TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Aili Juhasoo-Lawrence 184917IVSB

Social Media Scraping for Cybersecurity: Performing Open-Source Intelligence with Twitter

Bachelor's thesis

Supervisor: Kaido Kikkas
Doctor of Philosophy (PhD) in Engineering

Co-Supervisor: Siim Kurvits
Bachelor of Science in Gene Technology

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Aili Juhasoo-Lawrence 184917IVSB

Sotsiaalmeedia kaapimine küberturvalisuse eesmärkidel: andmekogumine avalikest allikatest Twitteri näitel

bakalaureusetöö

Juhendaja: Kaido Kikkas
Tehnikateaduste doktor

Kaasjuhendaja: Siim Kurvits
Geenitehnoloogia bakalaureus

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Aili Juhasoo-Lawrence

17.05.2021

Abstract

Currently the amount of solutions for extracting cyber threat intelligence from social media platforms are limited. It is clear that cyberspace has entered a rapid growth which is largely contributed to by social media. Proportionately, the impact and volume of cyberattacks has also seen an unprecedented increase. The research's purpose is to explore the potential of performing open source intelligence with social media for cyber defence. Furthermore, this thesis aims to propose a possible solution for extracting social media data for employment in cyber threat intelligence.

To test whether open-source intelligence techniques can be used effectively with social media to gather information about threats, this research uses a case study format by applying web scraping technologies to Twitter. The results suggest that Twitter can in some cases be a suitable platform for statistically tracking cybersecurity events, however in most cases the results are inconclusive. Moreover, it is found that cyber threat intelligence can be extracted from scraped Twitter data. The research concludes that although scraping and extracting cybersecurity information from Twitter is a satisfactory cyber defence technique, there are existing solutions that surpass this method in terms of processing and memory requirements.

This thesis is written in English and is twenty six pages long, including five chapters, seven figures and three tables.

Annotatsioon

Sotsiaalmeedia kaapimine küberturvalisuse eesmärkidel: andmekogumine avalikest allikatest Twitteri näitel

Interneti kasutajad satuvad tänapäeval aina tugevamate ja sagedasemate küberrünnakute alla. Üheks kaitsemeetmeks on teabekogumine küberohtude kohta - sellise luure eesmärk on leida infot vastase tuvastamiseks ja ohu kõrvaldamiseks. Käesoleval ajal on vähe lahendusi ohuteabe kogumiseks sotsiaalmeediast, ehkki seda kasutavad küberturbespetsialistid kogutud info levitamiseks laialdaselt. Käesoleva töö eesmärgiks on uurida andmete kogumise võimalusi avalikest allikatest Twitteri näitel ning pakkuda välja lahendus ohuteabe kogumiseks sotsiaalmeediast.

Meetodina kasutatakse siin veebikaapimist (tuntud ka veebikraapimise või -koorimisena), mille tõhusust hinnatakse kahes osas. Esmalt analüüsitakse seost Twitteris mainitud küberturvalisuse märksõnade kasutuse ning tõendatud küberintsidentide vahel. Seejärel uuritakse intsidentidele viitava materjali eraldamisvõimalusi Twitteri ülejäänud andmetevoost.

Analüüs näitab, et Twitterit on mõnel juhul võimalik kasutada küberjuhtumite jälgimiseks, ent valdavalt ei ole sealt kaabitud andmetel ning dokumenteeritud küberintsidentidel statistilist seost. Samuti leitakse, et intsidendiinfo eraldamine Twitterist on võimalik vaid juhul, kui kasutusel on täpsed reeglid kõikide andmetes esinevate anomaaliatega arvestamiseks. Töös järeldatakse, et kuigi Twitteri kaapimist on võimalik korrektsete protseduuride abil küberohuteabe kogumiseks kasutada, ei pruugi see viis olla kõige tõhusam, kuna nõuab arvestataval määral ressursse andmesideks, salvestuseks ja -töötluseks.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti kahekümne kuuel leheküljel, viit peatükki, seitset joonist, kolme tabelit.

List of abbreviations and terms

APT	Advanced Persistent Threat
CSV	Comma-Separated Values
CTI	Cyber Threat Intelligence
Cyberspace	The online computer networks, mainly the Internet
Cyberwarfare	Network-based conflict
Defanging	Refers to the process of changing a URL or IP address so that it cannot be clicked
DNS	Domain Name System
IOC	Indicator of Compromise
OSINT	Open-Source Intelligence
Scraping	Web scraping is the process of copying data from a web page resulting in scraped data
Tweet	Microblog post on the platform Twitter

Table of contents

List of figures

List of tables

1 Introduction

Cybersecurity is on the rise to becoming a key infrastructure component in company, government and military networks due to the rapid growth of cyberspace. One of the central constituents contributing to this growth are social networks which are exemplary illustrations of the thriving online population. One of the methods of combating destructive events in cyberspace is by gathering cyber threat intelligence (CTI) from publicly available sources, also known as open-source intelligence (OSINT), and applying it to cybersecurity mitigation techniques. Various social media platforms and channels can also be used for this purpose, as online cybersecurity communities are active in sharing their knowledge and findings with other members online.

This study seeks to retrieve cyber threat intelligence shared on the social media platform Twitter. In this research, one method of information gathering from publicly available sources is used with the goal of finding data applicable to the cybersecurity division. In order to achieve this goal, web scraping technologies are combined with data filtering and analysis tools. Therefore the focus of this study is to collect and analyse CTI information from social media. The final output of the research is an evaluation of whether web scraping is an effective method of extracting CTI from social media.

Due to rapid global increase in cybercrime, social media usage and online data volume, researchers and cybersecurity officials are in need of guides for utilizing said social media data to mitigate threats. An objective of this research is to be a contribution to the fast-growing cybersecurity industry and the challenges posed by the correspondingly rising number of cyber threats. Therefore the study aims to propose a solution to retrieving and extracting CTI from the vast amounts of data on social media for employment in proactive cyber defence.

1.1 Problem Statement

Social media today is a growing source of data in many fields, including cybersecurity. As cybersecurity is only becoming more relevant in the current information society, it would be beneficial for the community working in this field to have means to employ social media in their CTI operations. However, there are little proposed solutions on how to successfully extract and make use of CTI data found on social media platforms. The

aim of this research is to impart a possible solution for using social media as an OSINT tool, which would aid cybersecurity officials in conducting threat intelligence.

This thesis will explore the problem by attempting to answer the research question: **"How can Twitter be used for threat intelligence in order to enrich the capabilities of proactive cyber defence?"**

1.2 Hypothesis

The third chapter of this research paper investigates the following hypothesis in order to establish whether Twitter is a valid tool for social media OSINT:

The cybersecurity incident data scraped from Twitter is proportionate to recorded cybersecurity events.

1.3 Scope and Limitations

This research explores whether and how social media can be used in cyber defence in conjunction with OSINT. However only the Twitter platform is used for the research as a case study. Additionally, this study considers only web scraping OSINT techniques and does not cover all aspects of the discipline.

1.4 Thesis Outline

This thesis contains the following four chapters:

1. **Research Premises** gives context and background to this study. Its final subchapter is the research methodology.

2. **Analysis** evaluates historic Twitter data to judge its effectiveness as an OSINT tool for CTI. This chapter analyses Twitter's feed in relation to widespread attack types, aggressive threat actors and prevalent malware.

3. **Twitter Scraping for Defensive Cybersecurity** attempts to extract CTI that could aid in protecting systems from cyber threats.

4. **Summary** chapter discusses the conclusion and key takeaways of the research.

