Characterizing the Impact of Malware Infections and Remediation Attempts Through Support Forum Analysis

Sara Amini, Chris Kanich
University of Illinois at Chicago
samini3@uic.edu, ckanich@uic.edu

***Abstract*—Detection and removal of malware infections have always been significant concerns for every computer user. Countless people are victims of malicious programs per day all around the world despite substantial improvements in malware defense. Developing techniques to characterize the harm caused by these programs enables new defenses to counteract these behaviors. One way to create these metrics is to explore online remediation forums because so many people refer to them for help in finding solutions for their systems' malware-related problems.**

Here we describe and implement a technique to characterize and quantify the harm that victims encounter when their systems are infected with a specific malware strain. We analyze various malware families harmfulness by exploiting the user-generated data collected from Bleeping Computer, one of the most popular online malware remediation forums. Moreover, we quantify how successful and effective this type of online community is when it comes to addressing victims malware-related issues.

***Index Terms*—malware remediation, forums, malware harmfulness.**

I. Introduction

Despite significant improvements in malware defense, cybercriminals successfully deploy malware infections countless times per day. Understanding the profits of these cybercriminals allows us to better counteract their schemes and disincentivize this bad behavior, but the true goal of such an effort is to minimize the harm felt by the legitimate users of a given system. If we are able to develop techniques to characterize and quantify this harm, we can create metrics that correspond to the primary goal of cybersecurity, that of increased user safety.

Traditional quantification of malware's virulence and prevalence are typically denominated in terms of the number of infections: the more widespread a given family of malware is, the more damaging it is. The CVE system also includes a concept of severity for vulnerabilities, but this tracks what is possible, rather than what is achieved by an attacker. While number of infections works well as a metric for understanding the prevalence of a given infection, it does not capture the full experience of being infected by a given piece of malware. Different strains can cause widely varying effects on the victims' computers and the use thereof, which leads to differing impact on end users.

One website, bleepingcomputer.com [1], has a lively community that specializes in assisting users with malware infection remediation. This forum holds records of hundreds of thousands of attempts to remediate malware infections, including rich metadata like how long it took to perform the remediation, whether it was successful, and in many cases the name of the malware family that caused the infection. By mining this dataset, we are able to provide a rich characterization of the impact of these malware families' infections, which improves our understanding not only of how prevalent these infections are, but of how damaging they are in terms of users losing time or use of their computer due to such infections.

This project infers the harmfulness of various malware infections' by analyzing the data collected from Bleeping Computer, one of the most popular online malware remediation forums. Through a set of hand crafted heuristics, we mapped approximately 13% of all threads on the Bleeping Computer malware help sub-forum to malware family names, and found that this heuristics has 86% precision and 52% recall. While not comprehensive, this collection of 134,982 labeled threads serves as a dataset for analyzing the harmfulness of different viruses. Furthermore we investigate the harmfulness of the top 46 most frequent viruses (in terms of the number of distinct threads posted about that specific virus) in more detail throughout this paper. Finally, we discuss how successful and helpful Bleeping Computer is when it comes to malware remediation.

This work provides insights about the impact of various malware families' infection and harmfulness by exploring user-generated content information from actual victims attempting to remediate their own and others' malware infections. This work exists as a complement to previous measurement studies on remediation and underground forums [2], [3], analysis of cybercrime and fraud in online economic activity [4], [5], harm measurements and analysis of different kinds of loss users encounter due to different malicious activities [6]–[8].

Additionally, we have collected and made available the labeled dataset used in this paper[1] in the hope that it be used for further malware analysis and research on online remediation forums.

II. Data Collection

For the purposes of this project, we collected two different datasets. The first dataset is collected from one of the security

978-1-5386-2701-3/17/$31.00

[1]https://www.cs.uic.edu/~ckanich/datasets/BCLD.csv.gz

sub-forums of the Bleeping Computer website to capture the effort that victims expend to resolve malware-related problems they have with their computers. We chose Bleeping Computer because it is one of the most popular online forums for end users to assist each other with malware infections. The second dataset is a list of virus names collected from different sources such as VirusTotal [9], Symantec [10] and McAfee [11] which are used to map each thread on Bleeping Computer to one or more specific malware strains. A number of calibration steps are applied on the original collected datasets to get better results in terms of completeness and soundness.

A. Bleeping Computer online forum

This dataset is collected from the security forum of Bleeping Computer which is a resource site for asking and answering computer related technical questions. It consists of many forums which are organized by different computer related topics. People can readily register to the site using a valid email address with no fee and ask computer, security, and technical questions in a topic-related forum. There are experts who address these questions and offer expert opinions and suggestions to help virus victims resolve the issue.

We specifically used one of the sub-forums of the security forum which is mainly about malware issues named "Virus, Trojan, Spyware, and Malware Removal Logs" [12]. It includes 134,982 threads in total from 2004 to 2015 and each thread is followed by a number of comments. We collected the following fields about each individual comment on the forum:

- **Title of the thread:** Each victim initiates a thread by posting a title which includes a few number of words. Table II shows a few examples of titles.
- **Thread ID:** This ID uniquely identifies each thread.
- **Comment ID:** This ID uniquely identifies each comment.
- **Comment offset:** A one-indexed offset of the each individual comment within its containing thread.
- **Comment timestamp:** Timestamp that the comment was initially posted.
- **Comment body:** This is the actual content of the comment.
- **Author:** This indicates the author of each comment.
- **Views:** Number of views each thread has accrued.

