$, $p = 0.000$, and similar effects are found in other correlates as well. This suggests that certain socioeconomic conditions of the US make are not as correlated to cyber attack volume as the same conditions are in other countries. Removing the US as an outlier can help see trends it obscures.

The lower and upper confidence intervals of the happiness score shed some light on how the distribution of happiness scores correlate to cyber attack volume received. The narrower the interval is, the more uniform the distribution is. The lower confidence interval's positive correlation with cyber attacks received could be interpreted as a side effect of either a higher happiness score or a narrower distribution of happiness scores. To examine this idea more closely, I analyze the correlation between the size of the confidence interval of happiness score (Higher Confidence Interval of Happiness - Lower Confidence Interval of Happiness) and cyber attacks received. The result is $r = -0.242$, $p = 0.003$. This suggests that there is at least a weak inverse correlation between the size of the confidence interval for happiness and cyber attacks. If the happiness interval is very small, it means that most surveyed report similar happiness levels. Thus, societies in which most people are similarly happy have smaller confidence intervals of happiness, and receive more cyber attacks. This may be a counterintuitive example, as one may expect large confidence intervals with regard to happiness to correlate with income inequality and other social inequalities, which may in turn correlate to universally deleterious outcomes, of which cyber attacks are one. The idea that greater variance in happiness correlates with fewer cyber attacks may be explained by more complex socioeconomic forces. For instance, a wider confidence interval of happiness moderately correlates with the sentiment that people can be trusted with values $r = -0.445$, $p = 0.002$. And the sentiment that people can be trusted is moderately correlated with cyber attack volume. Intuitively, it can be observed that the more disparity between happiness recorded, the less people trust

each other. In a dystopia it could be the case that a few powerful people are very happy, and the rest are very unhappy, in which case the confidence interval would be very large, and people would not trust each other very much. In general, when these results seem unintuitive, one possible way to understand these correlations may be indirect results which can be explained by other more intuitive correlations. For another example, the happiness confidence interval size is inversely correlated with healthy life expectancy at birth ($r = -0.221$, $p = 0.008$), which in turn is correlated with increased cyber attacks ($r = 0.299$, $p = 0.000$). This is more intuitive, because a greater life expectancy is generally associated with wealthy countries who experience more cyber attacks. This intuition is confirmed in that life expectancy is strongly positively correlated with the log GDP per capita with $r = 0.811$, $p = 0.000$. Many of these seemingly disparate socioeconomic features of a society indicate the assets that a society has, and therefore, the value of launching a cyber attack.

Another important aspect about socioeconomic conditions is the unique patterns to attacks against certain countries. For instance, in 3 of 4 years in question in this study, Great Britain received its peak cyber attacks in November, as previously shown in figure 5.4 implying a heightened risk to GB leading up to holiday shopping season. In 2016 Great Britain had 2 maxima: one in the predictable November spot, and one in June, when the Brexit vote was held [39], it is possible that it is coincidental, or that general social tension is correlated with a greater volume of cyber attacks received.

Election cycles can also play a large role in cyber attack volume. India's most recent general election occurred in April - May 2014 [62], and 3 months prior to this, India's cyber attack volume reached a maximum for the entire 4 year period, also illustrated in figure 5.4. The US also experienced a similar influx of cyber attacks at a similar time relative to a major election. Over the 4 year period, cyber attack volume against the US reached a maximum 3 months before the November 2016

election, as shown in figure 5.2.

Alternatively, some attacks do not aim to change an election outcome, but to protest it. As shown in figure 5.6, Israel's maximum cyber attacks received occurs in April 2015. In 2 ways these attacks against Israel are different than both the attacks around the Indian and American elections. Firstly, the individual attacks are clearly politically motivated by their descriptions. Of the 9 attacks in April 2015, 6 attacks are coordinated under the #OpIsrael tag, 1 is carried out by the Palestinian hacker group "Gaza Team" defacing government sites with pro-ISIS propaganda, 1 attacks Israeli military networks, and 1 attacks the Israeli arms importer and manufacturer Fab-Defense. Secondly, these attacks follow the election, rather than precede it. These attackers are not trying to influence an election or taking advantage of the socioeconomic condition for profit; they are likely protesting the recent controversial election of Prime Minister Netanyahu.

5.4 Attack Classes

We break down attack classes into 4 groups based on Passeri's categorization: Hacktivism (H), Cyber Crime (CC), Cyber Espionage (CE), and Cyber Warfare (CW). These classes are not a universal set of accepted distinctions, and may sometimes overlap, but they describe the goal of a cyber attack fairly well. They correspond with the goals of a socio-political advantage (Hacktivism), a financial advantage (Cyber Crime), an informational advantage (Cyber Espionage), and an operational advantage (Cyber Warfare). The monthly attack volume of all 4 classes are illustrated in figure 5.25. Cyber crime accounts for most attacks reported, by a large margin. They peak just before and just after the 2016 US Presidential election, with a local minimum spanning 2.5 years (May 2015 - December 2018) between peaks. It is almost as if cyber criminals got tired and decided to take some time off.

