2016 European Intelligence and Security Informatics Conference

A Post-Mortem Empirical Investigation of the Popularity and Distribution of Malware Files in the Contemporary Web-Facing Internet

Jukka Ruohonen
University of Turku, Finland
juanruo@utu.fi

Sanja Šćepanović
Aalto University, Finland
sanja.scepanovic@aalto.fi

Sami Hyrynsalmi
University of Turku, Finland
sthyry@utu.fi

Igor Mishkovski
Aalto University, Finland & University Ss. Cyril and Methodius – Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

Tuomas Aura
Aalto University, Finland
tuomas.aura@aalto.fi

Ville Leppänen
University of Turku, Finland
ville.leppanen@utu.fi

***Abstract*—This short empirical paper investigates a snapshot of about two million files from a continuously updated big data collection maintained by F-Secure for security intelligence purposes. By further augmenting the snapshot with open data covering about a half of a million files, the paper examines two questions: (a) what is the shape of a probability distribution characterizing the relative share of malware files to all files distributed from web-facing Internet domains; and (b) what is the distribution shaping the popularity of malware files? A bimodal distribution is proposed as an answer to the former question, while a graph theoretical definition for the popularity concept indicates a long-tailed, extreme value distribution. With these two questions – and the answers thereto, the paper contributes to the attempts to understand large-scale characteristics of malware at the grand population level – at the level of the whole Internet.**

***Index Terms*—malware, web crawling, security intelligence**

I. Introduction

This short paper operates with two theoretical concepts: distribution and popularity, both of which are observed through downloadable files. A distribution rate is defined as the share of malware files to all files made available for download from a domain. Analogously, the popularity of a malware file is defined as the number of domains that have distributed the same unique file. These two concepts are utilized for investigating a dataset that covers over two million files from which well over hundred thousand are suspected to be malware. The empirical analysis operates under a so-called *post-mortem* setting, which is a common approach in network forensics research and practice [1]. In other words, it should be emphasized that the dataset provides only a snapshot for analyzing a historical period from late 2015 to early 2016. In general, the intention is to hypothesize about the population level characteristics behind the two subsequently operationalized concepts. For this purpose, the investigated snapshot is credible and even ideal.

A. Distribution

The primary data source is based on a web crawling framework. To quickly outline the underlying crawling framework from a software engineering perspective, consider three basic modules: a *crawler* $\mathcal{C}$, a *milker* $\mathcal{M}$, and a pool of *detectors* for malicious software, $\mathcal{D}_1, \ldots, \mathcal{D}_d$. Ideally, a single crawling snapshot would output a sample that is generalizable to the whole web-facing Internet – and theoretically even the Internet beyond WWW. The crawler thus crawls the Internet, using an initial seed of web sites for moving onward in a continuously updated graph comprised of WWW hyperlinks. For each hyperlinked host, whether a domain or an IP address, $\mathcal{C}$ passes the links for $\mathcal{M}$, which “milks” [2] the links for downloadable files, irrespective whether a subsequent download occurs via the hyper text transfer protocol, the file transfer protocol, or any other supported, conventional protocol for transferring files. Having milked all files for a given host, $\mathcal{M}$ then passes the downloaded files for $\mathcal{D}_1, \ldots, \mathcal{D}_d$, which classify the files as “clean” or “malicious” according to best of their abilities.

While not necessitated by the general framework, in this paper a further module, say *aggregator* $\mathcal{A}$, preprocess the raw sample by first (a) excluding hosts represented as IP addresses in the embedded hyperlinks, and then (b) aggregating the remaining fully qualified domain names to the second-highest level. That is, to use a so-called λ-notation [3], this aggregation is done by separating the 2-LD and the top-level domain (1-LD) from the associated uniform resource locators, and then summing the output from $\mathcal{D}_1, \ldots, \mathcal{D}_d$ accordingly.[1]

Then, for the i:th file, f_i, in a non-empty set of files, $f_i \in F_j$, downloaded from the j:th crawled, milked, and aggregated 2-LD/1-LD, the k:th detector outputs an integer:

$$\delta^k_{f_i \mid F_j} = \mathcal{D}_k(f_i \mid F_j) = \begin{cases} 0 & \text{if the file is clean, and} \\ 1 & \text{if the file is malicious.} \end{cases} \tag{1}$$

[1] Henceforth, the term *domain* is also a synonym for a 2-LD/1-LD.

