On the Evolution of Malware Species

Vasileios Vlachos[1], Christos Ilioudis[2], and Alexandros Papanikolaou[1]

[1] Department of Computer Science and Telecommunications,
Technological Educational Institute of Larissa,
Larissa, GR 411 10, Greece
{vsvlachos,alpapanik}@teilar.gr
[2] Department of Information Technology,
Alexander Technological Educational Institute of Thessaloniki,
P.O. Box 141 GR, 57400 Thessaloniki, Greece
iliou@it.teithe.gr

Abstract. Computer viruses have evolved from funny artifacts which were crafted mostly to annoy inexperienced users to sophisticated tools for industrial espionage, unsolicited bulk email (UBE), piracy and other illicit acts. Despite the steadily increasing number of new malware species, we observe the formation of monophyletic clusters. In this paper, using public available data, we demonstrate the departure of the democratic virus writing model in which even moderate programmers managed to create successful virus strains to an entirely aristocratic ecosystem of highly evolved malcode.

Keywords: malware, computer virus, phylogeny, cybercrime, malware writers.

1 Introduction

Malicious software is one of the most persistent threats to computer users. Earlier types of malcode debuted at the mainframes [1, 2], but a substantial rise could be attributed to the proliferation of home and personal computers [3]. Computer virology was theoretically and experimentally established by Fred Cohen and his supervisor Leonard Adleman [4]. Since then, computer viruses and other parasitic applications have became a common albeit annoyance for most computer users. As a result a multibillion world market for security applications has emerged and soared since then. Europe spent more than 4.6 billion EUR for security applications and services in 2008 [5]. According to antivirus vendors more than 4500 new malware species appear daily [6]. The effective handling of such a large number of threats requires substantial efforts and resources, human as well as computational, in order to provide timely remedies and protective measures. As consequence the absolute number of malware species constantly increases and at the time exceeds 2.6 million threats [6]. The overwhelming majority of the malware is either proof of concept code or flawed malicious programming attempts. Only a small number of viruses and worms manages to propagate in the wild (or in other words to reach and affect normal users), and merely a

H. Jahankhani et al. (Eds.): ICGS3/e-Democracy 2011, LNICST 99, pp. 54–61, 2012.

handful of them had the potential to become epidemics or pandemics. Therefore it is necessary to prioritize the imminent malware threats and devote the appropriate resources accordingly. In this paper we analyze a large data set of the computer viruses and other forms of malcode, that have been seen in the wild and we evaluate the current landscape so as to identify current hot spots that should trigger immediate attention. We believe that through the understanding of malcode evolution, a prioritization of current threats is both viable and beneficial. By extending the well established Darwinian theory, we find that the small percentage of computer viruses which is capable to mutate and adapt to the environment, is responsible for the majority of the security incidents.

The rest of this paper is organized as follows: Section 2 summarizes the related work, Section 3 presents and discusses our findings, whereas Section 4 concludes this paper along with possible future directions.

2 Related Work

A number of analogies between biological and computer viruses have been revealed [4, 7] in the past and more recently [8, 9]. An important outcome of this approach is the realization that the monocultures are particular harmful for the security of the software ecosystem [10–13]. Most of the work, however tackled the evolution of the security mechanisms from the defenders perspective [7, 14–16]. A more aggressive strategy would focus on reconnaissance of the weak points of the malware development process through biological analogies. *Phylogenetics* is the study of the relationships between organisms based on how closely they are related to each other. Researchers have applied similar methodologies to investigate the evolution of software and malware in particular, either using manual methodologies [17] or automated techniques [18–20]. It is reasonable to expect that only successful viruses will have the chance to mutate and eventually to create phylogenetic clusters. Therefore the WildList is better suited to become the basis of an evolutionary study. Though there is no reason to believe that the actual number of computer viruses differs from the estimation of major antivirus vendors, there is a clear difference between the malcode that has been developed for proof of concept purposes, *in vitro* environments and the number of malware strains that can be found *in vivo*. Moreover even if a virus circulates, it is not expected to cause significant damage given the total number of viruses in the wild. In our previous work we examined the factors that contributed to the success or the failure of a worm [8]. In this study we decided to utilize data from the WildList Foundation to capture the malware dynamics that have been seen in the wild. This list is somehow arbitrary as it is based on a limited number of participants, but as we will discuss, we believe that it provides significant advantages over other traditional approaches [21]. Despite the fact that some antivirus vendors [22] and researchers [23] do not agree with the methodology used by the WildList, still in general "*it is considered as an authoritative collection of the widespread malcode and is widely utilized as the test bench for in-the-wild virus testing and certification of anti-virus products by the* ICSA *and Virus Bulletin*" [24].