2 Research Premises

The purpose of this chapter is to provide background information on the topic in order to highlight the significance and relevance of this research. The first subchapter discusses the research subject matter in a broader context and attempts to discuss the issue from various points of significance. The second subchapter outlines the methodology of this research which is derived from the initial premise analysis of the topic.

2.1 Background

This section gives background information on the research topic by examining war in cyberspace, CTI, social media's relevance, OSINT and data gathering techniques. By doing so, this section of the thesis attempts to place the subject in context and explain its relevance.

2.1.1 Modern Warfare

The online population is on the increase and as a result, the number of cyberattacks alike. A report by Cybersecurity Ventures estimates that cybercrime damages could total to six trillion United States dollars (USD) in 2021 [1]. However while the costs of damages caused by cyberattacks are high, the business of carrying out cyberattacks is very profitable. Deloitte finds that a smaller campaign could cost as little as 34 USD per month while returning as much as 25,000 USD. They further estimate that a more expensive operation costing 3,800 USD per month could return up to one million USD [2]. Moreover, one does not even require a high level of technical skills anymore to carry out a cyberattack, various such services can be bought on the Dark Web. Editor in CSO Online, Dan Swinhoe, says "The low cost of entry, relative ease with which attacks can be deployed, and the high returns means the potential pool of threat actors isn't limited by technical skill level." [3] With expenses imposed by cyber threats and the growing trend of carrying out cyberattacks becoming more accomplishable, it is undeniable that cyberwarfare could become a key issue in the coming years.

In traditional warfare, it is usually up to a government's political powers to instate peace or war, however in the age of cyberwarfare, this has changed. Today not only governments can attack other governments in cyberspace, but also non-state sponsored citizen militias are able to launch cyberattacks against a country's infrastructure or even international companies. This means that "states are no longer the sole masters of international security" [4] and the definition of warfare is blurrier than ever. Furthermore, information operations are recognised as the fifth and newest dimension of warfare next to land, sea, air and space [5]. General Larry D. Welsh from the United States Air Force adds that cyberspace is embedded in all of the domains of warfare [6] which emphasises the scope of its influence. The result of this shift of conflict to cyberspace and the opportunities it has opened for normal citizens, further stresses the gravity of the impact that cyber threats pose.

Social media has equalised the battlefield when it comes to military and civil disagreements. Michael Erbschloe discusses in his book how social media has allowed warfare for "social, cultural, economic, and religious factions around the world" [7] as well as governments. Another book on the weaponization of social media extends this idea, by debating that the platforms are being used as "sophisticated weapon systems" [8]. While established that social media has become somewhat of a battleground for adversaries as it can be used for manipulation and propaganda, phishing and coordination, it also has great potential as a tool to combat cybercrime. This thesis will address uses of social media as a tool for defensive cybersecurity.

2.1.2 Cyber Threat Intelligence

Threat intelligence takes on an essential role in a present-day cybersecurity department of an organisation. CTI is collected information that is analysed and employed for reconnaissance of cyber threats, development of defence systems and security related decision making processes [9]. This data is held to high regard in the cybersecurity community due to its perceived benefits to various stakeholders. CTI gathering can be divided into three categories that describe the goal of said information [10]:

1. **Tactical CTI** focuses on mitigating near future or current threats. One of the primary methods for achieving this is by gathering indicators of compromise (IOCs) which can be directly applied to detection and prevention systems.

2. **Strategic CTI** is used for making decisions regarding future security techniques of an organisation and therefore focuses on emerging threat trends and analysis of the behaviour of adversaries.

3. **Operational CTI** attempts to understand the adversaries capabilities, motivations and associations. This knowledge can assist in resource allocation and prioritisation decisions.

This research mainly concerns tactical CTI as it aims to uncover data that can be directly applied to proactive cyber defence techniques.

In order to keep an organisation's CTI operations continually improving and ensure that an appropriate response is given to the current threat landscape, a framework is provided by the threat intelligence lifecycle [9]. Although varying in amount of steps, the intelligence cycle follows a logical structure:

1. The requirements for the CTI are established which is necessary for a plan of action.

2. The CTI data is gathered and processed.

3. The CTI data is analysed in order to determine its suitability to the organisation's systems.

4. An operational solution is produced.

5. The completed product is evaluated by stakeholders.

Within limits, this process is also followed in the research at hand with the exception, that the final product is evaluated by the author.

2.1.3 Social Media Data Scraping

Social media has become a vast ground of information exchange which only continues to increase in size. Social media is defined as "forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos)" [11]. As of January 2021, there is an estimated 4.2 billion active social media users, which is an increase of 13.2% since 2020 [12]. Moreover, Cisco predicts

that global IP traffic per month will reach 396.0 EB (exabytes) by the year 2022, marking a threefold increase of data volume compared to 2017 [13]. Therefore it can be assumed, that the growth of social media usage has also contributed to the growth of traffic volume.

Social media data is lucrative and described by researchers as “clearly the largest, richest and most dynamic evidence base of human behavior” [14]. It is possible to find a myriad of communities represented on social media, including the vast cybersecurity community, who are known for collaboration and information exchange. This points to the importance of having a means of using social media effectively as a tool in CTI.

One of the methods of extracting data from social media is web scraping. This process refers to copying data from a website and storing it locally in a structured format. The benefit of scraping data as opposed to streaming data directly from a website without saving it, is that the researcher can later return to the data to perform further analytics. There are some ethical concerns surrounding web scraping which predominately discuss the further dishonest application of the saved data, copyright issues and matters related to privacy [15]. However this research does not concern these issues, as no connections are made between an author and their posted content, making the dataset anonymous. Additionally, this study uses web scraping on a relatively small scale for academic purposes.

2.1.4 Social Media Open Source Intelligence

OSINT is a type of intelligence that is gathered from publicly available sources, such as the news, published reports and social media [16]. Although OSINT can have a variety of applications ranging from personal investigations to military operations, it is possible to categorise its employments into three parts [17]:

1. Social opinion and sentiment analysis.
2. Cybercrime and organised crime.
3. Cybersecurity and cyber defence.

One article argues that the reason OSINT has become so popular across different sectors, is that the risks and costs that come with gathering public information are low [18] compared to the value they bring to the table. A perceived limitation of OSINT is the

challenges posed by the quantity of data available in public sources [17] and the difficulties that come with finding methods of extracting what is necessary. This research mainly focuses on OSINT regarding cybersecurity and cyber defence and attempts to find a solution to the aforementioned limitation.

James M. Davitch, a lieutenant colonel in the United States and chief of the Intelligence Operations Division, stresses the importance of using public sources for intelligence gathering and its benefits in tactical response. In his article he highlights that there is a biased approach towards OSINT in the intelligence community, because classified information appears to be more advantageous due to its privileged nature [19]. However OSINT is not only a technique employed by the military where such biases may still exist, but also a tool that is largely used in circles that do not even have access to classified data. The costs of damage due to cybercrime are at an estimated growth rate of 15% a year [20]. Therefore it is more essential than ever for cybersecurity officials to make use of every resource available for developing and maintaining intrusion detection systems and implementing intelligence-driven incident response [16].