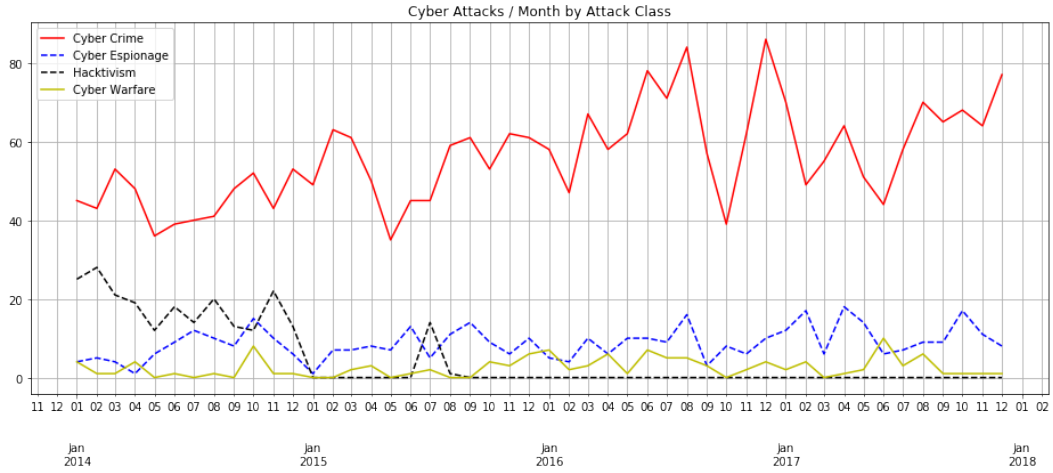


Figure 5.25: Attacks Timeline for All Attack Classes

Another interesting feature is that hacktivism reported plummets from a couple dozen attacks per month in the beginning of 2014 to zero reported hacktivism attacks since August 2015. This is probably not because of a lack of knowledge, as hacktivist activity is generally noticed in cyber security news. However, it may be difficult to classify an attack as related to hacktivism in certain situations. For example, there is an entry in the Hackmageddon data May 4, 2015, whose description reads, "A construction sign is hacked at the University of Montana, inviting users to "Smoke Weed Everyday"." The attack class is "Unknown". However, Montana had been a hotbed for political struggle over legalizing marijuana for many years [28]. This struggle came to a head in May, in which bipartisan legislation was introduced in the form of the Merkley-Daines amendment, which would allow doctors to recommend marijuana to veterans. It is reasonable to assume that the motivation was hacktivism, and that many other incidents are unnoticed, and confused for vandalism for vandalism's sake. Defacement is on the decline, and to some extent, DDoS attacks are also on decline. As a reminder, these numbers do not say anything about the number of individuals or machines affected; this means that it is possible that fewer attacks occur at a time when more widespread attacks occur, affecting more individuals and machines.

Cyber warfare is very minimal and erratic, while cyber espionage is slowly increasing. Cyber warfare pits very sophisticated government-sponsored groups known as "advanced persistent threats" (APT's) against another country. The aggressor APT does not want to be disclosed in fear the country may seem malicious. Often, the victim country will not want to disclose an exposure for fear of appearing weak. Because the sophistication of these attacks, and their mutual concealment, events of cyber warfare may be the most concealed. Thus the Hackmageddon data likely only discloses a very small sample of the actual attacks countries pit against one another.

5.5 Attack Vectors

We show the most common attacks are in table 5.6. Their monthly totals are shown in figure 5.26. Account hijacking is the exfiltration of login credentials. Malware is unauthorized software running on a machine. Targeted attacks are a broad category of attacks that seek to compromise a system, often a piece of infrastructure, as when, in 2015, it was revealed that in 2013, a group of Iranian hackers had penetrated the online control system of a New York dam through a cellular modem [44]. SQL injection attacks occur when a user is able to directly manipulate SQL commands through a web form, thereby injecting malicious statements. DDoS is a distributed denial of service, in which many machines make requests to a server at once, overwhelming the server, making it unavailable. Defacement is hacking a website to change its content, and is a frequent result of hacktivist activity.

Unknown attacks peak in August 2016, the peak of the attacks on the US, and the global maximum of all attacks. Unknown attacks are those in which the technological means of the attack are not specified. The reason for it being unknown may be from the victim's true ignorance, or the victim may kept it secret for fear of the vulnerability being taken advantage of again. Account hijacking has been slowly

increasing, and malware has increased dramatically, eclipsing unknown attacks as the dominant threat.

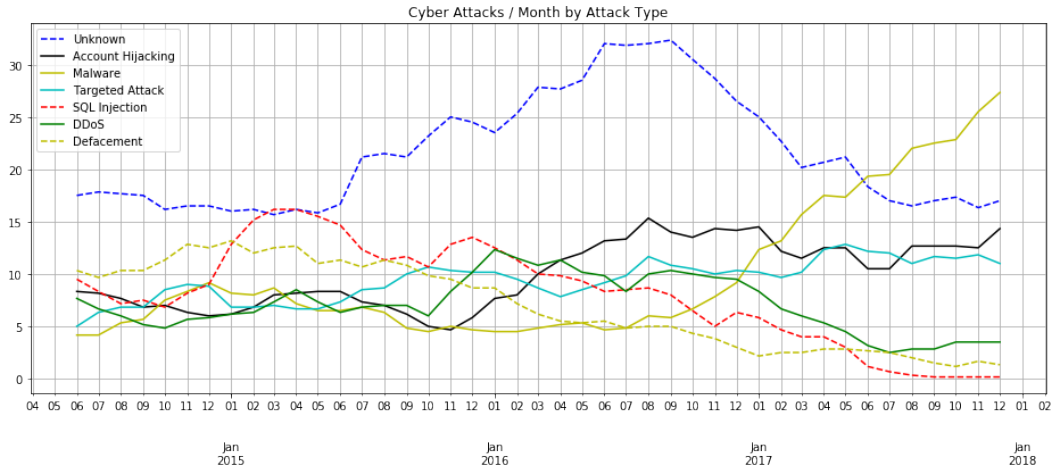


Figure 5.26: Attacks Timeline for All Attack Types, as a 6 month rolling average

5.6 Apriori Results

To answer the 3rd question of this thesis, "What else can we learn about trends in cyber attacks?", we use the Apriori algorithm to discover frequent grouping of attack variables. To prepare a dataframe for the Aprior algorithm, each of the most popular values of 'TargetClass', 'TargetCountry', 'AttackClass', and 'Attack' are represented in a column of a pandas dataframe. The US is excluded because its overwhelming prevalence obscures other relationships. The columns are, 'TargetClass: Education', 'TargetClass: Government', 'TargetClass: Individuals', 'TargetClass: Industry', 'TargetClass: Finance', 'TargetClass: HealthCare', 'TargetClass: News', 'TargetClass: Org', 'Attack: DDoS', 'Attack: SQLInjection', 'Attack: AccountHijacking', 'Attack: Targeted', 'AttackClass: Hacktivism', 'AttackClass: CyberCrime', 'AttackClass: CyberEspionage', 'AttackClass: CyberWarfare', 'TargetCountry: GB', 'TargetCountry: CA', 'TargetCountry: AU', 'TargetCountry: RU', 'TargetCountry:

KR', 'TargetCountry: FR', 'TargetCountry: UA', 'TargetCountry: JP', 'TargetCountry: IT','TargetCountry: DE', 'TargetCountry: PK', 'TargetCountry: BR', 'TargetCountry: TR', 'TargetCountry: NL', 'TargetCountry: SA', 'TargetCountry: India', 'TargetCountry: PH'. Each row represents 3 consecutive days from January 2014 - December 2017, a total of 487 rows. The data value '1' is placed in a cell if an attack with the target value occurred within that specific 3 day period, and '0' otherwise. The results, in tables A.2, A.3, and A.4, use an abbreviated form of original column names: 'AC' for 'AttackClass', 'A': for 'Attack', 'TC' for 'TargetClass', and 'TCo' for 'TargetCountry'. We focus on support, which is simply the frequency that the union of the antecedents and consequents occurs in a dataset. A full explanation of the interpretation of Apriori results can be found in previous works [23] [30] [24]. The results in table A.2 imply that within 3 day intervals January 2014 - December 2017, the most features of attacks most frequently occurring are that they are financially destructive incidents (Cyber Crime) against the private sector (Industry). In any 3 day period over the time line, both occur 56% of the time.