B. Virus names

We gathered a list of virus names from the following sources to search for virus names within threads' titles and eventually map each thread to a specific malware:

- **Symantec.** The Symantec website maintains a "Listing of Threads & Risks" which we collected in its entirety.
- **McAfee.** We downloaded recent virus names from McAfee Labs' Threat Library.
- **VirusTotal.** We used 1.1 million malware labels extracted from a dataset provided by VirusTotal for a previous research project. [13]

The union of these datasets consists of approximately 1 million unique virus names after calibration which we discus in detail in Section II-C1. While this list is in no way all-inclusive and representative of the whole malware types, it was sufficient to be used to form a sample of threads which discuss a variety of different viruses. The fact that we were able to map 13% (17,528 number) of threads to a specific malware with a precision of 86%, demonstrates the adequacy of our virus names' list for these purposes.

C. Data sanitization

1) Forum data calibration: We applied the following calibration steps to enhance the quality of the forum dataset and results.

- **Select frequent viruses.** To extract meaningful and interesting insights about each virus, we require sufficient information about them. In total, we map different conversation threads onto 3,162 distinct viruses; however, for many viruses there is very little data in only one or a few threads. Therefore, we decided to select the most frequent viruses for analysis to get more reliable results. We chose a cutoff of at least 40 threads for an individual virus to merit consideration, which limited our deeper analysis to 46 different malware names discussed in 10,132 threads. Figure 1 shows these viruses with the corresponding measurements. Measurements will be discussed in detail in Section IV.
- **Remove duplicates and outliers.** To further clean the dataset, we remove threads which are duplicates, which we define as started by the same author with the same exact title. Moreover, there are some threads with very low number of responses. In these cases often the victim initiates a thread (title plus the initial response) and never comes back. A professional would start helping the victim by posting a response to the thread. If the victim does not come back, the professional would post another response which says **"Due to the lack of feedback, this topic is now closed."** and close the topic which means no more responses can be posted. That leads to a total number of three for the number of responses in this type of scenario. In order to actually capture meaningful information about the difficulty that victims encounter when attempting to remove the malware-related issues, we eliminate those threads with three or fewer responses. After these cleaning steps, the number of distinct threads was reduced to 7,390.

2) Virus names data calibration: We applied the following calibration steps to enhance the quality of the list of virus names and results.

- **Remove general family names.** In order to search for virus names, we search for virus name *n-grams* within threads' titles. Given a sequence of words, *n-grams* are an adjacent sequence of *n* words. We removed some of the very general *n-grams* in order to improve precision. Although removing these *n-grams* led to a decrease in recall (meaning we would find fewer threads which include a virus *n-grams*), we were able to come up with

