Diversity with AntiVirus Products: Additional Empirical Studies

Ilir Gashi, Vladimir Stankovic
Centre for Software Reliability
City University London
London, United Kingdom
{i.gashi, v.stankovic}@csr.city.ac.uk

Michel Cukier, Bertrand Sobesto
Clark School of Engineering
University of Maryland
College Park, Maryland, USA
{bsobesto, mcukier}@umd.edu

Abstract— **In this paper we describe the design of a new set of empirical studies we will run to test the gains in detection capabilities from using diverse AntiVirus products. This new work builds on previous work on this topic reported in [1, 2, 3]. We describe the motivation for this work, how it extends the previous work and what studies we will conduct.**

Keywords - security assessment; anti-virus products; honeypots and honeynets; fault tolerance; instrusion tolerance; software diversity.

I. BACKGROUND

All systems, including those built from off-the-shelf components, need to be sufficiently reliable and secure in delivering the service that is required of them. Various ways in which this reliability and security can be achieved in practice ranges from the use of various validation and verification techniques to the use of software fault/intrusion tolerance techniques. Fault tolerance techniques range from simple "wrappers" of the software components [4] to the use of diverse software products in a fault-tolerant system [5]. This latter strategy of implementing fault tolerance was historically considered prohibitively expensive, due to the need for developing multiple bespoke software versions. However, the wide proliferation of off-the-shelf software for various applications has made the use of software diversity an affordable option for fault tolerance against either malicious or non-malicious faults.

A recent publication [1] has detailed an implementation of an AntiVirus (AV) platform that makes use of diverse AV products for malware detection. A similar architecture that uses diverse AV email scanners has been commercially available for several years [6]. Hence architectural solutions for employing diverse detection engines with AV products are already known and in some cases commercially deployed. Studies that provide empirical evaluation of the effectiveness of diversity for detection of malware are, on the other hand, much more scarce.

The following claim is made on the VirusTotal site [7]: "*Currently, there is no solution that offers 100% effectiveness in detecting viruses, malware and malicious URLs*". Given these limitations of individual AV engines, designers of security protection systems are interested in at least getting estimates of the possible gains in terms of added security that the use of diversity (e.g. diverse AV products) may bring for their systems.

Two authors of this paper have reported previously [2], [3] (together with other colleagues from CSR, City University London, Symantec Research and Institute Eurecom) results which analysed the detection capabilities of different AV products [1] and potential improvements in detection that can be observed from using diverse AV products. We observed that some AV products achieved high detection rates, but none detected all the malware samples in our study. We also found many cases of *regression* in the detection capability of the AV products: cases where an AV would regress from detecting the malware on a given date to not detecting the same malware at a later date(s). We saw significant improvements in the detection capability when using two or more diverse AV products. For example, even though no single AV product detected all the malware in our study, almost 25% of all the diverse 1-out-of-2 pairs[2] of AV products, and over 50% of all diverse 1-out-of-3 triplets of AV products successfully detected all the malware. We also observed significant potential gains in reducing the "at risk time" of a system from employing diverse AVs: even in the cases where AVs fail to detect a malware, there is diversity in the time it takes different vendors to successfully define a signature for the malware and detect it.

The results were intriguing. They concern, however, a specific snapshot in the detection capabilities of AV products against malware threats prevalent in that time period: the analysis was based on 1599 malware samples collected from a distributed honeypot deployment over a period of 178 days from February to August 2008.

Additionally, our dataset contains only confirmed malware. Hence we could measure failure of AV products to detect genuine malware (*false negatives*), but we could not measure cases where benign files are incorrectly identified as malware (*false positives*). False positive rate is an important measure when evaluating the effectiveness of any detection system, including AV products.

II. ADDITIONAL EMPIRICAL STUDIES TO TEST GAINS FROM DIVERSITY WITH DIVERSE ANTI-VIRUS PRODUCTS

With these limitations in mind, we are designing new studies collecting new data and performing further analysis

[1] For the sake of brevity, in the rest of the paper we will use the short-hand notation AV, or AV product to refer to the signature-based component of an AntiVirus detection engine.

[2] A failure of a 1-out-of-2 pair occurs when both of the constituent AVs fail.

of the potential benefits of diversity with AV products. The new work will also allow us to compare the new findings with the previous study and hence provide another viewpoint on the benefit, or otherwise, of using diverse AV products for malware detection.