Sector A	Sector B	r	p-value
Individuals	Industry	-0.595	0.000
Individuals	Health Care	0.562	0.000
Government	Industry	0.548	0.000
Industry	Org	0.476	0.000
Individuals	Finance	0.446	0.001
Health Care	News	-0.360	0.010
Government	Individuals	-0.348	0.013
Government	Org	0.346	0.014
Government	Finance	-0.345	0.014
Individuals	Org	-0.341	0.015
Industry	Finance	-0.339	0.016
Individuals	News	-0.326	0.021
Industry	Health Care	-0.283	0.046
Finance	Health Care	0.258	0.070
Industry	News	0.234	0.103
News	> 1	-0.233	0.104
Finance	Org	-0.225	0.116
Individuals	> 1	0.199	0.167
Government	News	0.151	0.294
News	Org	0.150	0.298
Education	Individuals	-0.140	0.332
Government	Health Care	-0.137	0.342
Education	News	0.099	0.496
Government	> 1	-0.092	0.523

Table 5.3: Cross-industry attack volume correlations among the top 7 sectors ("TargetClass" field), sorted in ascending order by p-value.

Socio-Economic Factor	r	p-value
Population	0.208	0.009
GDP	0.179	0.027
Lower Confidence Interval of Happiness	0.173	0.033
Happiness Score	0.172	0.034
Upper Confidence Interval of Happiness	0.170	0.036
Happiness Rank	-0.167	0.040
Log GDP per Capita	0.175	0.041
Healthy Life Expectancy at Birth	0.131	0.114
Generosity	0.133	0.125
Delivery Quality	0.408	0.147
Positive Affect	0.109	0.199
Freedom To Make Life Choices	0.098	0.247
Social Support	0.092	0.274
People Can Be Trusted	0.134	0.335
Confidence In Government	-0.062	0.488
Perception Of Corruption	-0.046	0.597
Democratic Quality	0.026	0.929
Negative Affect	-0.004	0.962

Table 5.4: Socio-Economic Correlates of Total Cyber Attack Volume Received 2014-2017 in ascending order by p-value.

Socioeconomic Factor	r	p-value
Population	0.360	0.000
Healthy Life Expectancy at Birth	0.299	0.000
Log GDP per Capita	0.297	0.000
Lower Confidence Interval of Happiness	0.281	0.000
Happiness Score	0.274	0.001
Happiness Rank	-0.273	0.001
Upper Confidence Interval for Happiness	0.268	0.001
GDP	0.238	0.003
People Can Be Trusted	0.383	0.005
Generosity	0.145	0.095
Perception Of Corruption	-0.144	0.097
Delivery Quality	0.408	0.147
Social Support	0.121	0.150
Freedom To Make Life Choices	0.111	0.192
Negative Affect	-0.075	0.372
Positive Affect	0.076	0.374
Confidence In Government	0.008	0.925
Democratic Quality	0.026	0.929

Table 5.5: Socio-Economic Correlates of Total Cyber Attack Volume Received 2014-2017 Excluding the US in ascending order by p-value.

Attack	Number of Recorded Attacks As
Unknown	1010
Account Hijacking	507
Malware	463
Targeted Attack	449
SQL Injection	349
DDoS	339
Defacement	317

Table 5.6: The most common attack types.

Chapter 6

DISCUSSION

In general, cyber attacks have a range of motivations, from monetary gain, political change, or simply spite, and other works have studied these [50]. In this study I cannot make a detailed conclusion of these motivations. Instead, I focus on the larger implications of each section. The main conclusions can be summarized in the following:

1. Countries with higher population, GDP, happiness, and other metrics that relate to resources are attacked somewhat more often.

2. There is an overall shift from attacking governments and businesses towards attacking individuals, even when political or financial gain is the object of the attack.

3. Malware has become the dominating threat in terms of number of attacks recorded in cyber security news.

The 1st point is supported by previous cyber security research, in which GDP per capita is moderately correlated ($r = 0.42$, $p < 0.001$) [33]. The 2nd point has not been examined. The 3rd point is supported by industry reports of a recent doubling of malware against IoT devices [20] and an overall decrease in DDoS attacks [11].

6.1 Countries

The data collected about attacks directed toward a country in particular is minuscule, and does not accurately estimate the true number of cyber attacks occurring. This is because most attacks are not recorded, and many of the recorded attacks are not

publicized for fear of making the target person, corporation, country, or government appear weak. Nevertheless, the biases inherent in the collection of these attacks seem self-consistent. There are no observable changes in these biases, and a small sample of attacks against countries with more recorded attacks can inform us of how cyber attack volume truly changed, and what can be expected in the future.

The US is the most consistent with other studies. By this study, the US is the most attacked country, and this conforms to Kumar's account in "Approaches to understanding the motivations behind cyber attacks" [50]. Kumar also found that China was attacked with DDoS the 2$^{\text{nd}}$ most. However, my findings indicate that China is only attacked by DDoS the 16$^{\text{th}}$ most, and is attacked overall the 11$^{\text{th}}$ most. As Kumar's findings use `www.digitalattackmap.com` [6] which directly records DDoS attacks as they happen, and mine rely on security reports, their numbers are much more accurate. However, from their limited recorded attacks, and the moderate correlation between attacks against China and Japan, we can say that their similar geography and economy make their cyber attacks tightly coupled. For instance, when the Yen and the Chinese stock market plummeted in January 2016 [2], China and Japan both saw their global maxima in cyber attacks 2014-2017 occur that month 5.13 . This demonstrates the surprising power of incomplete data when sampled consistently.

Great Britain is the 2$^{\text{nd}}$ most attacked country according to this study, and in 3 of 4 years, its annual maximum of attacks received occurs in November. This is again, using a very incomplete dataset, but it strongly implies that Great Britain's cyber attacks are cyclical, and are generally tied to the holiday shopping season.