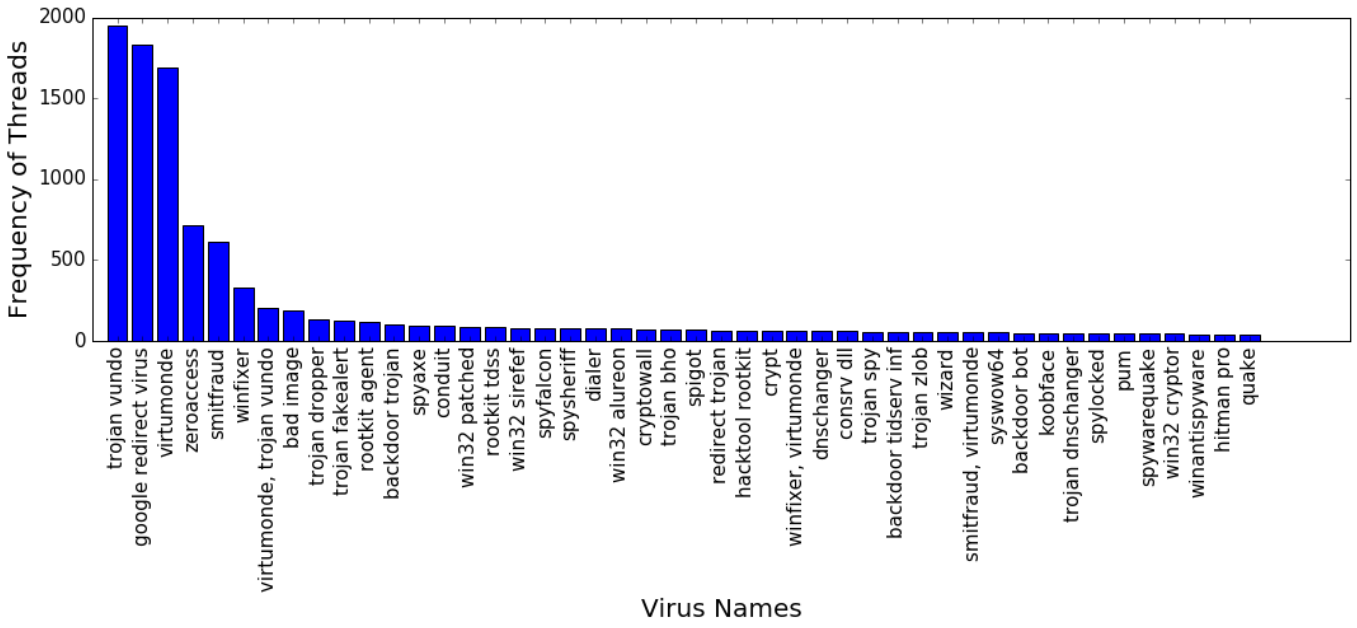

Fig. 1. Frequency of threads for each 46 top frequent viruses

an appropriate blacklist of general *n-grams* to hit an acceptable tradeoff between precision and recall at 86% and 52% respectively. Some of these general *n-grams* are shown in Table I.

TABLE I
BLACKLIST: GENERAL VIRUS *n-grams*

browser helper object	jump
microsoft corporation	ransom
malware	trojan
downloader	rootkit
proxy	adware

- **List expansion.** During the search phase, we noticed that there are a considerable number of viruses written in different ways which all are not included in the collected list. For example, viruses with the "win32" designation are usually written in two ways: "w32" and "win32". Another example is "Backdoor tidserv inf" virus which can also be written as "Backdoor tidserv !nf". Such virus family names can be found using edit distance and Jaccard similarity with an appropriate threshold. When each name is considered exactly as written, the attribution phase resulted in a pretty low accuracy. We expanded the list of virus names by including different spellings of the same virus to enhance performance. Our final list includes roughly one million virus names spellings.

III. METHODOLOGY

A. How we mapped threads

In order to map each thread to its relevant malware family, we search for virus name *n-grams* within titles of the threads. As a cleaning step, we removed all punctuation from both virus names and thread titles, converted all words to lower case, and tokenized both datasets into ordered word lists. Finally, we used a trie to efficiently search thread titles for virus names.

B. Evaluation and method accuracy

The first and the most significant objective of this project is to associate threads to the responsible malware family. To map each thread to a specific malware, we search for virus name *n-grams* within the threads' titles. A few examples of such titles are shown in Table II. We were able to map approximately 13% of all threads (17,528 out of 134,982 threads) to one or more specific malware names with 86% precision using our expanded virus name list. Also, precision and recall were computed by manually inspecting a 100 number of threads. We repeated the random inspection three different times with replacement. We found that our heuristic resulted in 86% precision and 52% recall, averaged over the three repetitions of manual 100 random thread inspections.

TABLE II
USER INITIATES A THREAD BY POSTING A FEW WORDS AS ITS TITLE. IN 25% OF ALL TITLES AT LEAST ONE VIRUS NAME IS EXPLICITLY INCLUDED. THIS TABLE SHOWS A FEW EXAMPLES OF SUCH TITLES.

I Am Infected With **Smitfraud**! Please Help!
Infected with **ZeroAccess** rootkit
ZeroAccess infection keeps putting nasty things in my machine
Help with removing **spigot** virus (Malware or Adware)
Trojan Vundo Help. I Give Up.....
infected by a **google redirect**

The second phase of the project is to quantify harmfulness of each virus and helpfulness of Bleeping Computer. For this

phase, we query the data to answer interesting and helpful questions such as:

- What are the common malware families during different periods of time?
- What is the distribution of occurrence time for different malware families?
- How harmful are different malware families, in terms of time lost to cleaning up a malware infection, success rate in cleaning infections, or quantity of users affected?

IV. Malware Externality Model

In order to analyze the amount of difficulty victims of malware attacks are encountered, we utilize data features extracted from Bleeping Computer as described in Section II-A. As discussed in Section II-C, we select threads discussing the top 46 most frequently referenced viruses. In this section, we first explain different measurements calculated using data. Then, we analyze interesting and significant insights inferred from these measurements and mainly discuss the following two main questions:

- How successful is Bleeping Computer and its users in helping victims alleviate the symptoms of malware infections?
- How harmful are the most frequent viruses in terms of the time and effort victims expend to resolve their issues?

A. Summary of measurements

Here we introduce various quantities to consider as proxies for the impact of different malware infections.

- **Frequency of threads.** This indicates the number of distinct threads per virus. It is calculated by counting the number of distinct threads for each label.
- **Average of all responses.** Each thread is followed by a number of responses. This value indicates the average of responses of threads for each virus. We compute this value by dividing sum of total responses by threads frequency for each virus.
- **Average of responses by initiator.** This measurement is the average of responses posted by the initiator of the thread for each virus. We compute this value by dividing sum of total initiator responses by threads frequency for each virus.
- **Average views.** This value is the average of views of threads for each virus. It is calculated by dividing the total number of views by the thread frequency for each virus.
- **Total views.** This value is the sum of views of all threads attributed to each virus.
- **Resolved percentage.** This value is the percentage of resolved threads for each virus. If a thread is resolved, the professional would close it by posting the last comments which says ***"It appears that this issue is resolved, therefore I am closing the topic."***. We compute this value by dividing the number of threads which last comment includes the word ***"resolved"*** by the total threads frequency for each virus.
- **Time duration.** This value shows the time duration in which each virus is discussed on the forum. For this value, we assume the duration of each virus is the time difference between the earliest and the latest threads in which a specific virus name is discussed.
- **Average time duration of threads.** This value is the average time duration of threads for each virus. We compute the difference between the first and last comment of each thread about a given virus, and then take the average of those time durations to approximate how long a user spends dealing with that virus on average.