Various AV vendors provide statistical data about the proliferation of computer malcode, paying more attention to the evolution of the malware codebase and the financial motives of their developers [6]. On the other hand researchers have focused on interviewing malware writers in order to explain their psychosynthesis [25–28]. These findings are important and useful, but have not been updated and correlated with the current trends. Our work shows that the development of malcode is no more a "democratic" activity, in which any individual with moderate skills (for fun, political, religious or other reasons) could develop a new strain of a computer virus and cause significant or widespread damage. Most modern malware incidents are the result of a few number of prominent malcode families which dominate the landscape and are responsible for most annoyances and damages. The rate of which improved versions of the specific families are rolled out predominates most of the malware activity.

3 Discussion

Although the current malware activity can be obtained through various sources, we deliberately choose to work with the WildList because we believe it represents better the observed malcode dynamics. According to their definition "*The list should not be considered a list of 'the most common viruses', however, since no specific provision is made for a commonness factor. This data indicates only 'which' viruses are In-the-Wild, but viruses reported by many (or most) participants are obviously widespread*". In other words, this list contains the viruses, worms and other types of malicious software that succeeded to propagate sufficiently to be detectable, which clearly excludes proof of concept prototypes, academic examples, or ill engineered malcode artifacts.

The WildList employes an arbitrary naming scheme to identify malware treats which is basically the name most used by different AV scanners or the name given a virus by the person who first reported it. For the purpose of identifying malicious code of the same malware family we analyze the archives of the Wild List Organization from July 1993 till June 2010 and we taxonomize them according to their name. For example during January 2008 we identified several worm strains as members of the W32/Feebs family. This approach which is based on the categorization of the WildList is not as detailed as the manual or automatic inspection of the malcode using "phylogeny model generators (PMGs)" [18] so as to discern their phylogenetic characteristics. Nonetheless we find the method of the WildList Organization sufficient to correctly categorize most of malcode species to malware families. Another issue with the Wild List is the fact that does not provide absolute numbers regarding the malevolent activity of the malware species. Therefore we are not able to know the number of infections so as to categorize the viruses and the worms according to their virulence. As a result a worm with a single entry in the WildLight might have caused more infections than all mutations of a malware family. On the other hand the fact that numerous mutations of a malcode phylogeny managed to propagate to a wide scale so as to be included in the WildList is indicative of its capabilities to exploit a large pool of victims.

In order to proceed with the classification we used a small bash script to download all the monthly archives form the WildList Organization. A Python program stripped all the unnecessary content of the archives and a subsequent Python application identified the malware families and performed analysis on the data. Our applications processed 175 files containing 238474 lines of text which were eventually stripped down to 69820 lines of data.

These data were the basis of the analysis for identifying the current threats in computer virology. The first and most observable trend indicates an important clusterization of the malicious software to a small number of malware families. From Figure 1 we can witness that the percentage of the malcode species that belong to a dominant malware family does not show significant change in respect to the first available data of the year 1993. Though one can observe evident increase for some months after the February of 1997 as well as for the period of the last years (after 2005), the latest measurements show a stabilization of the dominant malicious activity related to the dominant malcode family around 15% of all the viruses, worms, spyware families that were found in the wild each month. Far more important are the findings if we analyze the trends of the three, five or ten most dominant families in conjunction. In that case we can observe that according to the latest data (January 2010) the three most dominant families represent now the 40.81% of all malware species that have been actively circulating compared to a mere 24.04% of the first available data at the July of 1993. The five most dominant families at the same period show a serious increase from 28.85% to 58.77%, where for the ten most dominant malware families we recorded a substantial growth from 38.46% to 77.42%.