978-1-5090-2857-3/16 $31.00 © 2016 IEEE
DOI 10.1109/EISIC.2016.30

IEEE computer society

The overall *detection rate* can be thus defined as:

$$\overline{\delta}_{f_i \,\mid\, F_j} = s\left(\frac{1}{d}\sum_{k=1}^{d}\delta^{k}_{f_i \,\mid\, F_j}\right), \quad s(x) = \begin{cases} 0 & \text{if } x \leq \tau \\ 1 & \text{otherwise} \end{cases}, \qquad (2)$$

where $s(x)$ is used to dichotomize the relative rate according to a predefined threshold scalar $0 \leq \tau < 1$. The simplest case would be $\tau = 0$ and, hence, $s(x)$ would be a simple indicator function, outputting zero only in case all of the d detectors agreed upon the "cleanness" of a f_i associated with the set of downloaded files F_j from a given aggregated domain.

Analogously, the relative detection rate in (2) can be extended to the files distributed from the j:th domain:

$$r = \overline{\overline{\delta}}_{F_j} = \frac{100}{|F_j|}\sum_{i=1}^{|F_j|}\overline{\delta}_{f_i \,\mid\, F_j}. \qquad (3)$$

In other words, a domain's *distribution rate* of malware files is simplified to a percentage computed from

$$r = \textit{\# of malicious files} \Big/ \textit{\# of all files} \times 100, \qquad (4)$$

which is a simple but sufficient metric to proxy the hypothesized population level distributional characteristics. Even though the aggregation with $\mathcal{A}$ leads to the presence of encompassing domains, such as `co.uk`, malware is still only a tiny drop in the ocean of files distributed in the web-facing Internet. Therefore, it should be expected that a vector-valued $\mathbf{r} = [r_1, r_2, \ldots]'$ from a mass-scale crawling endeavor would show a large subsample of zero-valued distribution rates.

B. Popularity

A graph theoretical approach is well-suited for quantifying the popularity of a unique malware file. Motivated by recent work [4], the MD5 hashes associated with each file are used to construct an undirected bipartite graph as follows. First, for each milked file, a MD5 hash is added to the graph as a vertex with a binary-valued label from (2). Second, each 2-LD/1-LD is added with a label that identifies the added vertex as a domain. Third, edges are placed between domains and the distributed files (hashes) from these domains, such that two domains are "connected" via an additional file-labeled vertex in case these two domains have shared at least one unique file.

This labeled graph representation allows to define the *popularity* of a particular file: it is the degree of a file-labeled vertex; the number of domains that have distributed the same unique file. The label set of the MD5 vertices allows to further identify whether a given file is clean or malware. While differing slightly from existing operationalizations [5], this degree-based definition is intuitive and easy to use in practice.[2]

II. Experimental Results

The experimental empirical results are presented in two brief steps: after having outlined the dataset, the distribution and popularity metrics are evaluated with descriptive statistics.

A. Data

The dataset is assembled from two continuously updating data sources. The primary part – amounting to about 85 % of the files – comes from a module in the security intelligence infrastructure of a well-known security company, F-Secure. This proprietary data source is augmented with an open data feed from a project known as "CleanMX" [6], which has provided a valuable data asset also previously [7]. No attempts are made to separate different taxonomic types of malware files; every kind of "maliciousness" is covered from the "drive-by-download-style" (cf. [2], [5]) Flash exploits to traditional computer viruses embedded in Windows executables.[3]

B. Results

The results can be summarized by starting from the graph theoretical popularity metric. A basic breakdown can be conveniently represented by subsetting the degree distribution according to the labels that the vertices carry. This breakdown is shown in Fig. 1. The upper-left plot (a) refers to the whole graph, ignoring the subsequently illustrated label subsets.

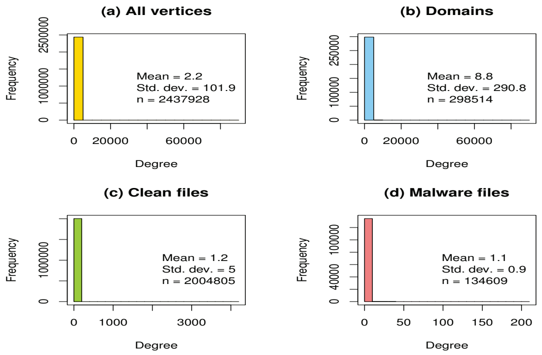

Fig. 1. Vertex Degree Histograms

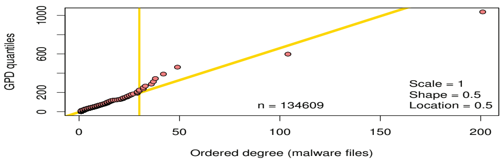

Fig. 2. The Popularity of Malware Files (GPD quantile-quantile)

[2] It should be emphasized that the graph construction is only a convenient abstraction; the interest is to observe the probability distribution characterizing the popularity of files – and not the bipartite graph construction as such.