Statistics of the aforementioned measurements for the top 46 most frequent viruses are summarized in Table III. As you can see in this table, the overall average view count for threads about these viruses was 2906.9 and the minimum number of views is 1459.8. The mean resolved rate is 51.5% which means on average about half of the threads pertaining to these viruses were resolved. Victims having issues related to these viruses spent 15 days on average to resolve the issue which is a considerable amount of time. In some cases it took them up to 25 days to resolve the issue. Among these frequent viruses, some malware families were constantly discussed on Bleeping Computer, with a time duration of 10 years, which is effectively the entire lifetime of the forum. On average, these viruses remain active for 5.7 years which shows their complexity and evolution. Many different threads are posted about the same families: up to 1951 different threads for a single virus. This again demonstrates that in some cases existing threads are not enough to address the issues and significant effort is expended to remove individual infections. The average number of responses, 15.6, depicts the valiant effort of professionals and victims to remove virus infections. Figure 3 presents boxplots of these measurements, which enable one to consider both the aggregate statistics of the distribution as a whole, as well as the value of outliers, which are shown as blue dots and will be discussed further later in this section.

B. Modeling how widespread malware families are

1) Total number of views: One of our approaches to explore the impact of each malware family in terms of being widespread, is to analyze the total number of views for each malware instance. While the number of views may have been inflated by advertisements or crawling bots, we believe the comparison of number of views can serve as a proxy for the relative popularity of different threads. Table IV shows the top six viruses with the highest number of total views. All of these most popular viruses appeared on Bleeping Computer from 8 to 9 years except for *zeroaccess*. In around 4 years, threads posted regarding *zeroaccess* got around 2 millions views which is an emphasis on its high level of prevalence. Moreover, Figure 4 shows that *zeroaccess* got released in 2011 and during 2013 it was featured in many threads. The resolved rate for this virus is around 68.8%. It is among top 5 most resolved viruses in Bleeping Computer due to our results. This considerably good rate of resolved threads may be because of various