CTI is a reconnaissance method used to "prepare, prevent, and identify cyber threats looking to take advantage of valuable resources" [21]. The key concept of CTI is intelligence, as this data is essential for combatting the adversary in the cyber landscape. Scott J. Roberts and Rebekah Brown claim in their book *Intelligence-Driven Incident Response* that the side that has allocated resources to intelligence gathering and inspection will almost always be at an advantage [16]. Their book further outlines how CTI plays an unprecedented role in cyber incident response and brings out OSINT as one of the main intelligence gathering methods out there [16]. Additionally, L. Kello even goes as far as to say "Information in no longer just a source of power; it has become force itself" [4], further highlighting the significant role that intelligence plays today. This suggests that besides OSINT being an effective way of gathering data, it also has significant applications in the field of cyber defence.

Social media and cybersecurity communities online are at a rise which gives grounds for driving focus towards using social media as a tool in cybersecurity and cyber defence. Moreover, it stresses the potential value of intelligence gathered on social media in various techniques applied towards the adversary. This is why this research proposes a

solution for using OSINT in social media and attempts to prove the value of this concept in the cybersecurity division.

2.1.5 Twitter Intelligence

Twitter is a social media platform and microblogging website ideal for OSINT in many different areas of study, including cybersecurity. Other social media platforms were also considered for this research, however after initial groundwork it became apparent that Twitter is unmatched for its potential in CTI. As of January 2021, there is an estimated 353 million active Twitter users and Twitter ranks in the top most used social media platforms [12]. Additionally, there is an average of 500 million tweets posted every day according to 2020 statistics [22]. Therefore the volume of data on Twitter is substantial which increases the possibility of finding information of merit.

One of the reasons for Twitter being a great tool for OSINT and even superior in some instances to other social media platforms is its search functionality. Twitter has the following search options [23]:

- **Words**: exact words, exact phrases (AND clause), any of a selection of words (OR clause), word exclusion, specific hashtag and language selection.
- **People**: from a specific account, replies to a specific account and mentions of a specific account.
- **Places**: tweets sent from specific geographic location.
- **Dates**: within a specified date range, before a specific date, after a specific date.

These comprehensive search filter options are uncommon for social networks and make Twitter unique in the OSINT field. Twitter also grants developer and academic research accounts which allow users to use Twitter's API. This indicates that Twitter, in a way, is designed to accommodate researchers, intelligence gatherers and data analysts. For these various reasons, this research uses Twitter as a case study to represent social media more broadly.

2.2 Methodology

The objective of this research is to firstly prove that Twitter is a social media platform that can be effectively used for CTI gathering. Furthermore, this study aims to propose a possible solution for OSINT with Twitter. For this purpose, the thesis employs a quantitative method by scraping Twitter data and performing data analytics. Before data is scraped, a manual evaluation of Twitter posts is done by searching for keywords relating to cybersecurity and mapping out relationships between them. For the purpose of this research, Twitter data is scraped based on keywords and date ranges. The copied data is then used to firstly, analyse Twitter's cybersecurity capabilities and secondly, to propose a means of its application to proactive cyber defence.

For scraping Twitter this research uses a Python package *snscrape* [24] which has the ability to copy data into comma-separated values (CSV) format based on command-line arguments or running scraping scripts. The latter was chosen for this research, where script files were used containing time period and keyword arguments. Initial manual groundwork of Twitter resulted in the following three keyword combinations for scraping, which retrieve tweets in the time range of 2016 to 2020:

1. "Cybersecurity" and "attack".
2. "Malware", and "campaign" or "bot" or "botnet" or "attack".
3. "Ransomware".

These data sets are used with different combinations in the analysis section of this thesis. For data analytics, filtering and intelligence extraction, this research uses Python 3.8.5 and predominantly Python packages *pandas 1.2.3*, *numpy 1.19.2* and *matplotlib 3.4.0*. In the first step of this process, copied data is quantified based on mentions of specific cyberattack types, especially aggressive threat actors, prevalent malware types and notorious ransomware. The results of these searches are compared to professional malware, incident and cybersecurity reports from various sources with the goal to discover, whether events that these sources deem significant reflect similarly in Twitter data. The second step involves using the same scraped data in combination with regular expression patterns to extract pieces of information applicable to proactive cybersecurity, such as IOCs and vulnerabilities.

3 Analysis

The purpose of this chapter is to evaluate data copied from Twitter for its content's correspondence with significant cybersecurity events, cyberattack types and most damaging threat actors. This research attempts to realise whether public knowledge of these types of occurrences in cyberspace correspond proportionately to their reactions and recordings on Twitter.

3.1 Widespread Attack Types

The most common cyberattack types according to Cisco are malware, phishing, Man-in-the-Middle, Denial-of-Service, SQL injection, zero-day and DNS tunneling attacks [25]. This section of the research evaluates the volume of mentions these attack types have on Twitter from 2016 to 2020. For the purpose of this investigation, Twitter data copied with the keywords "cybersecurity" and "attack" is used.

Using Python, the CSV files are searched for mentions of these attack types and the results are counted and aggregated. Figure 1 depicts a statistic derived from the results as a bar chart demonstrating the number mentions of each attack type in the dataset and grouped by year. The number of mentions does not necessarily correspond to the number of tweets, because mention of a specific malware type may occur several times in one post.

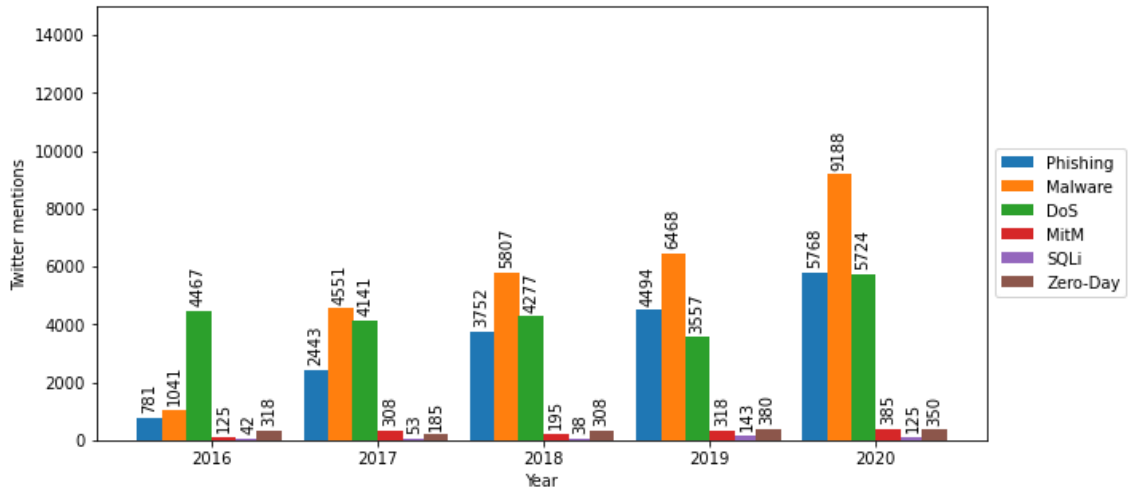

Figure 1. Twitter mentions of most common attack types grouped by year.

From the results it appears that mentions of Denial-of-Service attacks remain at a relative constant throughout the years while malware attacks see the biggest rise from 2016 to

2020. The mentions of DNS tunneling are so insignificant that they do not appear in Figure 1. Man-in-the-Middle attacks, SQL injections and Zero-Day exploits are mentioned at a continual low rate throughout this Twitter dataset.