[3] It is important to further point out a third "intermediate" source of empirical data, VirusTotal [8]. In other words, all of the 2,139,414 files milked by $\mathcal{M}$ were further passed to VirusTotal such that, for each file, the abstract detectors $\mathcal{D}_1, \ldots, \mathcal{D}_d$ refer to the tens of proprietary and open source detection engines that VirusTotal uses to scan a particular file type. The threshold scalar in (2) was set to the highest possible alert level. While this choice, $\tau = 0$, presumably increases the number of false positives and negatives, there exists a statistical trade-off: as $\tau \to 1$, the applied ground for statistical analysis reduces due to the diminishing amount of files detected as being malicious. Moreover, regarding practical applications, the value $\tau = 0$ is not necessarily a bad choice for logging entries in proxy servers, say. Finally, it is worth noting that only 8,627 (or about 3 %) hosts were represented as IPv4s, and, thus, the sample is not biased by the exclusion of these addresses.

As can be seen, the degree distribution is highly similar in all subsets. Although the majority of vertices have only a low degree, the distribution exhibits also a substantial positive skew caused by the outlying vertices at the right tail. Moreover, by turning the attention to the last two plots, (c) and (d), it is evident that no large differences exist between the files that were detected as clean and malware. On average, a single unique file is typically distributed from a single unique 2-LD/1-LD. While the standard deviations are also rather small, these are mostly caused by the few extreme outliers.

The popularity metric can be hypothesized to follow any of the so-called extreme value distributions (Gumbel, Fréchet, and Weibull) including the often used generalized Pareto distribution (GPD). This conclusion is enforced by Fig. 2, which uses the quantile function shipped in the *evir* R package [9] for computation. The two extreme outliers are suspected to be Flash exploits, although only by two detection engines. These two popular files – whether truly malicious or not – are located in the bottom-right disconnected subgraph in Fig. 3, which was constructed by first obtaining all malware file vertices with a degree of 25 or more, and then querying for all domain-labeled vertices to which these were adjacent to.

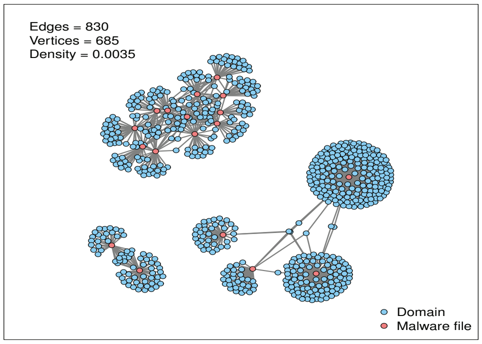

Fig. 3. Subgraphs of Popular Malware Files

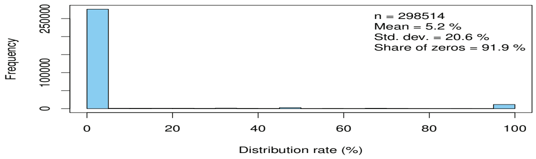

Fig. 4. Distribution Rate of Domains (2-LD/1-LDs)

The distribution rate across the about 0.3 million domains is summarized in Fig. 4. As was expected, the clear majority (about 92 %) of the aggregated domains have not distributed malware files at all. More interestingly, however, only a relatively few domains exhibit a rate in the interval $0 < r < 100$. That is to say: those domains that have distributed malware files have mostly distributed these without also distributing clean files. The total share of "100 % malware distribution" is about 3.8 %, but the share of $r = 100$ to $r > 0$ is as high as 47 %. Thus, for some supervised learning tasks, it may be reasonable to further dichotomize (4) according to a threshold.

III. Discussion

The remainder of this paper briefly summarizes the key empirical results, points out five limitations, and enumerates a few further directions for empirical malware research.

A. Findings

This short empirical paper operated with two theoretical concepts. The first was popularity, which was operationalized as the degree of labeled vertices representing malware in a file-based graph representation. The degree was observed to follow a typical long-tailed probability distribution; the well-defined GPD provides a decent reference point for the popularity of malware files. The other concept was distribution of malware files, which was defined as the rate of malware files to all files distributed from a given aggregated domain. This rate can be hypothesized to follow a bimodal probability distribution at the population level of the contemporary web-facing Internet.