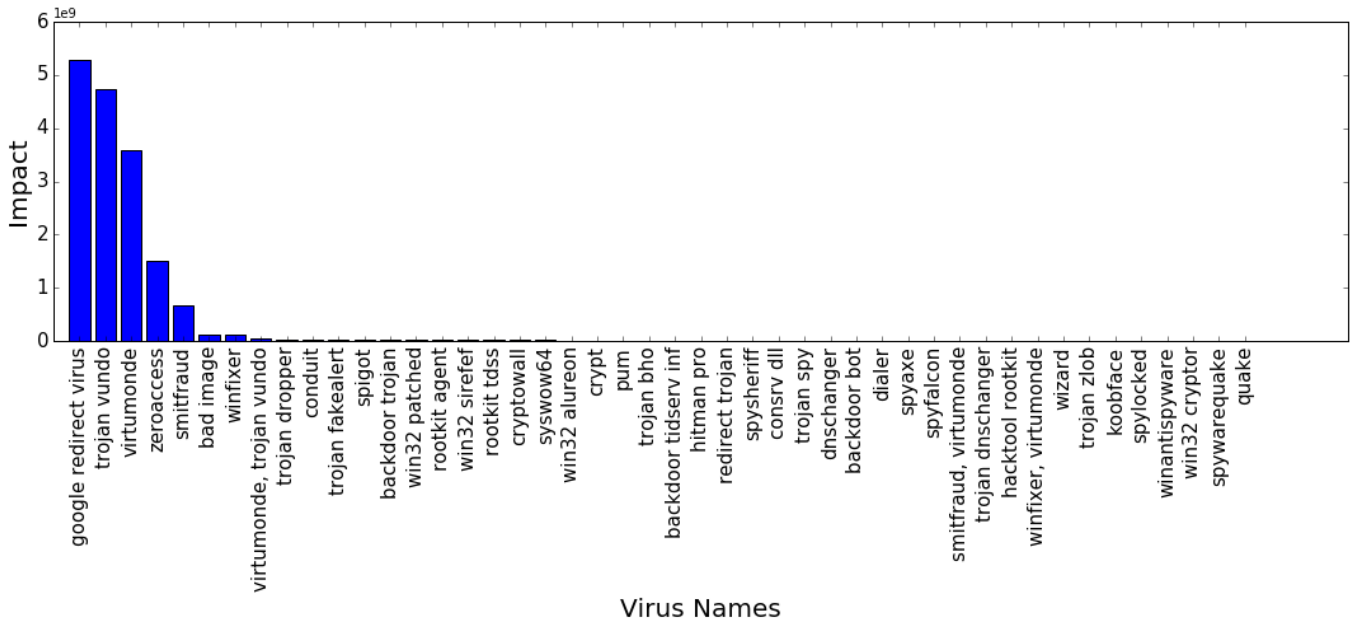

Fig. 2. The amount of impact for each 46 top frequent viruses

TABLE III
SUMMARY OF MEASUREMENTS FOR TOP 46 FREQUENT VIRUSES

Summary	Thread Frequency	AVG Responses	AVG Initiator Responses	AVG Views	Resolved Rate	Time Duration (year)	AVG Thread Duration (day)
mean	219	15.6	7.8	2906.9	51.6	5.7	15
stddev	436	3.7	1.9	1253.55	12.8	2.7	3
min	40	9.9	4.7	1459.8	21.6	0.35	8.8
max	1951	25.2	13	7095	73.1	10.2	21.7

reasons: perhaps *zeroaccess* is not hard to get resolved, its virulence encouraged experts to develop specialized removal tools, to or Bleeping Computer has shown a good performance on helping users removing *zeroaccess* and resolving related issues. We elaborate more on these potential explanations in Section IV-C.

The top 4 viruses listed in Table IV, the 4 outliers in Total Views subplot shown in Figure 3, also have the highest threads frequency among all frequent viruses according to Figure 1. *google redirect virus*, *trojan vundo* and *virtumonde* are also outliers in the Thread Frequency subplot shown in Figure 3. This is evidence of how virulent these malware instances are. The large number of threads along with even higher amount of views for these viruses means that even though there are already so many threads posted about them, there is still a need to post another one to seek for help and find a solution. Victims likely do view posted threads and attempt removing the virus using the existing solutions, but still cannot resolve the issue. Hence, they initiate another thread to get help from professionals specifically for their system's malware-related problems which due to IV results, takes around 16 days on average. Taking all of these factors into account, the average 54.2% resolved rate for these top 4 viruses shows the capability and effectiveness of online remediation forums such as Bleeping Computer in addressing victims concerns related to frequent viruses.

The *Google redirect virus* forcefully redirects users' Google web searches to malicious web pages, which is likely the side effect of a monetization scheme enacted by some other malware on their computers. Not surprisingly, *google redirect virus* is the top most viewed and prevalent virus on Bleeping Computer. First, web-based malware instances can be readily deployed on web sites [14]. Second, because this family name is a description of a symptom rather than an underlying infection, it is possible that it refers to several different underlying malware families. Even so, the resolve rate of 52% is roughly in line with other viruses, so the potential that this family is substantially more difficult to remove is low.

2) Most viewed viruses on Google: As mentioned in the previous section, we noticed there are some malware families with a very high number of views compared to their number of responses. We used the Pearson correlation to measure the linear correlation between the average number of views and responses. Unsurprisingly, as shown in Figure 5, the number of responses and views have a positive correlation meaning as the number of responses grows, there is an increase in the number of views. The one significant outlier is the *Spigot* virus, which is also shown in Views over Responses subplot of

