Identifying Malware Genera using the Jensen-Shannon Distance Between System Call Traces

Jeremy D. Seideman
The Graduate School and University Center
City University of New York
New York, USA
Email: jseideman@gc.cuny.edu

Bilal Khan
Dept. of Math & Comp. Science
John Jay College, CUNY
New York, USA
Email: bkhan@jjay.cuny.edu

Antonio Cesar Vargas
NacoLabs Consulting, LLC
New York, USA
Email: cesar@nacolabs.com

Abstract

The study of malware often involves some form of grouping or clustering in order to indicate malware samples that are closely related. There are many ways that this can be performed, depending on the type of data that is recorded to represent the malware and the eventual goal of the grouping. While the concept of a malware family has been explored in depth, we introduce the concept of the ***malware genus****, a grouping of malware that consists of very closely related samples determined by the relationships between samples within the malware population. Determining the boundaries of the malware genus is dependent upon the way that the malware samples are compared and the overall relationship between samples, with special attention paid to the parent-child relationship. Biologists have several criteria that are used to judge the usefulness of a genus when creating a taxonomy of organisms; we sought to design a classification that would be as useful in the world of malware research as it is in biology. We present two case studies in which we analyze a set of malware, using the Jensen-Shannon Distance between system call traces to measure distance between samples. The case studies show the genera that we create adhere to all of the criteria used when creating taxa of biological organisms.*

1 Introduction

Classification methods are extremely useful when studying malware. By finding ways to classify and group different malware samples, it is possible to compare malware in a useful way, such as by limiting research to groups that exhibit certain properties or features. Classifying malware samples also aids in malware defense, removal and analysis as similar malware may have similar weaknesses, structures and effects, all of which may be indicated through better detection methods [14, 23].

The classification of malware has been the subject of many research papers, such as [1, 2, 9, 14, 22], all of which approach the problem in a slightly different manner. The obvious problem with approaching classification in a different manner is that each study determines its own classification scheme, which leads to inconsistency among classification – a cluster or family determined by one study might not agree with those determined by another. That problem is not isolated to research studies; the anti-virus industry previously dealt with the same issue, leading to the 1991 New Virus Naming Convention (also called the CARO Convention). The CARO Convention attempted to standardize virus and malware names, based on family, group, and variant names [21]. This approach was successful in a limited fashion, as many vendors still use their own naming schemes.

All classification, however, is based on the assumption that there are common elements among malware samples. As the rate of malware release has been reported to be as high as 82,000 new malware samples per day [19], it follows that this would be the case. In the world of biology, it is extremely common to group organisms together based on "morphology, physiology, ecology and genetics" [4]. These groupings are called **taxa**, indicating a grouping at one of several levels of organization [3, p. 471]. The most common way that many organisms are identified is the **bino-**

978-1-4799-7329-3/14/$31.00 ©2014 IEEE

the correlation between the genera that we identify and malware families as identified by other clustering and classification methods. While our classification method can be used alone as part of a larger malware analysis, identifying genera that adhere to more restrictive conditions can also aid in evaluation of malware grouping methods and assist in determining shared characteristics that occur in multiple related malware samples. The author of [15] presented a generalized form of the Jensen-Shannon Divergence that could be applied to an entire genus or population in order to determine the degree to which the samples relate to one another.

References

[1] M. Bailey, J. Oberheide, J. Andersen, Z. M. Mao, F. Jahanian, and J. Nazario. Automated classification and analysis of internet malware. In *Proceedings of RAID 2007*, pages 178–197, 2007. http://dx.doi.org/10.1007/978-3-540-74320-0_10.

[2] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Krügel, and E. Kirda. Scalable, behavior-based malware clustering. In *Proceedings of NDSS 2009*, 2009. http://www.isoc.org/isoc/conferences/ndss/09/pdf/11.pdf.

[3] N. A. Campbell. *Biology*. The Benjamin/Cummings Publishing Company, Inc., New York, 4th edition edition, 1996.

[4] Classification of Species, 2009. http://classes.entom.wsu.edu/348/classification.htm.

[5] M. Elhadad. NLP09 Assignment 1: Computing KL Divergence, 2013. http://www.cs.bgu.ac.il/~elhadad/nlp09/KL.html.

[6] D. Endres and J. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, July 2003.

[7] D. Gao, M. K. Reiter, and D. X. Song. Behavioral distance for intrusion detection. In *Proceedings of RAID 2005*, pages 63–81, 2005. http://dx.doi.org/10.1007/11663812_4.

[8] Genus - definition from biology-online.org, 2014. http://www.biology-online.org/dictionary/Genus.

[9] M. Gheorghescu. An automated virus classification system. *Virus Bulletin Conference*, pages 294–300, Oct 2005.

[10] G. Jacob, H. Debar, and E. Filiol. Behavioral detection of malware: from a survey towards an established taxonomy. *Journal in Computer Virology*, 4(3):251–266, 2008.

[11] M. E. Karim, A. Walenstein, A. Lakhotia, and L. Parida. Malware phylogeny generation using permutations of code. *Journal in Computer Virology*, 1(1-2):13–23, 2005.

[12] J. Kinable and O. Kostakis. Malware classification based on call graph clustering. *Journal in computer virology*, 7(4):233–245, 2011.

[13] J. Z. Kolter and M. A. Maloof. Learning to detect and classify malicious executables in the wild. *Journal of Machine Learning Research*, 7:2721–2744, 2006. http://www.jmlr.org/papers/v7/kolter06a.html.

[14] T. Lee and J. J. Mody. Behavioral classification. In *Proceedings of EICAR 2006*, May 2006. http://secureitalliance.org/blogs/files/73/1244/Behavioral_Classification.doc.

[15] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, Jan 1991.

[16] D. Mutz, F. Valeur, G. Vigna, and C. Kruegel. Anomalous system call detection. *ACM Trans. Inf. Syst. Secur.*, 9(1):61–93, 2006.

[17] Norman Sandbox, 2009. http://www.norman.com/technology/norman_sandbox/.

[18] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[19] Annual Report Panda Labs – 2013 Summary, 2013. http://press.pandasecurity.com/wp-content/uploads/2010/05/PandaLabs-Annual-Report_2013.pdf.

[20] Y. Park, D. Reeves, V. Mulukutla, and B. Sundaravel. Fast malware classification by automated behavioral graph matching. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, page 45. ACM, 2010.

[21] C. Riau. A virus by any other name: Virus naming practices, Jun 2002. http://www.securityfocus.com/print/infocus/1587.

[22] A. K. Seewald. Towards autmating malware classification and characterization. In *Proceedings of Sicherheit 2008*, pages 291–302, 2008. http://alex.seewald.at/files/2008-01.pdf.

[23] J. Seideman. Recent advances in malware detection and classification: A survey. Technical report, The Graduate School and University Center of the City University of New York, 2009.

[24] Threat explorer - spyware and adware, dialers, hack tools, hoaxes and other risks, 2012. http://www.symantec.com/security_response/threatexplorer/.

[25] VirusTotal, 2008. http://www.virustotal.com.

[26] VX heavens, 2010. http://vx.netlux.org.