Analyzing Modern Trends in Malware

Daniel Sun, Joseph Torres, Katherine Wang, Jessica Zhu

***Abstract*—Malware has historically been, and continues to be, a pressing concern for computational systems. As various technologies are created and modified, individuals and corporations alike must find new ways of securing their data from malicious third parties. In this paper, we perform an analysis of modern trends in malware to determine the current state and future directions of malware development and targets. We have scraped over 2 million data points from 21 data repositories and drew patterns from various characteristics of malware, including but not limited to: its origin country, IP, SHA-256 hashes, and MD5 hashes. We compare our findings to current predictions made by antivirus companies to examine the accuracy of their results. Our results support some predictions but are inconclusive for others.**

I. Introduction

For as long as computational systems have been used, malicious third parties have explored methods of compromising and abusing these systems to accomplish personal goals. These malevolent actors have various methods of accomplishing such objectives, though one of the most prevalent is through malware. Malware is any type of "malicious software" that is intentionally designed to cause damage to computer systems. The mechanisms and intentions behind malware construction and distribution vary greatly (1). While some types of malware aim to cause as much damage as possible, others have ulterior motives, including Keyloggers which gain a user's credentials (2), rogueware which scams the user (3), or ransomware which holds a user's data hostage (4).

The nature of malware has shifted significantly over time. Early malware was often designed with damage more in mind than financial gain. As time has passed, attack vectors have become increasingly cunning in order to manipulate victims. Recently, malware developers have been particularly interested in creating programs which either fool or scare a user into giving sensitive information and money. Furthermore, the number of studies out there on malware are plentiful which makes it an excellent candidate for meta-analysis. Previous meta-analyses are growing stale, and can contribute to a more comprehensive modern one (5).

By performing analysis on malware, exploits that may be taken advantage of can be reverse engineered or examined. This can result in patches which render the malware ineffective, as well as insight into other vulnerabilities and exploits that can be fixed before potential zero-day attacks (6). Thus, in order to discover current and potential future trends, we have collected a holistic and comprehensive set of data composed of recent malware and modern malware. This information includes useful metadata as well as detailed information about the malware. In turn, we can make well-supported predictions about the direction of malware and grant well-advised precautions. In addition, we hope to perform our analysis in a way which makes it easily repeatable and extensible for subsequent usage.

II. Related Work

Companies such as Malwarebytes Labs, Symantec, and McAfee produce yearly reports on the state of malware based on malware their technologies have encountered over the span of the year. However, the number of malware is actually higher than ever (7) despite increasingly complex security systems, more defense in depth, and a greater consideration of security. Thus, these reports are not fully comprehensive, nor do they have the descriptive features which we are interested in. Though we will be conducting our analysis independent of these reports, it is worth noting some interesting findings from them.

A major trend in the past year was an increase in attacks on corporations, both to target the corporations themselves and the users that they served, likely due to increased profitability (8). These were commonly in the form of Backdoors, Miners, Spyware, and Information Stealers (9). As users become more conscious about attacks and scams, malware

has tended to become less and less detectable (8). One such example is Formjacking. When a user submits sensitive information to a Formjacked commercial website, their information is skimmed in real-time without a disruption to the intended transaction (10). A number of high-profile Formjacking attacks on e-commerce sites were carried out over the last year (9). In addition, taking advantage of known security vulnerabilities, the amount of Trojans, a form of disguised malware, increased as well (9). However, profitable methods such as Ransomware and Cryptojacking, especially as coin prices drop, are on a seeming decline, though still very much present.

In terms of attack medium, malware delivered through spam has re-emerged as the most popular option; before 2018, attacks via exploits were the most common. Engineering practices have been constantly improving and exploits are becoming harder to find, but social engineering is always executable (9).