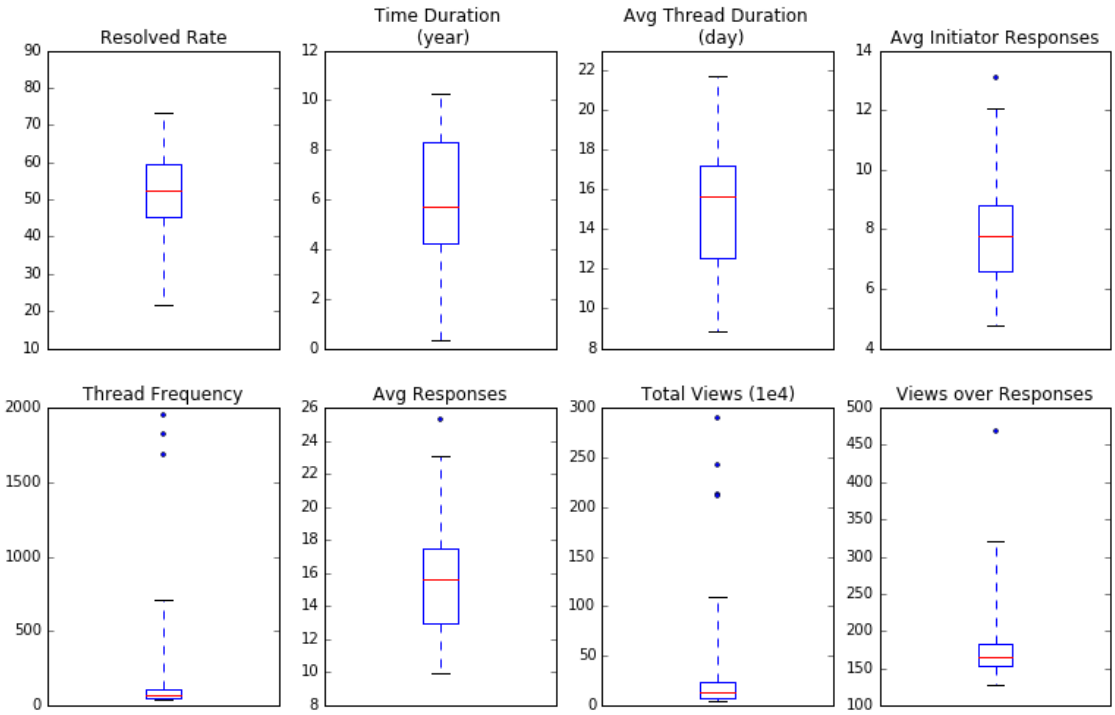

Fig. 3. Boxplots for different measurements. Outliers are shown as blue dots. Resolved Rate, Time Duration and Average Thread Duration subplots do not have any outliers. *zeroaccess* malware is the only outlier in Average Initiator Responses and Average Responses subplots. In Thread Frequency subplot, the 3 outliers are *google redirect virus*, *trojan vundo* and *virtumonde* from the highest to the lowest. These viruses are also outliers in the Total Views subplot along with *zeroaccess*. *spigot* is the only outlier in the Views over Responses subplot.

TABLE IV
SUMMARY OF MEASUREMENTS FOR THE MOST VIEWED AND FREQUENT VIRUSES

Malware	Total Views	Thread Frequency	Resolved Rate	Time Duration (year)	AVG Thread Duration (day)	AVG Responses	AVG Initiator Responses
google redirect virus	2,896,178	1,828	57.5	9.34	15.59	17.56	8.46
trojan vundo	2,430,233	1,951	45.3	9.75	17.15	13.76	6.99
virtumonde	2,124,203	1,687	45.3	8.83	14.49	13.19	6.70
zeroaccess	2,115,597	714	68.8	4.38	16.32	25.27	13.09
smitfraud	1,096,949	616	38.6	8.86	12.35	13.22	6.88
bad image	598,744	187	51.2	8.32	17.21	16.23	8.03

Figure 3. This high ratio of views over responses demonstrates the virulence of this type of virus among indirect victims, meaning those who have not posted on Bleeping Computer to directly ask question about their computer issued regarding *spigot*. They are those who have searched for a solution to this infection and eventually found Bleeping Computer threads regarding *spigot*. We manually inspected these threads (there are 54 of them). We observed that the symptoms for this virus are easy to identify and the removal steps suggested by professionals are easy to follow. *spigot* malware infections do not seem to be difficult resolve by the fact that the average number of responses by initiators is 7 and the high average resolution percentage of 70.4%. However, the high number of views indicate that this is likely a very common if not difficult to remove infection.

C. Modeling how difficult malware families are to get resolved

While the number of threads and views can be used to infer how widespread a virus is, we can model the difficulty victims have encountered removing these infections by exploring the number of responses and the duration of each thread. Specifically with respect to *zeroaccess*, we explore the difficulty of removing this virus in terms of how often victims continue to reply on their threads (under the assumption that users will continue to respond while their infection persists and they have not yet succeeded or given up on cleaning the infection). We refer to the average number of total responses and the average number of responses by initiator/victim. Using the

