Anti-virus Engine Analysis using Deep Web Malware Data

Igor Mishkovski, Miroslav Mirchev, Milos Jovanovik
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
Skopje, R. Macedonia
igor.mishkovski@finki.ukim.mk

ABSTRACT

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example.

AntiVirus products and tools are essential in every business deployment connected to the Internet. Nowadays, with the increase in the number and diversity of malware on the Web, there are also more AntiVirus Tools (AVT) becoming available to protect users and/or companies from malware. However, the quarterly growth at around 12\% for known unique malware samples, according to the *Intel Security Group's McAfee Labs Threat Report: August 2015*, and the fact that some AntiVirus companies use same or significantly similar AntiVirus engines leave us in some way vulnerable to the existing security threats.

In this work, using graph analysis and visualization methods, on one hand we will empirically infer detection engine similarity and existing groupings and/or overlapping between them, while on the other hand we will infer which Anti-Virus Tools (AVTs) differentiate from other AVTs and have greater advantage in detecting malware compared to others.

Using the AVT responses to our malware file set we will optimize the combination of AVTs in order to obtain maximum detection rate (i.e. coverage). We strongly believe that this approach can be used by companies who want to implement multi-scanning approach on their email gateways.

Finally, another novelty in this work is that we relate the source of the malware, i.e. the domain name where the malware is found, with AVTs. In this way, we will show the detection rate of AVTs across domains in which potential malware resides. The results will imply that certain AVTs have more detection capabilities on specific domains, whereas, others might have detection rate spread across multiple domains. All the analysis will be done on a malware file set provided by F-Secure and the AVTs responses on this file set obtained using the Virus Total API.

Based on the dataset we measure the similarity between different AVTs in order to see if there are some clusters or communities that share similar ``reaction`` to a certain malware files. Thus, we construct the *similarity network* $G^1 = (V, E, W^1)$ in order to characterize the similarity between different AVTs based on the shared files which they labeled them as malwares. The node set V consists of AVTs which were reported by Virus Total and the undirected edges set E contains the links between the AVTs that have labeled at least one common malicious file, with an edge weight w_{ij}^1 being defined through Jaccardi score of the sets of malware files detected by the two AVTs i and j. Thus, here we define the similarity between V_i and V_j as the co-occurrence strength. Let us assume that Fi and F_j denote set of files, labeled as malware by V_i and V_j, then we can define the Jaccardi similarity measure as a co-occurrence strength as follows.

$$sim(V_i, V_i) = \frac{|F_i \cap F_j|}{|F_i \cup F_i|} = w_{ij}^1 = w_{ji}^1, \ , \quad (1)$$

where |F| indicates the size of the set F. The value of wij1 is between 0 and 1 (where "0" indicates no co-occurrence relationship between two AVTs and "1" indicates a full co-occurrence).

The results show high similarity between certain AVT in their malware detection. Some of the AVT groups that show high similarity are i) **BitDefender, F-Secure, Emsisoft, MicroWorld-eScan** and **Ad-Aware**; ii) **Arcabit, eTrust-InoculateIT, UNA** and **T3**. This results clearly show that there might exist grouping in sense of structural communities and/or clusters between different AVTs. This kind of clustering or grouping might be as a consequence of the fact that different AVTs are specialized for certain type of malwares (Trojans, Adwares, Exploits, Rootkits, etc.), or malwares written for a given platform (such as Win32,

OSX, Android, etc.) or simply due to the fact that some companies use engines from other AV companies, such as *F-Secure* and *BitDefender*, *AVWare* and *VIPRE*.

Keywords— malware; community detection; anti-virus engines; data science; multi-scanning approach

ACKNOWLEDGMENT

Authors gratefully acknowledge the CyberTrust research project and F-Secure for their support. I.M. work was partially financed by the Faculty of Computer Science and Engineering at the University 'Ss. Cyril and Methodius' as part of the project "*AVADEEP: Anti-virus analysis using Deep Web malware files*".

REFERENCES

[1] M. Lindorfer, M. Neugschwandtner, L. Weichselbaum, Y. Fratantonio, V. v. d. Veen, and C. Platzer, "Andrubis - 1,000,000 apps later: A view on current android malware behaviors," in Proceedings of the Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BAD-GERS), 2014, pp. 3-17.

[2] M. K. Bergman, "White paper: the deep web: surfacing hidden value," Journal of electronic publishing, vol. 7, no. 1, 2001.

[3] A. Mohaisen and O. Alrawi, "Av-meter: An evaluation of antivirus scans and labels," in Proceedings of the 11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, ser. DIMVA '14. Springer International Publishing, 2014, pp. 112-131.

[4] "VirusTotal: Free service to analyze suspicious files and URLs," https://www.virustotal.com/en/, online; accessed 14 July 2016.

[5] I. Gashi, V. Stankovic, C. Leita, and O. Thonnard, "An experimental study of diversity with off -the-shelf antiVirus engines," in Proceedings of the 8th IEEE International Symposium on Network Computing and Applications, 2009.

[6] I. Gashi, B. Sobesto, V. Stankovic, and M. Cukier, "Does malware detection improve with diverse antivirus products? an empirical study," in Proceedings of the 32nd International Conference on Computer Safety, Reliability, and Security, ser. SAFECOMP '13. Springer Berlin Heidelberg, 2013, pp. 94-105.

[7] J. Canto, M. Dacier, E. Kirda, and C. Leita, "Large scale malware collection: lessons learned," in Proceedings of the 27th International Symposium on Reliable Distributed Systems, ser. SRDS '08, 2008.

[8] Q. Jerome, K. Allix, R. State, and T. Engel, "Using opcode-sequences to detect malicious android applications," in 2014 IEEE International Conference on Communications (ICC), 2014, pp. 914-919.

[9] M. Zheng, P. P. Lee, and J. C. Lui, "Adam: an automatic and extensible platform to stress test android anti-virus systems," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, 2012, pp. 82-101.

[10] F. Maggi, A. Bellini, G. Salvaneschi, and S. Zanero, "Finding non-trivial malware naming inconsistencies," in Proceedings of the 7th International Conference on Information Systems Security, ser. ICISS '11. Springer Berlin Heidelberg, 2011, pp. 144-159.

[11] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. D. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, ser. AISec '15. ACM, 2015, pp. 45-56.

[12] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM, 2015, pp. 45-56.

[13] J. Chang, K. K. Venkatasubramanian, A. G. West, and I. Lee, "Analyzing and defending against web-based malware," ACM Computing Surveys (CSUR), vol. 45, no. 4, p. 49, 2013.

[14] H. S. S.L., "Virustotal public api."

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.

[16] H. T. T. Truong, E. Lagerspetz, P. Nurmi, A. J. Oliner, S. Tarkoma, N. Asokan, and S. Bhattacharya, "The company you keep: Mobile malware infection rates and inexpensive risk indicators," in Proceedings of the 23rd International Conference on World Wide Web, ser. WWW '14. ACM, 2014, pp. 39-50.