According to 2017 statistics Denial-of-Service attacks were indeed higher than malware attacks in 2016 [26], but not by such a large degree as appears in Figure 1. Additionally, the same statistics present that SQL injection attacks were also relatively high which is not visible from Figure 1. An Online Trust Alliance report of 2017, estimates a 90% increase of ransomware attacks that year [27] which could explain the surge in malware Twitter mentions from 2016 to 2017. The 2018 and 2019 similarities depicted in Figure 1 are almost reflected in a 2021 SonicWall cyber threat report [28]. The same report however also estimates a 43% drop in malware attacks and a 66% increase in ransomware attacks from 2019 to 2020. Since ransomware is also a type of malware, then this statistic is difficult to apply to the results of this research, however a significant growth in ransomware attacks may explain the spike of Twitter mentions of malware in 2020.

Overall, the results of this section are inconclusive. Although there are some similarities in cyber threat reports and Twitter statistics of these years, there are also quite a few disagreements. In some cases the results may be perceived to reflect contents of cybersecurity reports, however mostly they do not seem to correspond. Therefore, it is likely, that the results depicted in Figure 1 can be attributed to other factors, such as the growth of the online population and social media users or overall rise of cybersecurity awareness.

3.2 Aggressive Advanced Persistent Threat Groups

Advanced Persistent Threat (APT) groups are described by Deloitte as "non-opportunistic" and "breaching organisations in a strategic, long-term manner with clear objectives" [29]. In practice, these groups are perceived to have unlimited resources due to them possibly being state-sponsored groups. An article defines the following APT groups as the most dangerous: APT1, APT41, APT35, APT33, APT38, APT37, APT28, APT29 and Equation Group [30].

This section of the research aggregates all of the data from Twitter gathered for this thesis and counts how many times these APT groups are mentioned on Twitter from 2016 to 2020. The APT groups also each have a number of other names they are known by, which are accounted for in this search. The synonymous names of the groups, which were also used in this research can be found in Appendix 2 – Alternative names for APT groups. The reason for the groups having several nicknames is that different CTI organisations categorise them with different schemes and cyber attacks cannot always be definitively attributed to one group.

Table 1. Number of mentions of APT groups on Twitter by year.

Year	**APT1**	**APT41**	**APT35**	**APT33**	**APT38**	**APT37**	**APT28**	**APT29**	**Eq.**
2016	0	0	0	0	1	3	26	5	6
2017	19	0	0	10	95	202	345	2	4
2018	511	0	4	231	1517	296	1577	128	121
2019	167	222	27	211	709	301	445	98	85
2020	68	117	61	43	828	54	758	260	83

According to Table 1 almost all of the APT groups see a significant spike in mentions in 2018. Additionally, some of the groups are not mentioned at all during the first three years of this Twitter dataset. The most mentions overall are won by APT38 and APT28. APT35 is mentioned the least over the years.

Because of the jump in mentions in 2018, Table 2 is brought as a reference and demonstrates the number of posts in the Twitter dataset and the number of unique users.

Table 2. Number of Twitter posts and number of different users posting by year.

Year	Nr. Of Posts	Nr. Of Users	Ratio of Users to Posts
2016	64822	22982	1 : 2.82
2017	113651	32466	1 : 3.49
2018	777358	155739	1 : 4.99
2019	673783	147652	1 : 4.56
2020	643857	156483	1 : 4.12

As seen in Table 2, the ratio of users to posts is on an average slightly larger during 2016 and 2017. However, the number of posts and users also sees a large increase during 2018 which could also explain the decrease in the ratio. Therefore the spike in APT mentions in 2018 is most likely caused by the overall increase in users tweeting about cybersecurity.

In 2017 and 2018 Lazarus group (APT38) became linked to the WannaCry 2.0 ransomware and in 2018 they were charged [31]. This event may be a contributing factor to the sudden surge in mentions of APT38 on Twitter in 2018. Russian APT group Cozy Bear (APT29) is linked to the supply chain attack in the end of 2020 [32] which may explain the increase in mentions of the group in 2020. The most Twitter mentions in 2016 is for Fancy Bear (APT28) where the group attempted to interfere with Hillary Clinton's campaign and also carried out extensive attacks against the World Anti-Doping Agency [33].

In spite of there being some evidence of the Twitter feed being successful at tracking APT activity and creating a timeline of the groups, it is not definitive. There are relatively few mentions of the groups in general considering the substantial role they play in the threat landscape.

3.3 Significant Ransomware Attacks

Since ransomware is one of the most prevalent malware types, the following section of the Twitter research attempts to map out mentions of infamous ransomware attacks conducted between 2016 and 2020. According to Kaspersky [34] and CrowdStrike [35], the following are the most influential ransomware campaigns of the past years:

- 2016 – Petya, Locky, Jigsaw, Dharma
- 2017 – Wannacry, Bad Rabbit, GoldenEye, BitPaymer and NotPetya
- 2018 – SamSam
- 2019 – DoppelPaymer, MedusaLocker and Revil
- 2020 – Ryuk

This information is used to filter through Twitter data from 2016 to 2020 which is scraped with the keyword “ransomware”. The different aforementioned ransomware attacks are searched for and the mentions are counted. This data is then used to create Table 3, which represents the ransomware mentions on Twitter in correspondence with the reported infamous cases from 2016 to 2020.

The integers in Table 3 show the number of mentions a ransomware had in the Twitter data. The column labelled “Total” shows how many times the specified ransomware was mentioned throughout the entire dataset. The row starting with “Total” shows the number of mentions of all searched ransomware attacks within a year. The percentages in Table 3 display the proportion that the mentioned ransomware has out of the total mentions of that year. The colours in Table 3 display how well the data retrieved from Twitter corresponds to the principal ransomware attacks carried out between 2016 and 2020 according to official reports. The red fields represent the years where the specific ransomware was recorded to have the most impact by cybersecurity reports. The dark blue fields speak for when the specific ransomware was reflected with the largest scale in terms of Twitter mentions. Finally, the purple fields constitute the overlapping of the red and the blue fields. Therefore it represents the instances where the reported impactful ransomware event is also mentioned the most on Twitter in the specified year.

Table 3. Most mentioned ransomware on Twitter from 2016 to 2020.

	2016	2017	2018	2019	2020	Total
Petya	12 727	55 532	3458	1259	628	73 604
	26.7%	22.4%	10.8%	6.2%	3.4%	
Locky	27 547	15 488	894	851	294	45 074
	57.9%	6.3%	2.8%	4.2%	1.6%	
Jigsaw	4989	306	398	299	119	6111
	10.5%	0.2%	1.3%	1.5%	0.6%	
Dharma	39	1591	1386	1188	756	4960
	0.1%	0.6%	4.3%	5.8%	4.1%	
Wannacry	0*	156 023	10 995	4555	1990	173 563
		62.8%	34.5%	22.4%	10.6%	
BadRabbit	0*	7776	236	18	13	8043
		3.2%	0.7%	0.1%	0.1%	
GoldenEye	1065	2452	28	15	8	3568
	2.3%	0.9%	0.1%	0.1%	0.1%	
BitPaymer	0*	349	719	1173	119	2360
		0.2%	2.3%	5.7%	0.6%	
NotPetya	0*	8318	1738	800	378	11 234
		3.4%	5.5%	3.9%	2.1%	
SamSam	1205	491	10571	715	130	13 112
	2.5%	0.2%	33.2%	3.5%	0.1%	
DoppelPaymer	0*	0*	0*	537	2237	2774
				2.6%	11.9%	
MedusaLocker	0*	0*	0*	212	102	314
				1.1%	0.5%	
REvil	2	23	18	1158	4457	5658
	0%	0%	0%	5.7%	23.8%	
Ryuk	6	11	1433	7594	7447	16 491
	0%	0%	4.5%	37.3%	39.8%	
Total	47 580	248 360	31 874	20 374	18 678	
	100%	100%	100%	100%	100%	
*	The ransomware did not exist at this point yet.					
	Most prevalent ransomware during specified year according to sources.					
	The ransomware was mentioned the most on Twitter in this year.					
	Most mentioned ransomware on Twitter matches data from sources.					