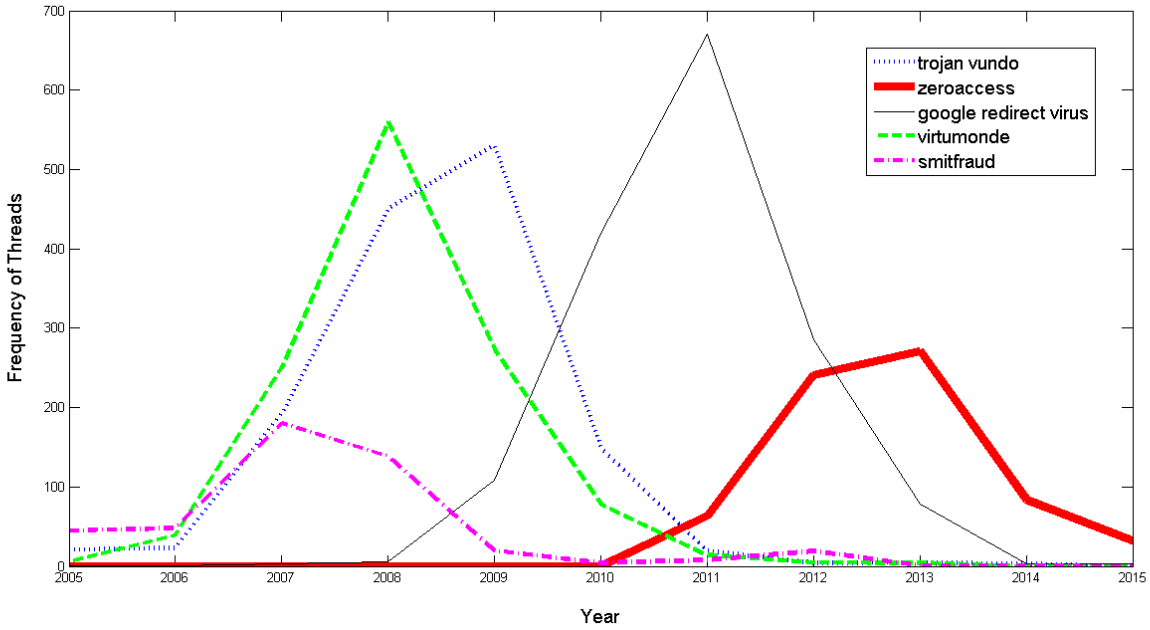

Fig. 4. Distribution of threads of top frequent viruses over time

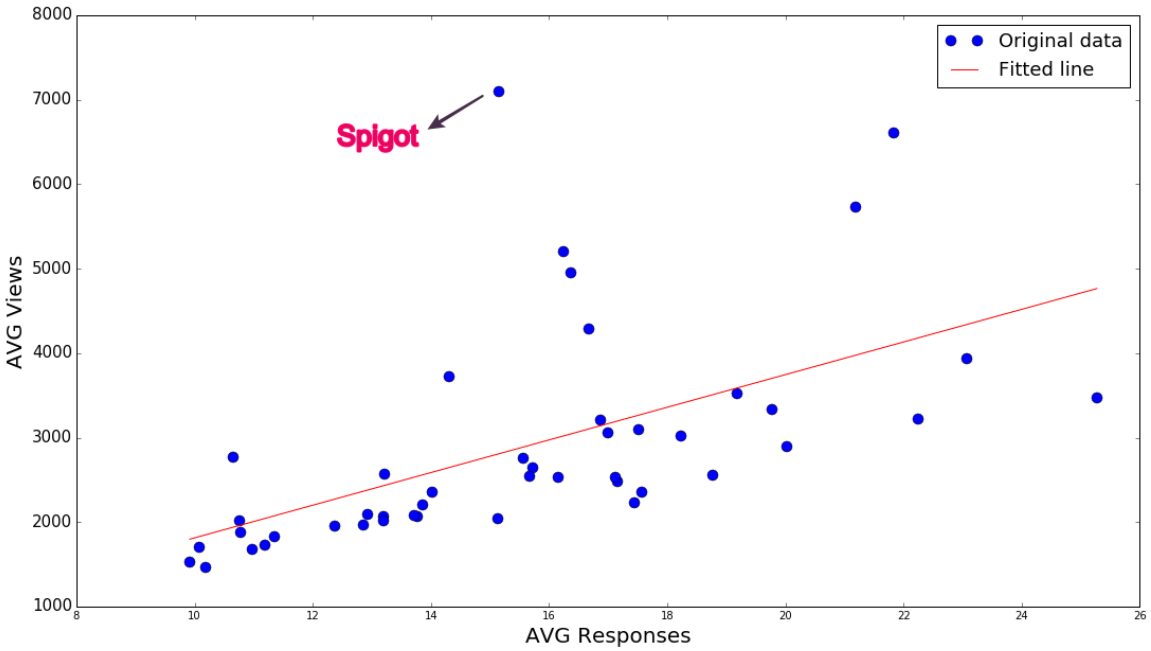

Fig. 5. Average number of responses vs. Average number of views for frequent viruses along with the Pearson correlation fitted line

boxplot for Avg Responses and Average Initiator Responses in Figure 3, we noticed that *zeroaccess* is an outlier in terms of the average number of total responses as well as the average number of responses by initiators. These numbers are 25.27 and 13.09 respectively as shown in Table IV. In other words, while *zeroaccess* is more likely to be resolved than a typical popular virus, our metrics show that it is actually more difficult to resolve. One potential reason for the relationship between this success rate and effort is that *zeroaccess* is an especially deep infection, which likely would completely incapacitate a computer and was especially resistant to removal.

D. Total impact

With the above described different aspects of malware infection harm, we build a model for the total impact of a virus infection, shown in Equation 1.

$$Impact \propto \|Number\ of\ Threads\| * \|Number\ of\ Views\| \quad (1)$$

There are two types of victims who seek help on Bleeping Computer: direct and indirect victims. Direct victims are those who directly initiate a thread on Bleeping Computer and seek help to resolve their system's malware-related issues. The other type of victims, indirect ones, are those who view already posted threads regarding their system's virus infection to find a way to remove the virus.

The *number of threads* reflects the impact of the virus's infections in terms of the number of different direct victims looking for help on Bleeping Computer. The larger the number of threads pertaining to a specific virus is, the more difficult the virus is to be solved. In other words, if the number of threads is increasing over time even though there are some threads already posted about that virus, it means that posted solutions are not sufficient to remove the virus and the need for starting a new thread still remains.

In addition, the *number of views* mostly demonstrates the prevalence of the virus among indirect victims. A higher number of views indicates that more people search for the removal of that specific virus and eventually view the relevant threads on Bleeping Computer.

Figure 2 shows the impact for top 46 frequent viruses. Following our analysis, *google redirect virus*, *trojan vundo*, *virtumonde*, *zeroaccess* and *smitfraud* malware families have an exponentially larger infection impact than the other popular viruses.