These companies, based on the trends they saw in 2018, also made predictions for 2019. Some of the predictions are out of the scope, like the rise of AI malware or criminal partnerships, but for some predictions, we were able to compare the results from our trends analysis to them. McAfee, for example, expected an increase in Banking Trojans, along with an increase in malware on mobile and IoT platforms (11). A continued decline in Ransomware instances is also expected for 2019 (9) (8). In general, security companies predict that the amount of supposedly less detectable malware, such as Botnets, Rootkits, Skimmers, and malware that uses "fileless" components, will continue to increase (9), (12).

III. Technical Approach

github.com/Finaris/6857-malware-analysis

A. Setup

Malware analysis can be a dangerous form of security research. Therefore, two principles were important for the execution of this project:

1) If we ran into suspected malware on what appears to be a legitimate platform, it was discarded from the sample, so no potential legal issues would arise from experimenting with malicious software.
2) The nature of malware is ambiguous, especially for new malware which has not yet been analyzed. Although severe zero-day vulnerabilities are rare, the assumption was made that each program had such capabilities. Therefore, code was only analyzed on an isolated machine through a secure, isolated virtual machine (VM).

For the purposes of this project, we gathered information about malicious files and domains that could encompass all devices. For static and dynamic analysis, we focused primarily on Windows machines and Windows Portable Executables (PEs). Because of this, the VM we set up had the following configuration:

- Windows 10 Enterprise x64
- 2 Processor Cores
- 2048 MB Memory
- 60 GB Persistent Memory

In addition, a new VM was used for each program tested to prevent confounding interactions during experimental dynamic analysis (which we accomplished by simply reloading a snapshot we took before doing anything). For information gathering and static analysis, the same VM state was used.

B. Web Scraping

Malware was collected by using databases and sources for modern malware and looking at old records. To compensate for malware which has been taken down, findings from previous researchers were aggregated, and these findings were collected in our meta-analysis. We also scraped hundreds of thousands of blacklisted urls as well. To scrape the URLs and the databases, we made 21 scrapers, inherited from an extensible base class scraper that we hope can be easily used by others that want to conduct malware analysis in the future. For the list of sources and specific implementations of these scrapers, please visit our Github page (https://github.com/Finaris/6857-malware-analysis) and look at the README in the web_scraper directory.

In order to conduct the analysis, malware was backlogged and the following characteristics were noted:

- **Category** (e.g. General malware, Ransomware, Botnet. etc.)
- **Origin Country**
- **CC** (country code, should correspond with origin country)
- **IP**
- **SHA256 and MD5** hashes (for malware database lookups)
- **Registrar** (where the domain the registered)
- **Date** (Recency)
- **Targeted Platform** (e.g. operating system, browser, application, etc.)
- **Source Vector** (e.g. executable, script, etc.)
- **Status** (e.g. online, offline, or unknown)
- **URL**

To help us fill in the fields, we used SHA256 and/or MD5 hashes to search for more information from VirusTotal, a large malware database. The VirusTotal API provides relevant meta-information about the malware by querying a multitude of different databases for information. However since we used VirusTotal's public API to retrieve information, we were constrained to 4 API calls an hour and thus limited in the amount of information we could garner. Using the API, we created an additional dataset of 4540 entries, sampled randomly from the initial database in order to prevent issues where clumps of domains with the same IP/malware were listed next to each other.

C. Preliminary Static and Dynamic Analysis

Though not a major face of our analysis, preliminary static and dynamic analysis was performed. We begin by discussing static analysis. For static analysis, we collected a list of URLs from our data set that pointed to PEs. On our Windows analysis VM, we downloaded all of these files and used the following tools:

- PEview (for viewing information about PE files)
- CFF Explorer (a PE editor)
- ILSpy (.NET decompiler)
- pefile (Python library for performing simple static analysis)

We did a very small analysis on around 200 PEs and gathered basic preliminary data (file size, type, hashes, source URL, basic categorization). Though we find no novel insights from these files, as they support our existing claims which we discuss momentarily, we hope further analysis will prove useful.