Table 3 demonstrates that Wannacry had an overwhelming number of mentions in 2017 with Petya as a follow up. Not only is it apparent from the data that Wannacry is the most mentioned ransomware on Twitter, but also the year of its release produced the largest amount of Twitter posts discussing ransomware in general. It is also clear that Wannacry is discussed years later after the main attack, indicating its long-lasting effect. Even though Locky was released in 2016 [36], its variants still can be found today which is also demonstrated in Table 3, as even though Locky mentions decrease, they do not drop

completely. Ryuk ransomware starts to make a more substantial appearance in 2018 [37], which is also apparent from this Twitter data.

As can be seen from Table 3, eight out of the fourteen searched ransomware occurrences are also reflected in the Twitter data with a corresponding magnitude. Furthermore, in most cases where the Twitter data of the ransomware attacks does not match the information gathered from reports, the mentions spike after the recorded event. This is a logical consequence, since malware often reoccurs after it's initial release. The only exception found in the data to this, is with the GoldenEye ransomware, where Twitter mentions of it are in a percentual manner highest before the sources suggest. However, even though mentions of GoldenEye make up a larger proportion in the 2016 data, there are still more counts of it in 2017. Since there are overall many more Tweets about ransomware in 2017, then as a result, GoldenEye makes up for a smaller division in that time.

The results of this search of the scraped ransomware Twitter data reflect the events consistently. Table 3 may not perfectly demonstrate the scale of impact these events had in cyberspace, however the media attention produced is mirrored in the results. Therefore it cannot be said with total certainty, that the factual impact of ransomware attacks are reflected on Twitter. However the societal reaction to these events can be retrieved from the data with a degree of accuracy.

4 Twitter Scraping for Defensive Cybersecurity

Effective cybersecurity requires intelligence directly applicable to defence mechanisms that detect and prevent intrusions. One of the primary artefacts of proactive cybersecurity are indicators of compromise (IOCs) which can be placed in automatic malware and threat detection systems, as well as aid information security officials to analyse malicious events [38]. IOCs are "one of the most common types of technical intelligence around intrusions" [16] which are shared within the cybersecurity community on various designated platforms as well as social media. IOCs can express in a wide range of forms, however the most commonly traded pieces of data are malicious file hashes, IP addresses, domains and similar information that can easily be implemented into intrusion detection systems.

Traditionally, IOCs are shared among professionals on platforms such as MISP – Threat Intelligence Sharing Platform, which is an open-source software for "collecting, storing, distributing and sharing cyber security indicators and threats" [39], where organisations are able to exchange detected information with trusted parties. There are also more publicly accessible platforms with similar information, such as VirusTotal, where it is possible to search through IOCs aggregated from "antivirus engines, website scanners, file and URL analysis tools, and user contributions." [40] However, such information can also be found on various social media platforms, most remarkably on Twitter. Besides organisations, there is also a community of independent threat hunters on Twitter sharing their discoveries. The following chapter will attempt to extract IOCs from the scraped Twitter data using regular expressions.

4.1 Indicators of Compromise Retrieval

Published IOCs often follow similar models and conventions, therefore they can be subject to pattern recognition. This thesis uses regular expressions in combination with Python to extract IOCs from the scraped Twitter data. The following possible IOCs are extracted:

- Domains, including defanged domains
- IP addresses, including defanged IP addresses

- MD5 hashes of files
- SHA256 hashes of files
- Bitcoin addresses
- Email addresses

This information can help proactively secure a network for example by adding them to firewall packet filtering, email inbox security or virus detection systems. Furthermore, these IOCs can help cybersecurity officials identify threats and conduct research about whether a specific threat has occurred before. Because the resulting tables containing extracted IOCs are of substantial size and also with relatively constant substance, the thesis presents only snippets of their contents.

4.1.1 Defanged Domains

Domains, which have malicious content such as malware download files or which are linked to command-and-control servers [16] are IOCs that can be used within a company network for web filtering. Content-control software is an essential network component, because a downloaded and executed malicious file can infect an entire network.

It is customary for malicious URLs and IP addresses to be posted online so, that they are not presented as a link and cannot be accidentally opened. Defanging is a process for preventing users from clicking on a malicious link [41], which is most commonly done by placing square brackets around a punctuation mark within the URL or obfuscating the protocol. In order to extract domains from the large Twitter data files, defanging has to be accounted for, therefore the following regular expression is used:

```
.(?:[a-zA-Z]+:\/\/)?[\w]+?(?:\[\.\]|\.)[\w]+
```

This regular expression attempts to account for defanged as well as fanged URLs. The pattern firstly optionally matches characters in the alphabet, a colon symbol and two forward slashes. After that it searches for a word character, which is followed by either a full stop or a full stop surrounded with square brackets.

Figure 2 is a sample output received after applying the previously outlined regular expression to scraped data with the “ransomware” keyword from the year 2020.

	Index	URL
2	26	:http://mrfixit[.]xyz
3	38	:http://chasiin[.]com
4	47	:94.245
5	65	:http://f0396918[.]xsph
6	85	https://iccreabc[.]com
7	88	23.253
8	119	U.S
9	139	http://mrfixit[.]xyz
10	146	U.S

Figure 2. Sample of domains extracted from Twitter scrape.

The Index column represents the row at which the domain name was positioned within the file it was extracted from. The URL column represents the extracted domain name. As seen on Figure 2, some of the rows do not contain URLs but do contain strings that match the same pattern. This suggests, that the regular expression requires adjustments to produce more precise results.

4.1.2 Defanged IPs

IP addresses are IOCs similar in function to domains and therefore also valuable pieces of data for a defensive cybersecurity system. The following regular expression was used to extract fanged and defanged IP addresses from Twitter data:

```
(?:[0-9]+:\/\/)?[\0-9]+?(?:\[\.\]|\.)[\0-9]+?(?:\[\.\]|\.)[\0-9]+?(?:\[\.\]|\.)[\0-9]+
```

This pattern attempts to match numbers from one to nine four times and separated by either a full stop or a full stop surrounded by square brackets. The sample output seen in Figure 3 is also from the “ransomware” keyword scrape from the year 2020.

	Index	IP
2	7269	"...7.2
3	24898	
4	92951	2/2. ...
5	269798	23.7.2020
6	329791	2.3.1.
7	393306	1.1.1.1+
8	496788	9.9.9.9
9	522285	9.9.9.9
10	531087	43.240.156[.]5\n

Figure 3. Sample of IP addresses extracted from Twitter scrape.

As seen from the Index column, IP addresses appear less in this set than domains. Additionally, out of the ten rows in Figure 3, only the last seems to be a valid IOC. It is clear that this pattern could be improved, so that it would match the desired results with more success. Character repetitions should be limited with stricter rules and there must be one to three numbers between each full stop. The 9.9.9.9 addresses are Quad9 IP addresses which occur frequently in this data set, as this is a Doman Name System (DNS) recursive service for security and privacy [42]. In order to achieve a better result, IP addresses such as these, should be excluded from the search.