V. Observations

Through our research of the Bleeping Computer dataset, we have found that *google redirect virus*, *zeroaccess*, *smitfraud*, *vundo* and *virtumonde* are the most harmful malware families in terms of not only their virulence but also their removal difficulty and the amount of time victims lose to resolve their infections. Although these malware instances are harder to remediate, the average resolution rate of threads discussed regarding these infections on Bleeping Computer is roughly 52%, which shows that while it is by no means a silver bullet for malware removal, posting a thread on Bleeping Computer is a worthwhile use of time when addressing malware-related issues.

It is also interesting to note that the *spigot* malware has an especially high ratio of views per responses among all frequent viruses. In other words, there have been many victims who have this type of infection who have found and viewed appropriate solution to remove it on Bleeping Computer. As the *spigot* resolution ratio is 70.4%, nearly the highest in our entire dataset, we believe that those viewers were able to successfully remediate their infections (and thus not open yet another thread on the topic of that malware).

As mentioned previously, we labeled each thread to a specific malware instance by searching virus names within their titles. In other words, if victims know what virus they are dealing with, they initiate a thread on the forum including the name of virus they suspect infected their computer. A few examples of such threads are included in Table II. By manual inspection, we have found that in 86% of cases the problem is actually about the virus name found within the threads' titles. All the information covered in this paper including the resolved rates is based on this assumption that by high confidence victims know what type of malware family they are dealing with and they actually include the name when asking for help on the forum. Through our research, we noticed there are some threads that include the *bi-gram unknown virus* in their title. These threads are actually among the most frequent ones which means they are more than 40 threads on the forum regarding *unknown virus*. We found this interesting that the average resolved rate for this kind of threads is around 48.2%. Comparing it to that of the frequent known viruses (roughly 52%), this demonstrates the ability of Bleeping Computer professionals in addressing malware issues and helping victims resolve them even though they are not informed by victims about a suspicious specific malware family.

VI. Conclusion

In this paper, we presented an approach to estimate the impact of various malware instances' infections based on the information collected from a lively community called Bleeping Computer that specializes in assisting users with malware infection remediation. First, we presented our methodology to map a descent number of this forum's threads to the relevant malware family. Afterwards, by mining this dataset, we provide an understanding not only of how prevalent these infections are, but also of how damaging they are in terms of users losing time or use of their computer due to such infections. Moreover, based on results of our research, we believe Bleeping Computer has been successful and advantageous to great extent in helping infected users to resolve various malware families' infections especially the most frequent ones.

VII. Acknowledgements

We would like to thank the anonymous reviewers for their extensive and helpful feedback. This material is based upon work supported by the National Science Foundation under

Grant Nos. 1405886 and 1351058. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] BleepingComputer, "Bleepingcomputer.com - news, reviews, and technical support," https://www.bleepingcomputer.com.

[2] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy, "Doppelgänger finder: Taking stylometry to the underground," in *Security and Privacy (SP), 2014 IEEE Symposium on*. IEEE, 2014, pp. 212–226.

[3] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 71–80.

[4] P. Snyder and C. Kanich, "Characterizing fraud and its ramifications in affiliate marketing networks," *Journal of Cybersecurity*, vol. 2, no. 1, p. 71, 2016. [Online]. Available: +http://dx.doi.org/10.1093/cybsec/tyw006

[5] S. Afroz, V. Garg, D. McCoy, and R. Greenstadt, "Honor among thieves: A common's analysis of cybercrime economies," in *eCrime Researchers Summit (eCRS), 2013*. IEEE, 2013, pp. 1–11.

[6] P. Kotzias, L. Bilge, and J. Caballero, "Measuring pup prevalence and pup distribution through pay-per-install services," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 739–756. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/kotzias

[7] M. T. Khan, X. Huo, Z. Li, and C. Kanich, "Every second counts: Quantifying the negative externalities of cybercrime via typosquatting," in *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, 2015, pp. 135–150.

[8] C. Kanich, N. Weaver, D. McCoy, T. Halvorson, C. Kreibich, K. Levchenko, V. Paxson, G. M. Voelker, and S. Savage, "Show me the money: Characterizing spam-advertised revenue." in *USENIX Security Symposium*, 2011, pp. 15–15.

[9] V. Total, "Virustotal - free online virus, malware and url scanner," https://virustotal.com/en/about/.

[10] S. Corp, "Symantec corp - a-z listing of threats & risks," https://www.symantec.com/security_response/landing/azlisting.jsp.

[11] McAfee, "Mcafee - computer virus attacks, information, news, security," https://home.mcafee.com/virusinfo.

[12] BleepingComputer, "Bleepingcomputer.com forums - virus, trojan, spyware, and malware removal logs," https://www.bleepingcomputer.com/forums/f/22/virus-trojan-spyware-and-malware-removal-logs/.

[13] B. Miller, "Scalable Platform for Malicious Content Detection Integrating Machine Learning and Manual Review (Dataset)," http://secml.cs.berkeley.edu/detection_platform/, 2016.

[14] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, N. Modadugu *et al.*, "The ghost in the browser: Analysis of web-based malware." *HotBots*, vol. 7, pp. 4–4, 2007.