We will use the network of honeypots deployed at the University of Maryland. Honeypots are virtual or physical hosts used for the sole purpose of collecting malicious activity. Thus, they have no production value and the network traffic they received is either due to misconfiguration or attacks. For two of the authors of this paper, the experience working with honeypots has been a multi-year learning process during which we incrementally improved the way we conducted honeypot-based empirical studies. The collected data were used in various research conducted (i.e., [8, 9]). The farm of hundreds of honeypots consists of low interaction honeypots (deployed at several organizations in the USA, Europe and Africa) and high interaction honeypots which goal is to characterize attacks and attackers. The data we will use will be collected on the low interaction honeypots that run dionaea [10]. dionaea was specifically developed to catch the complete attack payload, i.e., the malware. The malware collected will be sent daily to the VirusTotal [7] web service which allows the analysis by the signature-based engines of up to 37 different AV vendors. To test the evolution of the detection capability of the AV products we again plan to collect and send the malware over an extended period of time. In [2, 3] we did this for a period of up to 30 days. In the new experiments, in order to get more extensive and accurate measures of the "at risk time" (namely, the periods when no ("correct") signature exists for the detection of a given malware by a given AV product) we will continuously run the study over a period of several months (i.e., at least three months).

We will also collect data to assess the rate of false positives of AV products. However, we must stress that the exact experimental method we will use to obtain the data which would allow for this assessment is currently work in progress. We can, of course, just artificially create a large set of non-malicious files which we could then send to the AV products and obtain respective false positive rates. The representativeness of this kind of experiment is, however, difficult to justify: the non-malicious files chosen in this way are unlikely to be representative samples of the ones that AV products may be expected to inspect in normal operation. Choosing representative test loads and defining what constitutes a false positive is a matter of some debate in the AV community (see discussion in [11]). However, obtaining data on false positives remains high on our priority list to enable a more comprehensive assessment of the AV products detection capabilities and the possible benefits of diversity. We will study the confidentiality and data privacy issues with obtaining data on files that AV viruses inspect during normal operation in a small business or university campus. This will at least give us representative sets of both malicious and non-malicious files that an AV product would be expected to inspect in these environments. This data would complement the dataset on malicious files that we would collect from the honeypots, as explained above.

We also plan to research the architectural aspects of using diverse AntiVirus products and the performance overheads that this may bring, building on work that has been reported by other researchers [1] and commercial vendors [6].

Finally, we plan to do more extensive exploratory modeling, as well as modeling for prediction. In previous work [3], we observed that an empirically derived hyper-exponential model proved to be a remarkably good fit to the proportion of systems in each diverse setup that had a zero failure rate. The new dataset will allow us to test if the hyper-exponential distribution is more likely to be generic. If it is, it would be a useful means for predicting the expected detection rates for a system with a high degree of diversity (i.e., a high number of diverse AV products) based on measurements made with simpler diverse configurations (say, with 2 or 3 diverse AV products). We also plan to study to what extent these models can be extended to incorporate the time dimension (the "at risk time" from lack of signature definitions for a given product). This would allow a decision maker to more optimal trade-offs between detection rates and the "at risk time" when selecting diverse AV products for their configuration.

ACKNOWLEDGMENT

The initial research [2] on these results was done in collaboration with Corrado Leita and Olivier Thonnard, who are now with Symantec Research, and the work was supported by the European Union FP 6 via the "Resilience for Survivability in Information Society Technologies" (ReSIST) Network of Excellence (contract IST-4-026764-NOE), FP 7 via the project FP7-ICT-216026-WOMBAT, and a grant by City University Strategic Development Fund.

REFERENCES

[1] Oberheide, J., et. al. "CloudAV: N-Version Antivirus in the Network Cloud", in Proc. USENIX Security Symp., p. 91–106, 2008.

[2] Gashi, I., et al. "An Experimental Study of Diversity with Off-the-Shelf AntiVirus Engines", in Proc. IEEE NCA'09, p. 4-11, 2009.

[3] Bishop, P., et. al. "Diversity for Security: a Study with Off-The-Shelf AntiVirus Engines", in Proc. IEEE ISSRE'11, 2011, pp: 11-19.

[4] van der Meulen, M.J.P., et al. "Protective Wrapping of Off-the-Shelf Components". in Proc. ICCBSS'05, p. 168-177, 2005

[5] Strigini, L., "Fault Tolerance Against Design Faults", in Dependable Computing Systems: Paradigms, Performance Issues, and Applications, H. Diab and A. Zomaya, Editors, J. Wiley & Sons. p. 213-241, 2005

[6] GFi, "GFiMailDefence Suite", last checked 2012: http://www.gfi.com/maildefense/

[7] VirusTotal, "VirusTotal - A Service for Analysing Suspicious Files", last checked 2012: http://www.virustotal.com/sobre.html.

[8] G. Salles-Loustau et. al. "Characterizing Attackers and Attacks: An Empirical Study", in Proc. IEEE PRDC'11, 2011.

[9] Berthier, R., et. al. "Analyzing the Process of Installing Rogue Software", in Proc. IEEE DSN'09, 2009

[10] Dionea, "Dionea", last checked 2012: http://dionaea.carnivore.it/

[11] Leita, C., et al. "Large Scale Malware Collection: Lessons Learned". in Workshop on Sharing Field Data and Experiment Measurements on Resilience of Distributed Computing Systems, with Symp. on Reliable Distributed Systems (SRDS), 2008