That is, on one hand, the majority of domains distribute only clean files, but, on the other hand, there is supposedly a subpopulation that only distribute "near 100 % malware" files. In-between these two polar opposites seems to be a relatively small mixture compromised of domains to which malware is dropped to accompany other files distributed from the domains. The domain `dropbox.com` is a good example both empirically (with $r \simeq 2$ %) and metaphorically. Although it was beyond the scope of the paper to evaluate how many of the observed malware files refer to software designed to compromise web browsers, it thus seems reasonable to conclude that web surfing is not safe even in common and popular 2-LD/1-LDs. But this conclusion may also sound like a truism. Therefore – and since both operationalizations are also good at picking for statistical outliers, it can be also concluded that no web surfer should end up to some of the domains that represent the "near 100 % malicious" subpopulation within the hypothesized population level probability distribution mixture.

Thus, in terms of conventional blacklist-style solutions, the results reveal the enduring ground truth problem. On one hand, the results exemplify why there continues to be a demand for different blocking solutions. On the other hand, it continues to be difficult to classify domains, addresses, and files – not least because even Google may not always classify to the benign category. While different popularity lists (such as the one provided by Alexa) are frequently used to proxy the benign category [10], the results thus reinforce the argument about bad ground truths that these lists may induce [11]. Against this backdrop, it may be that further advances are more likely in the security intelligence domain, including security analytics, outlier detection, and related areas of research and practice.

B. Limitations

Five notable limitations can be enumerated, but these should be also balanced against the specified goal of hypothesizing about the population level characteristics. It can be started by acknowledging (i) the limitations imposed by a post-mortem analysis; nothing can be concluded regarding dynamics.

This paper analyzed over two million files distributed from nearly three thousand aggregated domains. While the amounts are large, (ii) care should be always used when attempting generalizations to the whole web-facing Internet. A more direct concern relates to the (iii) aggregation to 2-LD/1-LDs, which affects particularly the computed distribution rates. This limitation goes hand in hand with the crawling according to WWW hyperlinks: for instance, the denominator in (4) should make the rates negligible for domains such as `google.com` and `github.com`, which was not observed to be the case in the dataset. Moreover, (iv) the sample distribution for the observed $\mathbf{r}$ is directly affected by the scalar τ in (2). It may be that the utilized maximum alarm rate – the agreement of all of the $\mathcal{D}_1, \ldots, \mathcal{D}_d$ detectors – is not optimal due to the poor detection rate (i.e., false positives) of some of the engines aggregated by VirusTotal [8]. Finally, (v) malware is "always the same, never the same" [12], and, thus, it should be acknowledged that any modification to a malware file will change the associated MD5 hash, which will consequently disturb the simple popularity concept elaborated in this paper.

C. Future Research

The limitations of this paper also open plausible questions for further empirical research. For instance, to assess the severity of the limitation (ii), evaluation of the IP address space coverage would offer a good basic check [13]. Empirical evaluation of the limitations (iii) and (iv) would be also relatively straightforward to carry out. While conventional mixture modeling [14] could be adopted for pursuing the discussed points further, an evaluation of these intervening quantities would be generally beneficial for making advances with formal Bayesian models based on informative prior distributions.

It also seems reasonable to further continue the initiated work [4] for extending the scope from the conventional domain-based malware graphs. In this respect, the paper continues the work (here see [15]) for moving forward with different relational file-to-file representations [16]. To address the limitation (v), in turn, a theoretically motivated taxonomic approach to graph clustering could be adopted for examining the relational characteristics of malware species rather than individual malicious files. By considering how the so-called drive-by-download and related techniques work [2], [5], it seems reasonable to further hypothesize that average-to-high range distribution rates would be associated with a particular kind of malware species, for instance. This hypothesis also exemplifies the need for both active and passive harvesting techniques; the command and control channels are typically different from the channels that are used to download malware to compromised hosts [17]. Finally, it may well be that the so-called dark (deep) web would offer a more fertile ground for harvesting malware files – and the crawling of this Internet subpopulation is a non-trivial technical challenge in itself.

Acknowledgments

The authors gratefully acknowledge Tekes – the Finnish Funding Agency for Innovation, DIGILE Oy, and the Cyber Trust research program for their support. The authors also thank F-Secure for supplying proprietary data, the group behind "CleanMX" for supplying open data, and Rotarua Limited (d.b.a. VirusTotal) for enabling further computations.