4.1.3 Cryptographic hash functions

MD5 is most commonly used to hash malware samples so that they could be uniquely identified [43]. This type of IOC is a cryptographic checksum that acts as a fingerprint for a particular malware, making it an ideal vector for securing a network against malicious programs. The following regular expression is used to filter out MD5 hashes from the Twitter dataset:

```
[a-f0-9]{32}
```

This pattern matches any character from the lowercase letter “a” to “f “ and from numbers zero to nine exactly thirty two times.

SHA256 is another hash function with a longer digest than MD5. Like MD5, it is also used to uniquely identify malware. The regular expression pattern used to extract SHA256 hashes from the Twitter data set is the following:

```
[A-Fa-f0-9]{64}
```

This pattern matches any uppercase or lowercase letter from “a” to “f” and numbers zero to nine exactly sixty four times.

Figure 4 is an extract of the searching for the MD5 has pattern within the Twitter scrape from 2020 and with the keyword “malware”.

1	Index	MD5
2	47	50a3e9282c99a3d3656606892415fd27
3	65	50a3e9282c99a3d3656606892415fd27
4	189	2371ce0bd3e4c3da7aa1fe827df1ece6
5	202	f437080b5001266ddebafba6916db2d5
6	291	5d1d42631fb1363a3470f2d6201efb0d
7	307	d440584f4712f10e8ade8a603a943e31
8	325	81127b56d53f703de95ed97060dbfd42
9	332	641323d33b92d0d12bcb9fa78551a502
10	348	9f213490a28c7ea891e96ce79678a521

Figure 4. Sample of MD5 hashes extracted from Twitter scrape.

Figure 5 depicts and extract of SHA256 digests taken from the same sample of data as Figure 4.

1	Index	SHA256
2	6522	dda9f301fefb543235cd29166dd7bf306e2d52fa6126c887f12c1f4a2c8a3fb0
3	7469	dda9f301fefb543235cd29166dd7bf306e2d52fa6126c887f12c1f4a2c8a3fb0
4	9615	b441c390a566b60b9fcaf034269bcceb4554d81733215c3baff2b5f20a6e614e
5	9956	90acae3f682f01864e49c756bc9d46f153fcc4a7e703fd1723a8aa7ec01b378c
6	13871	266757bb15e4b7cffaa44045896c8ff4118a1d75e8ccbe4a9730bb6179f2bd32
7	13959	c8df39d2803d496902e4cec802079de739b20ad8db68bd0a03eeb0261bb6783f
8	14465	458f33a1dc34e0b587dda65f10238f590738abe8a453511fab0558144b919e37
9	14530	a6cc8bd23bbafd0b356404eb24b50236815a03abdfcf8d280dbedd5c45bf6282
10	15440	a18ad572ca6b8b53d45eef810fc116f9ea1e820528af97f2fbd970f252296fe5

Figure 5. Sample of SHA256 hashes extracted from Twitter scrape.

The output of this search is relatively frequent and accurate as can be seen from Figure 4 and 5. A couple of lookups of the SHA256 hashes in the VirusTotal database also yields matches. These checksums could be added to a company's malware identification database, which can become helpful when encountering a new malicious file.

4.1.4 Bitcoin Addresses

Bitcoin addresses can also be used as IOCs because certain ransomware campaigns can be linked to them by analysts [44]. Therefore, if an address shows up in a new attack which has been previously recorded then these incidents can be associated, which helps with future identification processes. The following is the regular expression used to search for Bitcoin addresses:

```
[13][a-km-zA-HJ-NP-Z1-9]{25,34}
```

This regular expression pattern matches character sequences starting with either a one or a three. Next it looks for combinations of certain uppercase and lowercase letters, and numbers from one to nine. The character sequences are of length twenty five to thirty four.

Figure 6 is an extract of Bitcoin addresses from the 2016 “ransomware” keyword Twitter dataset.

1	Index	Bitcoin Address
2	5856	11BE21DFA9AA67A9FBDB9AB757B4FDD
3	33731	3ccfd191dcceeae8e884f82f5c7ad
4	65025	1fb8aeb175de3828a29299df5233d4
5	91871	39fcc1c5f5a42722e8ee1554cba8
6	136828	3CC3397B57F1CFD3A14781719ACB7
7	144754	13E6DA27A2C95D3988714792DDE969
8	146210	35494aa6ce3ccef7346b548da5
9	162490	3239434398da123454635d8fdb
10	165054	16Ws5mAg87qwKBcXY3hRXhLtj9kkkDVisd

Figure 6. Sample of Bitcoin addresses extracted from Twitter scrape.

As seen from Figure 6, some of the results seem to be accurate, following the P2PKH and P2SH Bitcoin address formats [45]. However, this pattern could also retrieve some MD5 hashes, since the length of their character sequences fall into the range of a Bitcoin address.

4.1.5 Email Addresses

Email addresses that are used for phishing or spam are also a type of IOC because they may be distributing malicious files and content. Malicious email addresses can be blacklisted in a company network, so that users do not receive harmful letters. The following regular expression is used to filter out email addresses:

```
([a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+)
```

This pattern matches alphanumeric characters including some special characters up until the address sign. Next, it matches alphanumeric characters until a full stop, followed by additional domain characters.

Figure 7 demonstrates a sample of email addresses extracted from the year 2020 Twitter scrape with keyword “ransomware”.

1	Index	Email
2	392	sales@arrakisconsulting.com
3	425	sales@arrakisconsulting.com
4	524	itsecurity@pathcom.com
5	675	sophia.furber@spglobal.com
6	784	Ponce.lorena@aol.com
7	1012	symetrikk@protonmail.com
8	1069	sophia.furber@spglobal.com
9	1329	checkmail7@protonmail.com
10	1808	sqlsolutions@protonmail.com

Figure 7. Sample of email addresses extracted from Twitter scrape.

Although extracting email addresses is relatively easy, it is much more difficult to automatically deduce which of them are malicious. One way of deciding is based on the domain or service provider of the email address. None of the emails in Figure 7 appear to be suspicious, some of them seem to be addresses of cybersecurity companies and vendors. This suggest that the data scrape to find malicious emails should be performed with more specific keywords, as the word “ransomware” appears to be to general for this purpose.

4.2 Analysis of Efficiencies and Drawbacks

Extracting IOCs from Twitter data is an effective cybersecurity procedure when all anomalies in the data can be accounted for in the filtering rules. From the results of the separation of IOCs from data copied from Twitter, it is clear that some types of information are easier to draw out than others. In order for this method to work efficiently, a thorough and precise set of rules for extraction must be produced. Ideally, the values pulled out must be categorised correctly so they can be applied to the accurate components of a network’s infrastructure. Moreover, secure and trustworthy data points should not be added to IOCs as this may result in data loss. The following is an analysis of each type of IOC extraction executed in this chapter.

The malicious domain search proved successful to an extent where the desired results could be found in the data pulled. Domain names that are defanged are a clear indicator of an IOC, since this is a method commonly used by threat publishers to make the information they share safe for users. Therefore such domains could be directly added to an organisation’s web filtering mechanisms to prevent users from inadvertently infecting

a computer within a network. However there is no universal ruleset for publishing defanged URLs and therefore it is determined to be difficult to avoid filtering out necessary or extracting unnecessary data. Additionally, not all threat hunters even defang the information they publish, that is why it is also necessary to extract fanged URLs, but determining which of those are malicious requires context, not just the separated domain name.