The trends depict a significant change of the malware activity. Our interpretations of these findings agree with the work of S. Gordon [25–27], who examined the motivation of malware writers from a psychological perspective and that of S.Savage et al [29], which focused on the economic initiatives that drive the proliferation of computer crimes through the development and the maintenance of botnets. The earliest data (1993) depict a number of different malware strains that managed to propagate sufficiently so as to be included to the WildList. This trend eventually fades out as very few dominant malware families and their respective members represent the vast majority of the viruses that succeed to circulate at large. Therefore it is not as easy for a malicious entity to develop a new virus, worm or spyware as it used to be fifteen years ago. On the contrary one has much better chances to achieve widespread infection using a modified or extended version of a well maintained malware family. Based on the data analysis, an extended view of the malicious software landscape is available in Figure 2. Unfortunately, due to space limitations we had to include only the viruses, worms, spyware and bots that had more than 100 entries in the WildList in total and hence Figure 2 contains only 97 from the 821 malicious applications that were identified in the WildList.

Further analysis of the data indicates that the top ten malware families account for the 37.4% of the 817 total incidents that have been recorded in the WildList, while the top ten malware species are responsible for the 48.5% of all the incidents. In other words ten malware phylogenetic clusters are accountable for half of the cases that formulate the WildList so far. The common characteristic of the top ten entrants is that they have caused widespread problems and are also well known for their ability to mutate rapidly.

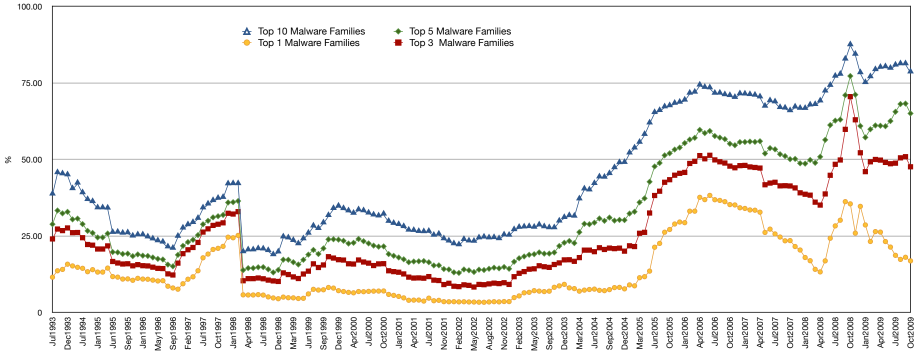

Fig. 1. Percentage of malware incidents attributed to top malcode families

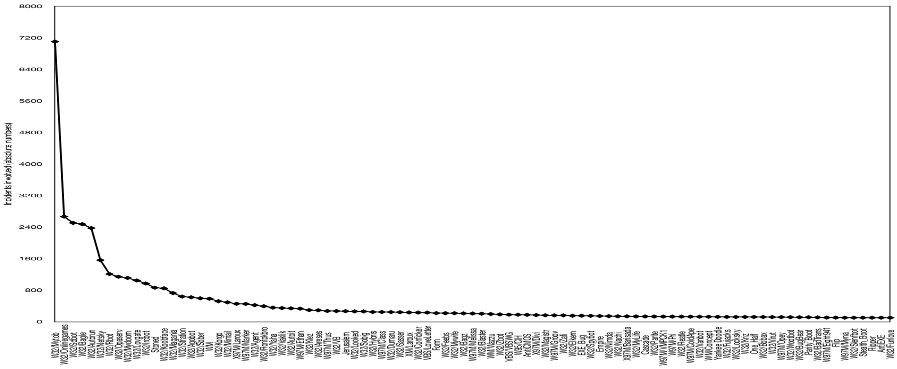

Fig. 2. Dominant Malware in the WildList

The implications of these findings are important as they suggest that most of the viruses, worms, spyware, do not manage to propagate in the wild and remain *in vitro* samples of malicious code. Even the malcode that manages to infect a sufficient number of victims so as to be included in the WildList, either mutates and evolves rapidly, or eventually diminishes and vanishes. Therefore only well written malcode, which offers high degree of upgradability or can be easily mutated, has improved chances to survive in the wild for a sufficient period.