References

[1] S. Khan, A. Gani, A. W. A. Wahab, M. Shiraz, and I. Ahmad, "Network Forensics: Review, Taxonomy, and Open Challenges," *Journal of Network and Computer Applications*, vol. 66, pp. 214–235, 2016.

[2] A. Nappa, M. Z. Rafique, and J. Caballero, "The MALICIA Dataset: Identification and Analysis of Drive-by-Download Operations," *International Journal of Information Security*, vol. 14, no. 1, pp. 15–33, 2015.

[3] A. Berger, A. D'Alconzo, W. N. Gansterer, and A. Pescapé, "Mining Agile DNS Traffic Using Graph Analysis for Cybercrime Detection," *Computer Networks*, vol. 100, pp. 28–44, 2016.

[4] A. Boukhtouta, D. Mouheb, M. Debbabi, O. Alfandi, F. Iqbal, and M. El Barachi, "Graph-Theoretic Characterization of Cyber-Threat Infrastructures," *Digital Investigation*, vol. 14, no. Supplement 1, pp. S3–S15, 2015.

[5] C. Grier, L. Ballard, J. Caballero, N. Chachra, C. J. Dietrich, K. Levchenko, P. Mavrommatis, D. McCoy, A. Nappa, A. Pitsillidis, N. Provos, M. Z. Rafique, M. A. Rajab, C. Rossow, K. Thomas, V. Paxson, S. Savage, and G. M. Voelker, "Manufacturing Compromise: The Emergence of Exploit-as-a-Service," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS 2012)*. Raleigh: ACM, 2012, pp. 821–832.

[6] Clean MX, "Realtime Database," 2016, Data feed available online in April 2016: http://support.clean-mx.de/clean-mx/viruses.

[7] C. Chen, J. Huang, and Y. Ou, "Efficient Suspicious URL Filtering Based on Reputation," *Journal of Information Security and Applications*, vol. 20, pp. 26–36, 2015.

[8] Rotarua Limited (d.b.a. VirusTotal), "VirusTotal," 2016, available online in April 2016: https://virustotal.com/.

[9] B. Pfaff, "evir: Extreme Values in R," 2012, R Package Version 1.7-3. Available online in May 2016: https://cran.r-project.org/web/packages/evir/index.html.

[10] A. K. Jain and B. B. Gupta, "A Novel Approach to Protect Against Phishing Attacks at Client Side Using Auto-Updated White-Lists," *EURASIP Journal on Information Security*, no. 9, pp. 1–11, 2016.

[11] A. Pinto, "Secure Because Math: A Deep-Dive on Machine-Learning-Based Monitoring," in *Black Hat Briefings*. Las Vegas: Black Hat, 2014, pp. 1–11, available online in May 2016: http://ubm.io/1ygkG8V.

[12] M. Ramilli and M. Prandini, "Always the Same, Never the Same," *Security & Privacy*, vol. 8, no. 2, pp. 73–75, 2010.

[13] J. Fritz, C. Leita, and M. Polychronakis, "Server-Side Code Injection Attacks: A Historical Perspective," in *Proceedings of the 16th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2013), Lecture Notes in Computer Science (Volume 8145)*, S. J. Stolfo, A. Stavrou, and C. V. Wright, Eds. Rodney Bay: Springer, 2013.

[14] Z. Zivkovic and F. van der Heijden, "Recursive Unsupervised Learning of Finite Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 651–656, 2004.

[15] E. Nissan, "An Overview of Data Mining for Combating Crime," *Applied Artificial Intelligence*, vol. 26, no. 8, pp. 760–786, 2012.

[16] L. Chen, W. Hardy, Y. Ye, and T. Li, "Analyzing File-to-File Relation Network in Malware Detection," in *Proceedings of the 16th International Conference on Web Information Systems Engineering (WISE 2015), Lecture Notes in Computer Science (Volume 9418)*, J. Wang, W. Cellary, D. Wang, H. Wang, S.-C. Chen, T. Li, and Y. Zhang, Eds. Miami: Springer, 2015, pp. 415–430.

[17] C. Rossow, C. Dietrich, and H. Bos, "Large-Scale Analysis of Malware Downloaders," in *Proceedings of the 9th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA 2013), Lecture Notes in Computer Science (Volume 7591)*, U. Flegel, E. Markatos, and W. Robertson, Eds. Heraklion: Springer, 2013, pp. 42–61.