The weak inverse relationship ($r = 0.294$, $p = 0.038$) between attacks received by Pakistan and attacks received by Canada may seem perplexing. Because of its small sample size and surprising result, it is tempting to consider this finding in figure 5.12

a fluke . However, accurate relationships have been found in other small sample sets, and the correlation is weak, so perhaps it is not so unusual after all.

On average, Pakistan and India seem periodic, maximizing on average in October in figure 5.9. They are geographically similar, and in mutual conflict, which is not surprising. This average is disproportionately influenced by the cyber attack volume of October 2014, which was a global maximum for both countries as implied by figure 5.14. Another interesting result in figure 5.14 is that the Indian elections, which only occurred once 2014-2017, occurred March through April [16], and one of India's 2 attack volume maxima occurred 3 months prior, in January. Four of January 2014's 8 attacks against India target the government directly. This maximization of cyber attacks received around 3 months prior to a national election also occurred in the 2016 US Presidential election. So it is reasonable to expect that 3 months before a contentious election, cyber attacks may increase.

6.2 Sectors

The clearest observation is that attacks against various industries are dereasing and attacks against individuals are increasing, as figure 5.20 illustrates. Attacks against individuals and industry are also the most correlated by absolute value, ($r = -0.595$, $p = 0.000$). There seems to be a systematic shift towards attacking individuals, the largest share of which are account hijacking (31%), followed by malware (30%). Of account hijacking against individuals, (14%) were Twitter account hijacking. This is indicative of the types of trends to anticipate in the coming years.

Cyber attack volume against various industries and government sectors, as shown in figure 5.22, are very similar ($r = 0.548$, $p = 0.000$). For the first 6 months they look like y-shifted versions of one another, and have the same sign of their first derivative; an increase in one means an increase in another. All this indicates that attackers

target governments and the private sector at similar times, perhaps due to a tight coupling of politics and the economy of the US, the most attacked nation.

Attacks against several sectors (> 1) make up a small portion of attacks in figure 5.18, but are on the rise, quadrupling between 2014 and 2017. In addition to the graphs in section 5.2, the change in attacks against the most targeted sectors can be understood through table 6.1. For clarification, the "Organizations" category includes political parties and religious groups, which are neither private sector nor government agencies.

Sector	2014 January-June	2017 July-December
Individuals	4%	31%
Industry	34%	18%
Government	19%	10%
Healthcare	5%	8%
> 1	2%	8%
Finance	3%	5%
Organizations	10%	4%
Education	7%	4%

Table 6.1: Contrast of the attacks by sector at the beginning half of 2014 and the last half of 2017.

Another way to profile these attacks against different sectors is to examine the type of attack most likely to occur in each sector, as shown in tables 6.2, 6.3 6.4

These data all imply that substantial portion of future attacks will target individuals, and this trend will continue to grow.

Attack	% of Attacks Against Individuals
Account Hijacking	31%
Malware	30%
Unknown	9%
Targeted Attack	7%
Malvertising	5%
Defacement	2%
Mobile Malware	1%

Table 6.2: Types of attacks against individuals, by percentage.

Attack	% of Attacks Against Industry
Unknown	31%
Account Hijacking	10%
SQL Injection	10%
Malware	10%
DDoS	9%
Targeted Attack	7%
Point of Sale Malware	7%

Table 6.3: Types of attacks against industry, by percentage.

6.3 Socioeconomic Factors

From the results obtained, there are no strong correlations between any of the socioeconomic factors and an increase cyber attacks received, with a moderate correlation for only 3 factors. This difficulty is corroborated in Solano's "Socio-economic factors in cybercrime", which also does not find any strong correlations between socioeco-

Attack	% of Attacks Against Governments
Targeted Attack	23%
Defacement	22%
Unknown	18%
DDoS	14%
SQL Injection	6%
Malware	5%
Account Hijacking	5%

Table 6.4: Types of Attacks Against Governments, by percentage.

nomic factors and cyber attacks received across all countries [64]. The top 3 socioeconomic factors that elevate the likelihood of cyber attacks are population ($r = 0.360$, $p = 0.000$), healthy life expectancy at birth ($r = 0.299$,$p = 0.000$), and log(GDP) per capita ($r = 0.297$, $p = 0.000$). Some of these categories that imply more cyber attacks are also at least weakly correlated to GDP closely tied to GDP. Healthy life expectancy is strongly correlated with log GDP per capita with ($r = 0.811$, $p = 0.000$), and happiness score is weakly correlated with log GDP per capita with ($r = 0.171$, $p = 0.035$). However, population is not correlated with log GDP per capita at all ($r = -0.032$, $p = 0.714$), indicating that attackers view people as a distinct resource on their own. More generally, the factors that matter are resources that malefactors can subvert. Those resources are the number of human beings and the money that a country holds. This makes sense, as the target of cyber attacks is often direct monetary gain, or account hijacking, which can then be used for monetary gain by using card payment information or through extorting the victims. Happier countries are attacked more often, with a correlation between happiness score and number of attacks received. Including the US, the correlation is ($r = 0.293$, $p = 0.001$); excluding the US the correlation is ($r = 0.183$, $p = 0.035$). In general, countries with more

citizens, happier citizens, and more wealthy citizens are targeted more often.

6.4 Attack Classes

Over the span of the study 2014-2017, cyber crime has always dominated other attack classes, and continues to grow, as shown in 5.25. This asymmetry is likely because hacktivism is truly does represent a small slice of the pie, and cyber espionage and cyber warfare are not generally reported. Cyber crime is most likely a large source of cyber attacks, but the large gap between cyber crime and other attack classes is indicative of the culture of secrecy around cyber warfare and cyber espionage.

6.5 Attack Vectors

The peak of all attacks which occurred 3 months prior the US Presidential elections is also where cyber activity starts to change dramatically 5.26. Unknown attacks plummet, DDoS, SQL Injection, and Defacement attacks also taper off. Targeted attacks and account hijacking do not deviate much, but malware increases dramatically, so much so that December 2017 sees 5 times the number of malware reports as the average from the start of the study in 2014 until August 2016. Because malware is more associated with attacks against individuals, this larger trend implies as much about who is targeted as it implies about how they are being targeted.