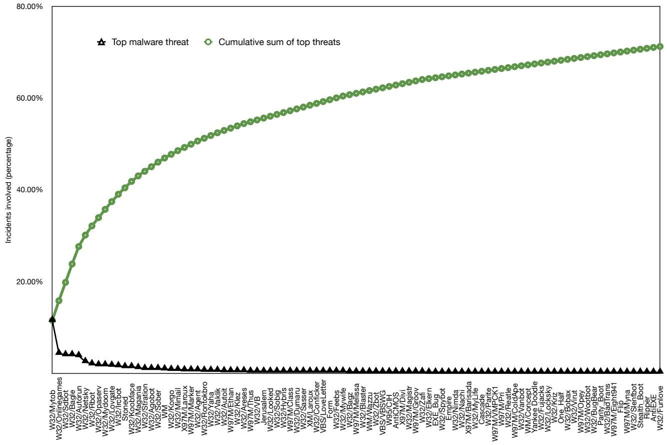

Fig. 3. Top family per month

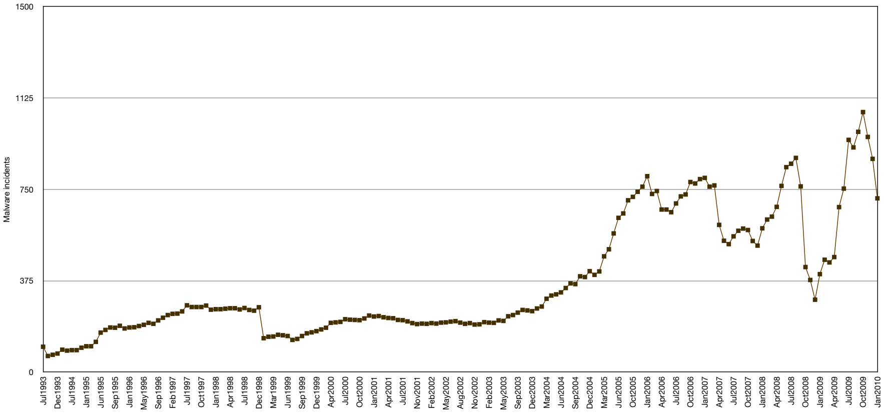

Fig. 4. Number of incidents per month

4 Future Work and Concluding Remarks

Our analysis is targeted to identify the dominant malware phylogenetic clusters and to indicate via statistical means that virus writing has become a professional activity in which amateurs with moderate skills are no more eligible to participate. Of course the available data of the WildList Organization could reveal other significant characteristics of computer virology. Our intention is to work in the future towards the prediction of the imminent threats by implementing econometric models and technical analysis on security data. Specifically, known models such as AR, MA and ARMA could be used to predict future threats depending on past data by finding self-similarities and periodicity.

The latest highly sophisticated malcode of the largest malware families indicates an escalation of the security arms race between malware writers and security researchers. The analysis of the WildList data emphasizes on the fact that malware writing is not any longer a trivial task. Gone are the days when disgruntled teenagers, activists or college dropouts could wreak havoc using simplistic programing tricks and earn their 15 minutes of fame. Competent malware should be able to mutate rapidly so as to propagate sufficiently and overcome the creation of effective signatures and evade other security mechanisms. The available data on the other hand signalize that the spreading of a virus or a worm in a wide scale is far from a trivial task. Therefore from a malware perspective it is better to work on a well maintained malicious code base than to develop new virus strain from scratch. Security professionals might found more promising an approach which prioritizes and concentrates their efforts against the most dominant malware phylogenies rather than trying to neutralize an overwhelming number of threats. For that reason if the available recourses are not adequate, it would be more productive for the research community to focus on the largest malware families, to monitor closely all the related developments and disseminate as fast as possible any findings of this activity. For years malcode developers exploit the monoculture weakness of modern IT in order to perform their vicious acts. By turning our attention to the most common and widely used malcode, we can exploit their tactics for our benefit.