We have yet to collect useful data from dynamic analysis, but we have a basic setup we've experimented with that use the following tools:

- Wireshark (for analyzing network traffic made from mawlare)
- TCPView (observing TCP connections made from a program)
- Process Explorer (an in-depth system monitor)
- Process Monitor (an additional system monitor which has different features from Process Explorer)

D. Data Processing and Analysis

We used the Python's Pandas and Matplotlib libraries to process, analyze, and graph over 2 million compromised domains. After collecting the initial database of information, we began normalizing the data in order to fill in empty fields and make the analysis simpler. For example, we standardized the categories, normalized the country code and origin country, and created year and year-month columns for easier aggregation. We conducted an exploratory analysis of the data, generating statics and graphs about the distribution of the data and trends over time.

We extended the data fields in our analysis to include the geolocation information, which includes the city, country, continent, and autonomous system organization (ASO) of each IP address in our dataset. Some of the data from the original scraper already included country and ASO information, and so we merged any additional information that the geolocation would provide. In order to geolocate IPs, we used the Geoip2 module, a Python library that allows us to locally query our large database of IPs and receive locational data. The module utilizes the GeoLite2 City, Country, and ASN databases from MaxMind, which we downloaded locally to query our hundreds of thousands of IPs.

For the smaller VirusTotal dataset, since some of the sites we scraped our data from might overlap domains, we wanted to eliminate duplicates. In order to remove duplicates, we first queried the VirusTotal API with each individual malware's MD5 or SHA256 hash, whichever was available

from the original scrape. The response contained a SHA256 hash, regardless of whether the API was queried with MD5 or SHA256, which allowed us to drop data duplicates.

In order to determine the composition of our dataset, we first had to standardize the responses returned by various databases. Each had their own format for malware representation and some had differing results as well. We defined a set of keywords that we would want to investigate (eg. Trojans, Viruses, Rootkits, Spyware, etc.). Parsing the result responses along with some manual work allowed us to find representations of each of our keywords, which we used to determine a phrase of keywords for each malware entry. We then combined the keywords into one result column, by choosing keywords that were present in three of more individual result responses. We also performed a similar process for date and year, except in this case we choose the most frequent entry to be the single representative value. Consolidating the data in this way simplified the analysis process. For analysis, we generated statistics and trend graphs for the various keywords.

IV. Findings

A. Exploratory Analysis

1) Overall Dataset: Our dataset contained 2,892,956 entries. Within these entries, we had over one million MD5 and SHA256 hashes and over one million unique URLs. Combining the geolocated IP information and the originally scraped information, we were able to determine the source country for 600,000 entries in our dataset. We gathered data about the most common countries that the malware originated from. The top 20 are shown in Figure 1. We were also able to determine the breakdown of the data among overarching categories and different filetypes, shown in Figure 2 and Figure 3. To conduct on analysis on trends, we generated graphs of various category trends over time; an example with malware as the category is shown in Figure 4a respectively. About 10% of the malware in our dataset is still online.

2) Hash Dataset: From our smaller dataset taken from VirusTotal, we found that 111 out of 4540 data points were duplicates, or about 2.44%. VirusTotal also gave us information about whether or not multiple virus scanners detected the domain as malicious.

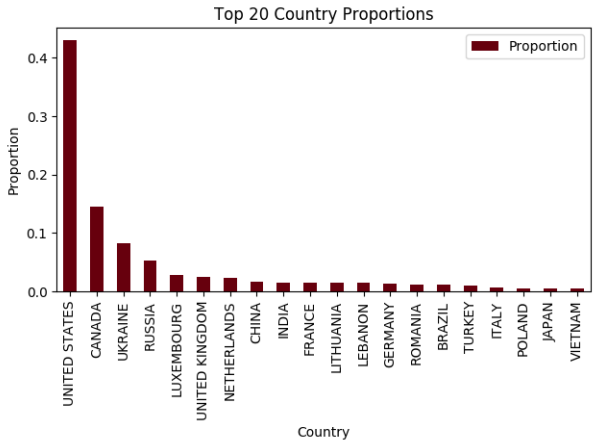

Fig. 1: We see the vast majority of our data originates in the US, possibly due to selection bias of our original sources.