Chapter 7

CONCLUSION

Throughout this thesis I have analyzed cyber attack volume as a function of time, attack type, method of attack, target country, target sector, country to country relationships, socioeconomic factors, and have discovered tentative but meaningful correlations in various factors of the victims of cyber attacks. Furthermore, I accomplished this using only freely available data sources. I discovered that annually sampled socioeconomic statistics are difficult to correlate with cyber attack volume received, but GDP per capita and population play the strongest roles. Despite this tenuous relationship, the largest spikes in cyber attack volume are tied to socioeconomic events, namely contentious elections, political corruption, and stock market dives. The volume of future attacks will continue to increase, more often targeting individuals directly, affecting wealthier more populous nations including the US and China, and more often take the form of malware. The imminent ubiquity of the Internet of Things (IoT) will present new challenges to security as easier systems to compromise, a wider attack vector, and higher value objectives, such as home surveillance, and interference with physical systems, such as power plants, as well as consumer devices. Therefore, it is extremely important that the security research community knows as much as possible about trends in cyber attacks.

From the results which point to increased cyber attacks preceding contentious elections, I expect the US and the world to experience the most attacks of 2018 in August, preceding the US midterm elections. For the same reason, I also expect August 2020, preceding the US presidential election to have the highest volume of cyber attacks in the US and worldwide. The economy has been rocked recently, and that is often a predictor of cyber attacks. As President Trump has pulled out of the

Paris Climate Accord, has considered pulling out of NAFTA, and recently imposed large tariffs on steel, aluminum, and other goods to allies in the EU, Canada, and Mexico [53]. The announcement of these tariffs recently dipped dramatically, as the Wall Street Journal notes, "Nasdaq was down 6.54 percent, marking their biggest weekly percentage falls since January 2016" [32]. Dramatic US economic changes like these will likely precipitate a higher volume of cyber attacks, as I occurred with the Chinese stock market crash described in section 6.1.

In the future, it would be very helpful obtain more accurate estimates of the numbers and locations of cyber attacks, either using a private database, similar to Carley's work [33], or by devising a way to reproduce similar results through open means. Additionally, correlating cyber attack volume with socioeconomic data on more than an annual basis, perhaps weekly, would bring greater clarity to how a changing socioeconomic climate directly affects cyber attack volume. Additionally, it may be helpful to store cyber attacks in a graph database, like Neo4j, which values relationships between entities over fast, uniform data access. In graph databases, entities are represented by vertices, and relationships are represented by edges. For instance, it is possible to use Neo4j's Cypher query language to detect the longest cycles of aggression in the following way:

```
MATCH p = (attacker:entity)-[launches]->(a:Attack)
-[targets]->(victim:entity)-[launches]->[b:Attack]
->[targets]->(attacker:entity)
WHERE ALL(a.attackDate < (b.attackDate)
AND ALL(b.attackDate - a.attackDate < 30)
RETURN p
ORDER BY length(p) DESC;
```

This of course, requires knowledge of the attacker, which is generally unknown in

the data from Hackmageddon, and may also be unknown in the Symantec telemetry database used in Carley's work [33]. However, DDoS reports from Digital Attack Map [6] generally do identify at least the country of origin, which could be taken advantage of with a graph database.

BIBLIOGRAPHY

[1] About the correlates of war project correlates of war. http://www.correlatesofwar.org/. (Accessed on 06/03/2018).

[2] Chinas stockmarket crashes again - open the door to green. https://www.economist.com/finance-and-economics/2016/01/04/chinas-stockmarket-crashes-again. (Accessed on 05/27/2018).

[3] Country analysis, industry analysis - market risk assessment. https://www.eiu.com/home.aspx. (Accessed on 06/03/2018).

[4] Credit insurance group — credendo. https://www.credendo.com/. (Accessed on 06/03/2018).

[5] Cyber crime costs projected to reach $2 trillion by 2019. https://www.forbes.com/sites/stevemorgan/2016/01/17/cyber-crime-costs-projected-to-reach-2-trillion-by-2019/#168014803a91. (Accessed on 05/19/2018).

[6] Digital attack map. http://www.digitalattackmap.com/#anim=1&color=0&country=ALL&list=0&time=17661&view=map. (Accessed on 05/27/2018).

[7] Downloads — world happiness report. http://worldhappiness.report/download/. (Accessed on 04/27/2018).

[8] Freedom house — championing democracy. https://freedomhouse.org/. (Accessed on 06/03/2018).

[9] Frequently asked questions — world happiness report. http://worldhappiness.report/faq/. (Accessed on 04/27/2018).

[10] International labour organization. http://www.ilo.org/global/lang--en/index.htm. (Accessed on 06/03/2018).

[11] report-ddos-trends-q42017.pdf. https://www.verisign.com/assets/report-ddos-trends-Q42017.pdf. (Accessed on 06/09/2018).

[12] Russia used twitter bots and trolls to disrupt brexitvote — news — the times. https://www.thetimes.co.uk/article/russia-used-web-posts-to-disrupt-brexit-vote-h9nv5zg6c. (Accessed on 05/17/2018).

[13] Transparency international - the global anti-corruption coalition. https://www.transparency.org/. (Accessed on 06/03/2018).

[14] Welcome to rsf website — rsf. https://rsf.org/en. (Accessed on 06/03/2018).

[15] World bank open data — data. https://data.worldbank.org/. (Accessed on 06/03/2018).

[16] eci.nic.in/eci_main1/current/press note ge-2014_05032014.pdf. http://eci.nic.in/eci_main1/current/Press%20Note%20GE-2014_05032014.pdf, May 2014. (Accessed on 05/27/2018).

[17] Statistical appendix for chapter 2 whr. https://s3.amazonaws.com/happiness-report/2015/StatisticalAppendixWHR3-April-16-2015.pdf, April 2015. (Accessed on 05/03/2018).

[18] Global device growth traffic profiles. https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-

forecast-highlights/pdf/Global_Device_Growth_Traffic_Profiles.pdf, 2016. (Accessed on 05/19/2018).

[19] Iot: number of connected devices worldwide 2012-2025 — statista. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/, November 2016. (Accessed on 05/31/2018).

[20] Amount of malware targeting smart devices more than doubled in 2017 — kaspersky lab. https://www.kaspersky.com/about/press-releases/2017_amount-of-malware-targeting-smart-devices-more-than-doubled-in-2017, June 2017. (Accessed on 06/09/2018).

[21] Cisco's vni predicts global annual ip traffic to exceed — the network. https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1853168, June 2017. (Accessed on 05/31/2018).

[22] Appendix1ofchapter2.pdf. https://s3.amazonaws.com/happiness-report/2018/Appendix1ofChapter2.pdf, March 2018. (Accessed on 05/17/2018).

[23] R. Agarwal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, 1994.

[24] M. Al-Maolegi and B. Arkok. An improved apriori algorithm for association rules. *arXiv preprint arXiv:1403.3948*, 2014.

