Poster: Benchmarking Computer Security Using WINE

Tudor Dumitras, Darren Shou, Corrado Leita
Symantec Research Labs
Email: {tudor_dumitras,darren_shou,corrado_leita}@symantec.com

1. Challenges for security benchmarking

The data sets available today are often insufficient for conducting scientifically rigorous experimentation in computer security. This is the result of a number of scientific, ethical and legal challenges, which must be considered when constructing a benchmark for computer security:

- **Field data.** Some benchmarking results in the past have addressed privacy concerns by generating *synthetic data*, based on the observed statistical distributions of raw data samples collected [1]. However, because the cyberthreat landscape changes frequently, it is difficult to relate such benchmarking results with the real-world performance of the system-under-test [2]. We need to build upon real-world field observations and consequently deal with the associated ethic and legal concerns.
- **Reproducibility.** Many security data sets have been mentioned in a single publication and then forgotten. Furthermore, researchers often do not have access to the data used in experiments conducted in other groups, which represents an obstacle for the independent verification of results, one of the basic tenets of scientific research. The data sets used in such experiments must be archived for future reference, and they should be made available to the research projects attempting quantitative comparisons against the prior art.
- **Representativeness.** Due to the dynamic nature and the complexity of the threat landscape, it is expensive to build and maintain a sufficiently large and detailed snapshot that is representative of the current trends. Many research contributions are consequently proposed and validated on inappropriately-sized or biased data sets, and this can cast doubts on the real-world applicability of the proposed ideas and on the evaluation of their performance.
- **Ethics.** A benchmark for computer security must necessarily include sensitive code and data. These artifacts could damage computer systems or could reveal personally identifiable information about the users affected by the cyber attacks recorded in the data set. For instance, the IP address of hosts initiating network-based attacks implicitly leaks information about the security state of individuals and organizations. Similarly, binary samples of malware must not be made available on the Internet to avoid the generation (even accidental) of infections.
- **Legal.** Compliance with privacy laws often restricts the data collection, storage and exchange. We have previously expressed the need to build upon field data generated from real networks and users. This has multiple implications with respect to laws that limit access to the traffic or that address the storage of this information.

These conflicting challenges increase the cost for creating and maintaining scientifically-sound security benchmarks, and they also raise important barriers for sharing this information among research groups. Consequently, it is common practice for research groups to maintain private benchmarks (e.g. malware collections) of different size and nature, with little attention to the challenges identified and reported in the past.

2. The WINE framework

We aim to fill these gaps by (i) making representative field data, which covers the entire lifecycle of malware, available to the research community, and (ii) developing a platform for repeatable experimentation around these data sets. Central to our approach is the Worldwide Intelligence Network Environment (WINE) [3], Symantec's program to share its data with the research community[1].

[1] More information on accessing the WINE data is available at http://www.symantec.com/WINE

Data set	Sources	Description
Binary reputation	3.5M machines	Information on unknown binaries-i.e., files for which an A/V signature has not yet been created-that are downloaded by users who opt in for Symantecs reputation-based security program. This data can indicate for how long a particular threat has existed in the wild before it was first detected.
A/V telemetry	90M machines	Records occurrences of known threats, for which Symantec has created signatures and which can be detected by antivirus products. This data set includes intrusion detection telemetry
Email spam	2.5M decoy accounts	Samples of phishing and spam emails, collected by Symantecs enterprise-grade systems for spam filtering. This data set includes samples of email spam and statistics on the messages blocked by the spam filters.
URL reputation	10M domains	Website-reputation data, collected by crawling the web and by analyzing malicious URLs (a simplified interface for querying this data is available at http://safeweb.norton.com/).
Malware samples	200 countries	A collection of both packed and unpacked malware samples (viruses, worms, bots, etc.), used for creating Symantecs A/V signatures.

Table 1. WINE data sets

WINE aims to create a rigorous benchmark for computer security, based on comprehensive field data collected by Symantec from 240,000 sensors worldwide. Instead of developing *micro-benchmarks*, which systematically test all the features of a security tool, WINE will provide a *macro-benchmark* that is representative for the real-world workloads of these tools.

We are currently developing a data storage and analysis platform, which aims to ensure *experimental repeatability* by archiving snashots of the data used in each experiment and by providing researchers tools for recording all the information required for reproducing the results. To protect the sensitive information included in the data set and to ensure the reproducibility of experimental results, all the experiments and empirical studies will be conducted on the WINE platform hosted by Symantec Research Labs at our Culver City, CA or Herndon, VA locations. A snapshot of the data requested will be frozen, for future reference, and all the analysis and experimentation will be conducted on the platform. Researchers will have access to the raw data collected, and they will retain all right, title and interest to the research results.

WINE will provide access to a large collection of malware samples, and to the contextual information needed to understand how malware spreads and conceals its presence, how it gains access to different systems, what actions it performs once it is in control and how it is ultimately defeated. WINE will initially include five data sets, summarized in Table 1. These data sets enable two research directions: (i) empirical studies for *understanding each phase in the lifecycle of cyberattacks*, and (ii) quantitative evaluations and comparisons of attack prevention or detection techniques, for *benchmarking security systems*.

The selection of the initial data sets for WINE was guided by our goal to establish a benchmark for computer security and by the needs expressed in the security community [4]. However, the access to the WINE data is not restricted to security researchers. WINE aims to aggregate the data feeds collected by Symantec in order to enable experimental research across a broad spectrum of disciplines, e.g., dependability, machine learning, software engineering, networking, economics, visual analytics.

We believe WINE to be an important contact point between the worlds of academic and industrial research. By presenting the ongoing work on the framework in the context of the S&P symposium, we hope to engage discussions with some of the best researchers in the field, both to receive constructive feedback on the initiative and to engage in new collaborations that will lead to future advances to the state of the art to security and privacy research.

References

[1] R. Lippmann *et al.*, “Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation,” in *DARPA Information Survivability Conference and Exposition, 2000. DISCEX’00. Proceedings*, vol. 2. IEEE, 2000, pp. 12–26.

[2] J. McHugh, “Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory,” *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 262–294, 2000.

[3] T. Dumitras and D. Shou, “Toward a standard benchmark for computer security research,” in *Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS workshop)*, 2011.

[4] J. Camp *et al.*, “Data for cyber-security research: Process and “wish list”, http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf.”