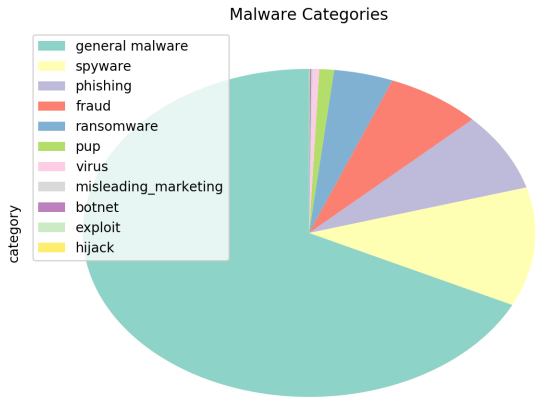

Fig. 2: Categories of our malware data. Note that general malware is such a large section because a lot of data was not specifically categorized, and there is a wide array of possible malware types.

We found that 326 out of 4540 domains, 7.36%, were not found to be malicious by any virus scanner. Our analysis revealed that 99.4% of our data was detected to be some type of Trojan by at least one malware scanner, 17.4% were Ransomware, and only 1.5% were worms. Of course, it is important to note that our sample size is still quite small compared to our full dataset, and our findings are biased based on our original sources. Additionally, Trojan attacks occupy such a large percentage because it is often used as an umbrella term for many different types of malware.

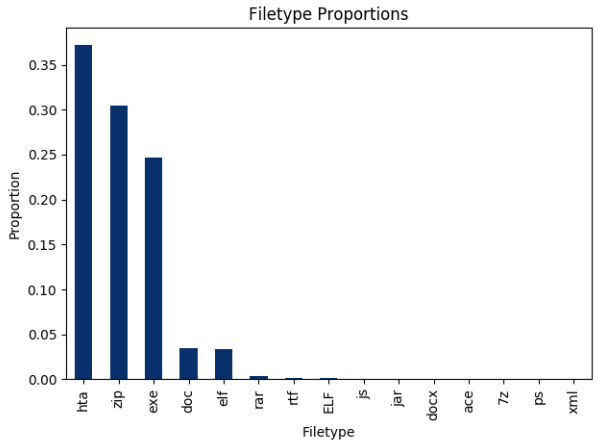

Fig. 3: Proportions of filetypes within our malware data.

We also used this analysis technique to track these keyword trends over time. We graphed the proportions of different observed malware types over the past 5-6 years. We found that some keywords, like Trojan or Injector, consistently showed up with high percentages in all of our data. Proportions for terms like Spy, Worm, and Ransom rose and fell throughout the years. Having keyword phrases also helped us establish relations between the keywords (especially between specific keywords and more general keywords) along with providing valuable information about each malware entry. One note on this dataset however, is that it may not be completely representative of our overall dataset, since only entries that had hashes could be selected.

Compared to predictions made by various companies about the direction of malware in 2019, our findings support some predictions while they vary with others. For example, we found that Ransomware seemed to be on a decline since the second half of 2018, a correct prediction by MalwareBytes Labs. Additionally, companies expected an increase in Banking Trojans and Rootkits, which our analysis also supported.

V. Conclusion & Future Work

We scraped 21 sites for over 2 million data points on potential malware domains. We performed analysis on our data that gave us insights into the accuracy of industry predictions for 2019. We intend