Extracting malicious IP addresses has similar issues with the URL filtering, however due to the straightforward structure of an IP, this study finds that it is simpler to perform. Nevertheless this type of IOC separation from the Twitter dataset also comes with its complications. There are a number of IP addresses that must be whitelisted, since there are a number of DNS resolution services whose secure addresses appear frequently in Twitter content that discusses cybersecurity issues. Another concern that becomes apparent in this research, is that data such as date and time can have a similar pattern to IP addresses. Therefore the regular expression has to be more precisely constructed for this method to be effective. If a system were to implement false positives in this form into web filtering processes, it would most likely not have destructive results, however it would be a misuse of memory and storage.

Out of the IOC types extracted in this chapter, malware hash and Bitcoin address separation gave the most productive results. However, due to the similarities of MD5 hashes and Bitcoin addresses, the filtering process must be more precise than this study applied. For example the regular expression pattern could stay the same for MD5 and Bitcoin address extraction, however the tweet must also include a defining keyword such as “hash” and “Bitcoin” respectively. The SHA256 separation from the rest of the pulled Twitter data can be deemed a success, as these hashes can also be matched on VirusTotal.

Email extraction is the least productive of the IOC retrievals performed in this research. The pattern is successful at generating a list of emails from the Twitter data, however determining which of these are malicious is complicated. Firstly, many email domains must be whitelisted, however there are numerous private enterprise domains as well as email service provider domains to account for. Additionally, even if the domain is legitimate, that does not directly rule out that the address is not a spam or phishing account. In this case the email address extracted requires context and what else is said about it in the tweet it originates from, but this would require either an advanced natural

language processing program to evaluate the text surrounding the possible IOC or manual intervention.

Although there is potential for retrieving IOCs from Twitter by firstly scraping its data according to cybersecurity related keywords and secondly searching that information with pattern matching, this method comes with many flaws. The patterns used must account for a myriad of anomalies in the data and even then the separated information may have to be evaluated for its integrity and probability of being an IOC. In practice, extracting IOCs with this type of procedure may require adequate natural language processing capabilities, because some of the indicators entail context. This means that an automated interpreter should have the capability to determine whether the data is benign or actually an IOC. Furthermore, in order to implement this method on a scale that would be beneficial to the cybersecurity industry, it would require a substantial amount of resources such as good bandwidth for the copying of Twitter content, a substantial amount of memory for storing the data and strong processing capabilities for filtering and evaluating the data (according to the author's experience, many of the scraping sessions took too much time to be of practical use, as did the post-processing and analysis in Python).

5 Summary

Social media undergoes a continuous growth each year and is a contributing factor to the equally sizable increase of cyberspace. These expanding grounds in cyberspace open doors for various adversary activities at an unprecedented scale, which cybersecurity officials must confront in an efficient and productive manner. In order to tackle these incessant cyber threats, a number of mitigation techniques must be employed. One of these techniques is gathering CTI from publicly available sources. Social media is essentially one of the largest public knowledge bases, which aggregates a wide range of information, including data about cybersecurity events, threats, actors and alleviation procedures. Therefore in order to better cybersecurity capabilities, there must exist solutions for filtering and extracting such intelligence from social media.

This thesis proposes a method for using social media to benefit the cybersecurity field by extracting necessary information employed in threat mitigation efforts. The solution suggested by the research firstly exercises the use of scraping tools on the social media platform Twitter and using cybersecurity related keywords for retrieving relevant data. Secondly, the copied data is searched, filtered and the desired data is quantified in order to evaluate Twitter's overall potential for tracking cybersecurity events. Finally the scraped data is searched with the goal of extracting information applicable to proactive cyber defence, such as IOCs and vulnerabilities.

The solution put forward by this thesis provides cybersecurity agents with a possible method for making use of Twitter in their cyber defence procedures. The results demonstrate that Twitter's data stream is successful in tracking major cybersecurity events. Moreover, it is possible to extract information, such as IOCs, directly applicable in preventative methods of cyber defence. However, the large amounts of data processed and stored in memory exhibit a downfall to this method and as a real-world implementation this could be deemed impractical. Finally the hypothesis posed in Chapter 1 stating that the cybersecurity incident data scraped from Twitter is proportionate to recorded cybersecurity events, proved incorrect. The results from Chapter 3 were mostly inconclusive.

In terms of future developments, Twitter scraping could be automated and the data forwarded to a database where it could further undergo automatic processing with the

extraction of CTI as the goal. This type of application would greatly benefit cybersecurity researchers who are interested in analysing time series and historic data, but also are interested in current cyber threat identification, prevention and mitigation. A more imminent case for future research, would be to improve the regular expressions used for pattern matching IOCs or to discover a superior way of extracting desired data.

In the case that statistics of historic data is not of interest to a cybersecurity official, this solution could be changed to directly streaming Twitter data through filters and data processing algorithms. The benefit of this method would be significantly lower storage requirements and a more rapid result output.

References

[1] S. Morgan, “2021 Report: Cyberwarfare in the C-Suite”, Cybersecurity Ventures, 2021.

[2] Deloitte Risk & Financial Advisory, “Deloitte Puts the Spotlight on the Cost of Cyber-Crime Operations in New Threat Study,” Deloitte United States, 2021. [Online]. Available: https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-announces-new-cyber-threat-study-on-criminal-operational-cost.html. [Accessed: 17 Apr. 2021].

[3] D. Swinhoe, “How much does it cost to launch a cyberattack?,” CSO Online, 2020. [Online]. Available: https://www.csoonline.com/article/3340049/how-much-does-it-cost-to-launch-a-cyberattack.html. [Accessed: 17 Apr. 2021].

[4] L. Kello, The virtual weapon and international order. New Haven: Yale University Press, 2018.

[5] K. Benedict, “Information Operation: The Fifth Dimension of Warfare”, SAP Insider, 2012.

[6] L. D. Welch, “Cyberspace – The Fifth Operational Domain”, IDA, 2011.

[7] M. Erbschloe, Social Media Warfare: equal weapons for all. S.l.: CRC PRESS, 2020.

[8] T. E. Nissen, #TheWeaponizationOfSocialMedia: @Characteristics_of_ Contemporary_Conflicts, Royal Danish Defence College, 2015.

[9] K. Baker, “What is Cyber Threat Intelligence? [Beginner's Guide],” *Crowdstrike,* 18-Feb-2021. [Online]. Available: https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/. [Accessed: 16-May-2021].

[10] “Cyber Threat Intelligence 101,” *FireEye*. [Online]. Available: https://www.fireeye.com/mandiant/threat-intelligence/what-is-cyber-threat-intelligence.html. [Accessed: 16-May-2021].

[11] Merriam-Webster, “Social media.” Merriam-Webster.com Dictionary. [Online]. Available: https://www.merriam-webster.com/dictionary/social%20media. [Accessed 10 Apr. 2021].

[12] Hootsuite and We Are Social, “Digital 2021: Global Overview Report”, 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-global-overview-report/. [Accessed 10 April 2021].

[13] Cisco Systems, “Cisco Global 2022 Forecast Highlights”, 2018. [Online]. Available: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2022_Forecast_Highlights.pdf. [Accessed 10 Apr. 2021].

[14] B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” AI & SOCIETY, vol. 30, no. 1, pp. 89–116, Jul. 2014.

[15] V. Krotov and L. Silva , “Legality and Ethics of Web Scraping”, 2018.

[16] S. J. Roberts and R. Brown, Intelligence-driven incident response: outwitting the adversary. O'Reilly, 2017.

[17] J. Pastor-Galindo, P. Nespoli, F. Gomez Marmol, and G. Martinez Perez, “The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends,” IEEE Access, vol. 8, pp. 10282–10304, 2020.