[25] G. Aviles. How us political and socio-economic trends promotes hacktivist activity, 2015.

[26] T. W. Bank. Gdp per capita, ppp (current international $) — data. https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD, 2016. (Accessed on 04/27/2018).

[27] T. Bergin and N. Layne. Special report: Cyber thieves exploit banks' faith in swift transfer network — reuters. https://www.reuters.com/article/us-cyber-heist-swift-specialreport-idUSKCN0YB0DD, May 2016. (Accessed on 05/15/2018).

[28] M. Bodley. Why montana is going backward on medical marijuana. https://www.nbcnews.com/news/us-news/why-montana-going-backward-medical-marijuana-n410081, August 2015. (Accessed on 05/27/2018).

[29] D. Borak and K. Vasel. The equifax hack could be worse than we thought. http://money.cnn.com/2018/02/09/pf/equifax-hack-senate-disclosure/index.html, February 2018. (Accessed on 05/31/2018).

[30] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record*, 26(2):255–264, 1997.

[31] W. J. BROAD, J. MARKOFF, and D. E. SANGER. Stuxnet worm used against iran was tested in israel - the new york times. https://www.nytimes.com/2011/01/16/world/middleeast/16stuxnet.html?pagewanted=all, January 2011. (Accessed on 05/31/2018).

[32] S. Carew. Wall street nosedives as investors flee on trade war fears — reuters. https://www.reuters.com/article/us-usa-stocks/wall-street-nosedives-as-investors-flee-on-trade-war-fears-idUSKBN1GZ1O8?il=0, March 2018. (Accessed on 06/13/2018).

[33] K. M. Carley, G. Mezzour, and L. Carley. Global mapping of cyber attacks. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA PITTSBURGH United States, 2014.

[34] B. Chappell. U.s. says north korea 'directly responsible' for wannacry ransomware attack : The two-way : Npr. https://www.npr.org/sections/thetwo-way/2017/12/19/571854614/u-s-says-north-korea-directly-responsible-for-wannacry-ransomware-attack, December 2017. (Accessed on 05/31/2018).

[35] J. Clark. Why cyber security ignorance is bliss for iot hackers - computer business review. https://www.cbronline.com/internet-of-things/cyber-security-ignorance-bliss-iot-hackers/, March 2017. (Accessed on 05/31/2018).

[36] J. Cohen. Statistical power analysis for the behavioral sciences 2nd edn, 1988.

[37] C. Debeck. *The Correlates of Cyber Warfare: A database for the modern era*. Iowa State University, 2011.

[38] L. Downes. A victory for medical marijuana - the new york times. https://takingnote.blogs.nytimes.com/2015/05/22/a-victory-for-medical-marijuana/, May 2015. (Accessed on 05/27/2018).

[39] S. Erlanger. Britain votes to leave e.u.; cameron plans to step down - the new york times. https://www.nytimes.com/2016/06/25/world/europe/britain-brexit-european-union-referendum.html, June 2016. (Accessed on 05/21/2018).

[40] D. B. Figueiredo Filho, R. Paranhos, E. C. d. Rocha, M. Batista, J. A. d. Silva Jr, M. L. W. D. Santos, and J. G. Marino. When is statistical significance not significant? *Brazilian Political Science Review*, 7(1):31–55, 2013.

[41] C. Forrest. 80% of iot apps not tested for vulnerabilities, report says - techrepublic. https://www.techrepublic.com/article/80-of-iot-apps-not-tested-for-vulnerabilities-report-says/, January 2017. (Accessed on 05/31/2018).

[42] R. Gandhi, A. Sharma, W. Mahoney, W. Sousan, Q. Zhu, and P. Laplante. Dimensions of Cyber-Attacks: Cultural, Social, Economic, and Political. *IEEE Technology and Society Magazine*, 30(1):28–38, 2011.

[43] I. Gutzmer. Equifax announces cybersecurity incident involving consumer information — equifax. https://investor.equifax.com/news-and-events/news/2017/09-07-2017-213000628, September 2017. (Accessed on 05/31/2018).

[44] K. Hall. Iranian hackers targeted new york dam, had a quick nosy around the register. http://www.theregister.co.uk/2015/12/21/iranian_hackers_target_new_york_dam/, December 2015. (Accessed on 05/28/2018).

[45] A. Hearn. Hacking risk leads to recall of 500,000 pacemakers due to patient death fears — technology — the guardian. https://www.theguardian.com/technology/2017/aug/31/hacking-risk-recall-pacemakers-patient-death-fears-fda-firmware-update, August 2017. (Accessed on 05/31/2018).

[46] J. Helliwell, R. Layard, and J. Sachs. World happiness report. 2012.

[47] Hiscox. https://www.hiscox.com/sites/default/files/content/2018-hiscox-cyber-readiness-report.pdf. https://www.hiscox.com/sites/default/files/content/2018-Hiscox-Cyber-Readiness-Report.pdf, February 2018. (Accessed on 05/30/2018).

[48] I. Ilascu. Anonymous dumps email, facebook accounts and card data of israelis. https://news.softpedia.com/news/Anonymous-Dumps-Email-Facebook-Accounts-and-Card-Data-of-Israelis-477710.shtml, April 2015. (Accessed on 05/21/2018).

[49] R. L. John Helliwell and J. Sachs. World happiness report 2015 — world happiness report. http://worldhappiness.report/ed/2015/, 2015. (Accessed on 04/27/2018).

[50] S. Kumar and K. M. Carley. Approaches to understanding the motivations behind cyber attacks. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 307–309. IEEE, 2016.

[51] S. Kumar and K. M. Carley. Ddos cyber-attacks network: Who's attacking whom. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, pages 218–218. IEEE, 2016.

[52] D. K. Lambert and B. A. McCarl. Risk modeling using direct solution of nonlinear approximations of the utility function. *American Journal of Agricultural Economics*, 67(4):846–852, 1985.

[53] H. Long. Trump has officially put more tariffs on u.s. allies than on china - the washington post. https://www.washingtonpost.com/news/wonk/wp/2018/05/31/trump-has-officially-put-more-tariffs-on-u-s-allies-than-on-china/?noredirect=on&utm_term=.c82d97290afc, May 2018. (Accessed on 06/13/2018).

[54] M. G. Marshall. Polity iv project: Home page. http://www.systemicpeace.org/polity/polity4x.htm. (Accessed on 06/03/2018).

[55] W. McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.", 2012.