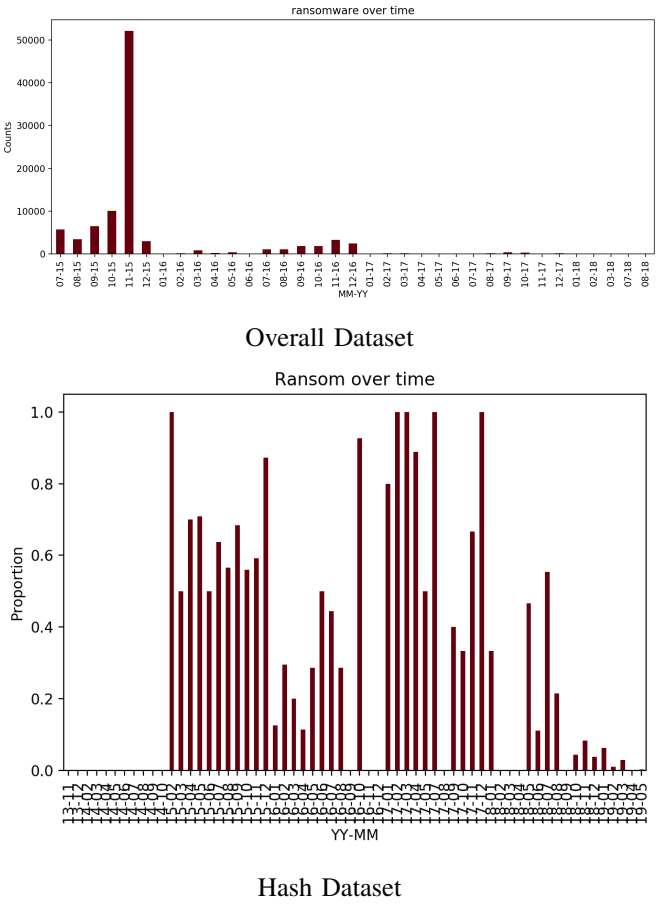

Fig. 4: We can see Ransomware rates dropping off over the last year.

on expanding our static and dynamic analysis data sets.

For future work, a deeper analysis on the specific types of malware pertaining to the file content and relevant machinery can be done. The VirusTotal public API gives limited information in the response, which allows us to detect duplicates and get virus type. However, if one searches manually or pays for the private API, then information about file type, size, target machinery, program resources and imports is also available. By either paying for the API or finding other databases online that contain such information, we could discover even more trends that are more critical for the user, such as types of machines targeted (Windows, OSX versions, Intel processors), as well as what processes are created or injected. In addition, more available databases would allow us to find even more samples of IPs for detecting duplicates and malware types.

REFERENCES

[1] N. DuPaul, "Common malware types: Cybersecurity 101," 2012. [Online]. Available: https://www.veracode.com/blog/2012/10/common-malware-types-cybersecurity-101

[2] McAfee, "What is a keylogger?" 2013. [Online]. Available: https://securingtomorrow.mcafee.com/consumer/family-safety/what-is-a-keylogger/

[3] C. Pickard and S. Miladinov, "Rogue software: Protection against potentially unwanted applications," in *2012 7th International Conference on Malicious and Unwanted Software*, Oct 2012, pp. 1–8.

[4] CISA, "Ransomware." [Online]. Available: https://www.us-cert.gov/Ransomware

[5] M. Gehem, A. Usanov, E. Frinking, and M. Rademaker, "Assessing cyber security: A meta-analysis of threats, trends, and responses to cyber attacks," Hague Centre for Strategic Studies, Tech. Rep., 2015. [Online]. Available: http://www.jstor.org/stable/resrep12567

[6] D. Last, "Consensus forecasting of zero-day vulnerabilities for network security," in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, Oct 2016, pp. 1–8.

[7] AV-TEST, "Malware." [Online]. Available: https://www.av-test.org/en/statistics/malware/

[8] "2019 internet security threat report," Tech. Rep.

[9] MalwareBytes, "2019 State of Malware," MalwareBytes Labs, Tech. Rep., 2019.

[10] Symantec, "Formjacking: Major increase in attacks on online retailers," 2018. [Online]. Available: https://www.symantec.com/blogs/threat-intelligence/formjacking-attacks-retailers

[11] McAfee, "McAfee Labs 2019 Threats Predictions Report," Tech. Rep., 2019.

[12] TrendMicro, "Mapping the Future: Dealing With Pervasive and Persistent Threats," Tech. Rep., 2018.