[18] Y. Ju, Q. Li, H. Y. Liu, X. M. Cui, and Z. H. Wang, “Study on application of open source intelligence from social media in the military,” Journal of Physics: Conference Series, vol. 1507, p. 052017, 2020.

[19] J. M. Davitch, “Open Sources for the Information Age: Or How I Learned to Stop Worrying and Love Unclassified Data,” National Defence University Press, 2017.

[20] INTRUSION Inc., “Cybercrime To Cost The World $10.5 Trillion Annually By 2025,” GlobeNewswire News Room, 2020. [Online]. Available: https://www.globenewswire.com/news-release/2020/11/18/2129432/0/en/Cybercrime-To-Cost-The-World-10-5-Trillion-Annually-By-2025.html. [Accessed: 11 Apr. 2021].

[21] “What is Threat Intelligence?,” Forcepoint, 2020. [Online]. Available: https://www.forcepoint.com/cyber-edu/threat-intelligence. [Accessed: 11 Apr. 2021].

[22] David Sayce, “The Number of tweets per day in 2020,” David Sayce, 2020. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/. [Accessed: 20 Apr. 2021].

[23] “How to use advanced search – find Tweets, hashtags, and more,” Twitter. [Online]. Available: https://help.twitter.com/en/using-twitter/twitter-advanced-search. [Accessed: 20 Apr. 2021].

[24] JustAnotherArchivist, “JustAnotherArchivist/snscrape,” GitHub. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape. [Accessed: 28 Apr. 2021].

[25] Cisco, “Cyber Attack - What Are Common Cyberthreats?,” Cisco, 19-Feb-2021. [Online]. Available: https://www.cisco.com/c/en/us/products/security/common-cyberattacks.html. [Accessed: 29 Apr. 2021].

[26] P. Passeri, “2016 Cyber Attacks Statistics,” HACKMAGEDDON, 2017. [Online]. Available: https://www.hackmageddon.com/2017/01/19/2016-cyber-attacks-statistics/. [Accessed: 29 Apr. 2021].

[27] OTA, “2017 Cyber Incidents & Breach Trends Report”, Internet Society, 2018.

[28] SonicWall, “2021 SonicWall Cyber Threat Report”, SonicWall, 2021.

[29] Deloitte, “Advanced Persistent Threat,” Deloitte Switzerland, 07-Aug-2020. [Online]. Available: https://www2.deloitte.com/ch/en/pages/risk/articles/advanced-persistent-threat.html. [Accessed: 29 Apr. 2021].

[30] E. Seker, “Top Famous and Active APT Groups who can Turn Life to A Nightmare,” Medium, 2020. [Online]. Available: https://medium.datadriveninvestor.com/top-famous-and-active-apt-groups-who-can-turn-life-to-a-nightmare-5d130168f43. [Accessed: 29 Apr. 2021].

[31] “North Korean Regime-Backed Programmer Charged With Conspiracy to Conduct Multiple Cyber Attacks and Intrusions,” The United States Department of Justice , 06-Sep-2018. [Online]. Available: https://www.justice.gov/opa/pr/north-korean-regime-backed-programmer-charged-conspiracy-conduct-multiple-cyber-attacks-and. [Accessed: 29 Apr. 2021].

[32] “Supply Chain Attack on SolarWinds Orion Platform Affecting Multiple Organizations Worldswide (APT29),” FortiGuard, 2020. [Online]. Available: https://www.fortiguard.com/threat-signal-report/3770/supply-chain-attack-on-solarwinds-orion-platform-affecting-multiple-organizations-worldwide-apt29. [Accessed: 29 Apr. 2021].

[33] APT28, 2017. [Online]. Available: https://attack.mitre.org/groups/G0007/. [Accessed: 29 Apr. 2021].

[34] Kaspersky, “Ransomware Attacks and Types – How Encryption Trojans Differ,”.[Online]. Available: https://www.kaspersky.com/resource-center/threats/ransomware-attacks-and-types. [Accessed: 29 Apr. 2021].

[35] CrowdStrike, “12 Notorious Ransomware Examples,” 2021. [Online]. Available: https://www.crowdstrike.com/cybersecurity-101/ransomware/ransomware-examples/. [Accessed: 29 Apr. 2021].

[36] “Locky Ransomware,” ENISA, 2016. [Online]. Available: https://www.enisa.europa.eu/publications/info-notes/locky-ransomware. [Accessed: 29 Apr. 2021].

[37] A. Hanel, “What is Ryuk Ransomware? The Complete Breakdown,” CrowdStrike, 2019. [Online]. Available: https://www.crowdstrike.com/blog/big-game-hunting-with-ryuk-another-lucrative-targeted-ransomware/. [Accessed: 29 Apr. 2021].

[38] Trend Micro, “Indicators of compromise,” Definition. [Online]. Available: https://www.trendmicro.com/vinfo/us/security/definition/indicators-of-compromise. [Accessed: 29 Apr. 2021].

[39] Misp, “MISP/MISP,” GitHub. [Online]. Available: https://github.com/MISP/MISP. [Accessed: 29 Apr. 2021].

[40] VirusTotal, “How it works,” VirusTotal. [Online]. Available: https://support.virustotal.com/hc/en-us/articles/115002126889-How-it-works. [Accessed: 29 Apr. 2021].

[41] IBM, “Email Security – Defanging URLs,” IBM. [Online]. Available: https://www.ibm.com/docs/en/rsoa-and-rp/32.0?topic=SSBRUQ_32.0.0%2Fcom.ibm.resilient.doc%2Finstall%2Fresilient_install_defangURLs.htm. [Accessed: 29 Apr. 2021].

[42] Quad9, “A public and free DNS service for a better security and privacy,” Quad9. [Online]. Available: https://www.quad9.net/. [Accessed: 29 Apr. 2021].

[43] M. Sikorski and A. Honig, “Practical Malware Analysis,” O'Reilly Online Learning. [Online]. Available: https://www.oreilly.com/library/view/practical-malware-analysis/9781593272906/ch02s02.html. [Accessed: 29 Apr. 2021].

[44] N. Kseib, “CryptoLocker Deep-Dive: Why We Use Bitcoin Addresses as an IOC,” TruSTAR, 09-Aug-2018. [Online]. Available: https://www.trustar.co/blog/why-we-use-bitcoin-addresses-as-an-ioc. [Accessed: 29 Apr. 2021].

[45] All Private Keys, “Bitcoin address formats and prefixes,”. [Online]. Available: https://allprivatekeys.com/bitcoin-address-format. [Accessed: 29 Apr. 2021].

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Aili Juhasoo-Lawrence

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis Social Media Scraping for Cyber Security: Performing Open-Source Intelligence with Twitter, supervised by Kaido Kikkas
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

17.05.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – Alternative names for APT groups

Advanced Persistent Threat Group	Alternative Names
APT1	PLA Unit 61398, Comment Crew, Comment Group, Comment Panda
APT41	Double Dragon
APT35	Charming Kitten, Newscaster
APT33	Elfin, HOLMIUM
APT38	Lazarus Group, HIDDEN COBRA, Guardians of Peace, ZINC, NICKEL ACADEMY
APT37	ScarCruft, Reaper, Group123, TEMP.Reaper, Ricochet Chollima
APT28	Fancy Bear, SNAKEMACKEREL, Swallowtail, Group 74, Sednit, Sofacy, Pawn Storm, STRONTIUM, Tsar Team, Threat Group-4127, TG-4127
APT29	YTTRIUM, The Dukes, Cozy Bear, CozyDuke
Equation Group	Tilded Team, Labert, EQGRP, Longhorn, PLATINUM TERMINAL