[56] G. Mezzour, L. R. Carley, and K. M. Carley. Longitudinal analysis of a large corpus of cyber threat descriptions. *Journal of Computer Virology and Hacking Techniques*, 12(1):11–22, 2016.

[57] S. Morgan. Cybercrime damages $6 trillion by 2021. `https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/`, November 2017. (Accessed on 05/30/2018).

[58] S. D. S. Network. World happiness report — kaggle. `https://www.kaggle.com/unsdsn/world-happiness/data`, 2017. (Accessed on 04/27/2018).

[59] P. Passeri. About me hackmageddon. `https://www.hackmageddon.com/about/`. (Accessed on 04/20/2018).

[60] P. Passeri. Hackmageddon information security timelines and statistics. `https://www.hackmageddon.com/`. (Accessed on 05/27/2018).

[61] P. Passeri. 1-15 january 2018 cyber attacks timeline hackmageddon. `https://www.hackmageddon.com/2018/02/06/1-15-january-2018-cyber-attacks-timeline/`, February 2018. (Accessed on 05/06/2018).

[62] N. Sadan. Press note ge-2014_05032014.pdf. `http://eci.nic.in/eci_main1/current/Press%20Note%20GE-2014_05032014.pdf`, May 2014. (Accessed on 05/17/2018).

[63] A. Sharma, R. Gandhi, Q. Zhu, W. R. Mahoney, and W. Sousan. A social dimensional cyber threat model with formal concept analysis and

fact-proposition inference. *International Journal of Information and Computer Security*, 5(4):301–333, 2013.

[64] P. C. Solano and A. J. R. Peinado. Socio-economic factors in cybercrime: Statistical study of the relation between socio-economic factors and cybercrime. In *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–4, June 2017.

[65] I. Thompson. Hackers can turn web-connected car washes into horrible death traps. `https://www.theregister.co.uk/2017/07/27/killer_car_wash/`, July 2017. (Accessed on 05/31/2018).

[66] N. Watson. Mitigating losses from a trillion dollar cybercrime industry. `https://www.entrepreneur.com/article/304971`, November 2017. (Accessed on 05/30/2018).

[67] K. Zetter. An unprecedented look at stuxnet, the world's first digital weapon — wired. `https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/`, November 2014. (Accessed on 05/31/2018).

[68] K. Zetter. Inside the cunning, unprecedented hack of ukraine's power grid — wired. `https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/`, March 2016. (Accessed on 05/31/2018).

APPENDICES

Appendix A

TABLES

Table A.1: **The country-to-country correlation and p-values of the top 20 targets of cyber attacks, arranged in ascending order by p-value.**

Country A	Country B	r	p-value
TR	PH	0.458	0.001
AU	IT	-0.366	0.009
IT	DE	-0.356	0.011
JP	CN	0.339	0.016
US	JP	0.326	0.021
SA	CN	0.324	0.022
IT	PK	-0.318	0.024
US	CN	0.309	0.029
RU	JP	0.305	0.031
IL	PK	0.297	0.037
CA	PK	-0.294	0.038
IN	PK	0.289	0.042
IN	CN	0.268	0.059
KR	NL	0.268	0.059
US	TR	0.265	0.063
IN	TR	0.264	0.064
IL	IT	-0.264	0.064
US	DE	0.251	0.079

Country A	Country B	r	p-value
US	RU	0.246	0.086
FR	PH	0.244	0.088
IT	CN	-0.241	0.091
RU	NL	0.241	0.092
GB	CA	0.23	0.109
GB	JP	-0.227	0.113
PK	NL	-0.216	0.131
RU	UA	0.215	0.133
IT	PH	0.215	0.134
RU	KR	0.214	0.135
IN	FR	-0.213	0.137
US	INDIA	0.212	0.14
KR	FR	-0.211	0.142
IN	IT	-0.21	0.143
US	PH	0.209	0.145
IL	SA	0.202	0.16
IL	KR	-0.197	0.17
DE	NL	0.195	0.176
IL	CN	0.194	0.177
UA	IT	0.191	0.185
TR	CN	0.19	0.186
PK	CN	0.19	0.186
IN	SA	0.188	0.191
KR	CN	0.186	0.196
UA	PK	0.185	0.197

Country A	Country B	r	p-value
GB	AU	0.182	0.205
AU	IL	0.178	0.217
KR	UA	0.174	0.227
AU	JP	-0.174	0.228
IL	DE	-0.174	0.228
RU	CN	0.172	0.231
IN	UA	0.17	0.237
JP	NL	0.169	0.241
UA	BR	-0.167	0.246
GB	KR	-0.167	0.247
GB	DE	0.165	0.252
IL	NL	-0.165	0.253
AU	UA	-0.164	0.254
FR	JP	-0.161	0.265
NL	PH	-0.158	0.274
CA	NL	0.157	0.275
IL	BR	0.157	0.277
CA	AU	0.156	0.279
GB	SA	0.148	0.306
GB	CN	-0.147	0.308
BR	SA	0.146	0.31
UA	CN	0.146	0.312
AU	DE	0.145	0.314
NL	SA	0.143	0.322
CA	KR	-0.143	0.323

Country A	Country B	r	p-value
KR	TR	-0.14	0.331
IT	SA	-0.139	0.334
IN	DE	0.138	0.34
TR	SA	0.138	0.34
CA	IT	0.134	0.354
GB	TR	-0.132	0.359
AU	SA	0.131	0.366
CA	UA	-0.125	0.388
FR	NL	-0.124	0.393
JP	IT	-0.123	0.395
AU	RU	-0.123	0.396
GB	RU	-0.12	0.408
CA	SA	0.116	0.421
GB	NL	-0.116	0.422
FR	TR	0.114	0.429
CA	TR	-0.113	0.436
IN	BR	-0.112	0.44
IT	TR	0.109	0.45
CA	BR	-0.109	0.452
RU	PK	-0.108	0.454
DE	TR	-0.106	0.466
PK	PH	-0.104	0.471
AU	KR	-0.103	0.477
KR	JP	0.101	0.485
AU	FR	0.098	0.5

Country A	Country B	r	p-value
IN	JP	0.098	0.5
IN	RU	0.097	0.501
CA	DE	0.097	0.504
CA	PH	0.095	0.512
US	FR	0.091	0.528
DE	PH	-0.091	0.53
GB	IL	-0.09	0.533
RU	TR	0.09	0.534
KR	PH	-0.09	0.535
KR	BR	-0.088	0.543
GB	INDIA	0.088	0.545
RU	IL	0.086	0.552
CA	FR	-0.086	0.555
KR	IT	-0.085	0.557
AU	PK	0.082	0.571
GB	UA	-0.079	0.586
US	SA	0.078	0.591
PK	BR	0.078	0.592
AU	BR	-0.077	0.595
US	IT	-0.077	0.595
US	UA	-0.077	0.595
DE	BR	0.076	0.598
IL	PH	-0.076	0.598
PK	TR	-0.076	0.6
FR	SA	-0.075	0.603