References

1. Ferbrache, D.: A Pathology of Computer Viruses. Springer, NY (1992)
2. Szor, P.: The Art of Computer Virus Research and Defense. Addison-Wesley, Upper Saddle River (2005)
3. Skoudis, E.: Malware: Fighting Malicious Code, 6th edn. Computer Networking and Distributed Systems. Prentice Hall, NJ (2004)
4. Cohen, F.: Computer Viruses: Theory and Experiments. In: Proceedings of the 7th National Security Conference, pp. 240–263 (1984)
5. Anderson, R., Böhme, R., Clayton, R., Moore, T.: Security Economics and the Internal Market. Technical report, European Network and information Security Agency (ENISA) (2008)
6. Turner, D., Blackbird, J., Low, M.K., Adams, T., McKinney, D., Entwisle, S., Wueest, M.L.C., Wood, P., Bleaken, D., Ahmad, G., Kemp, D., Samnani, A.: Symantec Global Internet Security Threat Report. Trends for 2008. Technical report, Symantec (2009)
7. Forrest, S., Hofmeyr, S., Somayaji, A.: Computer Immunology. Communications of the ACM 40(10), 88–96 (1997)
8. Vlachos, V., Spinellis, D., Androutsellis-Theotokis, S.: Biological Aspects of Computer Virology. LNICST, vol. 26, pp. 209–219 (2010)
9. Li, J., Knickerbocker, P.: Functional Similarities Between Computer Worms and Bilogical Pathogens. Computers & Security 26, 338–347 (2007)
10. Geer, D.: Monoculture on the Back of the Envelope. Login 30(6), 6–8 (2005)
11. Goth, G.: Addressing the Monoculture. IEEE Security & Privacy 1(6), 8–10 (2003)

12. Geer, D., Bace, R., Gutmann, P., Metzger, P., Pfleeger, C.P., Quarterman, J.S., Schneier, B.: Cyber Insecurity: The Cost of Monopoly. Technical report, Computer & Communications Industry Association (2003)
13. Geer, D.: The Evolution of Security. ACM Queue, 31–35 (2007)
14. Somayaji, A., Hofmeyr, S., Forrest, S.: Principles of a Computer Immune System. In: Meeting on New Security Paradigms, September 23-26, pp. 75–82. ACM, Langdale (1997)
15. Anagnostakis, K., Greenwald, M., Ioannidis, S., Keromytis, A., Li, D.: A Cooperative Immunization System for an Untrusting Internet. In: Proceedings of the 11th IEEE International Conference on Networks (ICON), pp. 403–408 (2003)
16. Sidiroglou, S., Keromytis, A.: A Network Worm Vaccine Architecture. In: IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Workshop on Enterprise Security, Linz, Austria (2003)
17. de la Cuadra, F.: The Geneology of Malware. Network Security, 17–20 (2007)
18. Hayes, M., Walenstein, A., Lakhotia, A.: Evaluation of Malware Phylogeny Modelling Systems Using Automated Variant Generation. Journal in Computer Virology 5(4), 335–343 (2009)
19. Karim, M., Walenstein, A., Lakhotia, A., Parida, L.: Malware Phylogeny Using Permutations of Code. Journal in Computer Virology 1(1), 13–23 (2005)
20. Seewald, A.K.: Towards Automating Malware Classification and Characterization. In: Konferenzband der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik (German-Language Proceedings), Saarbrücken, pp. 291–302 (2008)
21. Gordon, S.: What is Wild? In: Proceedings of the 20th National Information Systems Security Conference (1997)
22. Bustamante, P.: The Disconnect Between the WildList and Reality. Technical report, PandaLabs (2007)
23. Marx, A., Dessman, F.: The WildList is Dead, Long Live the WildList! In: Virus Bulletin Conference, pp. 136–146 (2007)
24. The WildList Organization International: Wildlist, `http://www.wildlist.org/WildList/201001.htm`
25. Gordon, S.: Inside the Mind of Dark Avenger. In: Virus News International (1993)
26. Gordon, S.: Generic Virus Writer. In: 4th International Virus Bulletin Conference, Jersey, UK (1994)
27. Gordon, S.: Generic Virus Writer II. In: 6th International Virus Bulletin Conference, Brighton, UK (1996)
28. Gordon, S.: Understanding the adversary. IEEE Security & Privacy 4(5), 67–70 (2006)
29. Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G., Paxson, V., Savage, S.: Spamalytics: an empirical analysis of spam marketing conversion. Commun. ACM 52(9), 99–107 (2009)