Country A	Country B	r	p-value
BR	TR	0.072	0.618
IN	AU	0.071	0.626
US	AU	0.07	0.628
SA	PH	-0.069	0.634
DE	PK	-0.068	0.64
RU	SA	-0.065	0.654
FR	BR	-0.064	0.658
CA	RU	0.058	0.688
AU	PH	-0.058	0.69
GB	IT	-0.057	0.695
JP	TR	0.056	0.7
DE	SA	0.055	0.703
GB	PK	-0.054	0.709
CA	CN	-0.053	0.713
US	IL	-0.052	0.717
FR	UA	0.052	0.72
IN	PH	-0.052	0.721
RU	DE	-0.051	0.723
CA	IL	-0.051	0.723
RU	PH	-0.051	0.726
US	NL	0.05	0.732
JP	SA	0.049	0.735
KR	DE	-0.048	0.743
GB	BR	-0.045	0.758
PK	SA	-0.044	0.761

Country A	Country B	r	p-value
IN	NL	-0.041	0.775
US	PK	-0.04	0.784
UA	SA	0.04	0.785
NL	CN	0.039	0.788
BR	PH	0.039	0.79
JP	PH	-0.037	0.801
IN	IL	0.036	0.803
AU	TR	-0.035	0.808
JP	DE	-0.032	0.826
IT	NL	-0.031	0.829
UA	DE	-0.031	0.831
US	GB	0.031	0.832
FR	CN	0.028	0.846
US	CA	0.027	0.852
CA	JP	-0.027	0.853
IT	BR	-0.026	0.856
AU	NL	0.026	0.856
FR	DE	0.025	0.861
US	BR	0.025	0.862
BR	CN	0.025	0.866
TR	NL	-0.023	0.872
IN	KR	0.023	0.876
UA	JP	-0.022	0.878
BR	NL	0.021	0.884
FR	PK	-0.019	0.896

Country A	Country B	r	p-value
RU	BR	-0.019	0.898
UA	PH	0.018	0.902
IN	CA	-0.017	0.908
RU	FR	-0.016	0.915
RU	IT	0.016	0.915
JP	BR	-0.015	0.916
FR	IT	0.015	0.916
UA	NL	-0.014	0.921
IL	FR	-0.014	0.924
UA	TR	-0.014	0.924
KR	SA	-0.013	0.928
DE	CN	-0.012	0.936
AU	CN	0.011	0.939
JP	PK	0.011	0.942
GB	PH	0.01	0.943
US	KR	0.01	0.947
KR	PK	-0.008	0.954
PH	CN	0.008	0.954
IL	JP	0.008	0.957
IL	UA	0.005	0.971
GB	FR	0.002	0.988
IL	TR	-0.001	0.993

A	Cons	A.S.	C.S.	S	Conf	Lift	Lev	Conv
TC: Industry	AC: Cyber Crime	0.581	0.854	0.563	0.968	1.133	0.066	4.584
TC: Government	AC: Cyber Crime	0.522	0.854	0.472	0.906	1.060	0.027	1.543
AC: Cyber Crime	Attack: Targeted	0.854	0.448	0.411	0.481	1.074	0.028	1.064
AC: Cyber Crime	AC: Cyber Espionage	0.854	0.429	0.390	0.457	1.064	0.024	1.051
AC: Cyber Espionage	Attack: Targeted	0.429	0.448	0.382	0.890	1.988	0.190	5.019

Table A.2: Top 5 apriori Rules associations for antecedent groups of size 1, in descending order by support. Column labels: "A" = antecedents, "Cons" = consequents, "A.S." = antecedent support, "C.S" = consequent support, "S"= support, "Conf" = confidence, "Lev" = leverage, "Conv" = conviction. Value labels: "AC" = Attack Class, "TC" = Target Class

A	Cons	A.S.	C.S.	S	Conf	Lift	Lev	Conv
AC: Cyber Espionage, AC: Cyber Crime	A: Targeted	0.390	0.448	0.347	0.889	1.987	0.172	4.998
AC: Cyber Crime, TC: Government	TC: Industry	0.472	0.581	0.337	0.713	1.227	0.062	1.460
AC: Cyber Crime, TC: Government	A: Targeted	0.472	0.448	0.273	0.578	1.292	0.062	1.310
TC: Industry, AC: Cyber Crime	TCo: GB	0.563	0.357	0.271	0.482	1.348	0.070	1.240
AC: Cyber Crime, A: Targeted	TC: Industry	0.411	0.581	0.271	0.660	1.136	0.032	1.232

Table A.3: Top 5 apriori Rules associations for antecedent groups of size 2, in descending order by support. Column labels: "A" = antecedents, "Cons" = consequents, "A.S." = antecedent support, "C.S" = consequent support, "S"= support, "Conf" = confidence, "Lev" = leverage, "Conv" = conviction. Value labels: "TCo" = Target Country.

A	Cons	A.S.	C.S.	S	Conf	Lift	Lev	Conv
AC: Cyber Espionage, AC: CC, TC: Gov	A: Targeted	0.261	0.448	0.240	0.921	2.058	0.124	7.0150
TC: I, AC: CC, AC: CE	A: Targeted	0.248	0.448	0.226	0.909	2.031	0.115	6.076
TC: I, AC: CC, TC: Gov	A: Targeted	0.337	0.448	0.187	0.555	1.240	0.036	1.241
TC: I, TC: Gov	AC: CC	0.345	0.854	0.337	0.976	1.143	0.0421	6.123
AC: CE, AC: CC, TC: Gov	TC: I	0.261	0.581	0.179	0.685	1.179	0.027	1.330

Table A.4: Top 5 apriori Rules associations for antecedent groups of size 3, in descending order by support. Column labels: "A" = antecedents, "Cons" = consequents, "A.S." = antecedent support, "C.S" = consequent support, "S"= support, "Conf" = confidence, "Lev" = leverage, "Conv" = conviction. Value labels: "TCo" = Target Country, "CC" = Cyber Crime, "E" = Espionage, "Gov" = Government, "I" = Industry .