Field Data Available at Symantec Research Labs: The Worldwide Intelligence Network Environment (WINE)

Invited Paper

Tudor Dumitraș
Symantec Research Labs
tudor_dumitras@symantec.com

Abstract

The data sets available today are often insufficient for conducting representative experiments or rigorous empirical research. The Worldwide Intelligence Network Environment (WINE) aims to fill this gap by providing access to sampled data feeds, which are used internally at Symantec Research Labs, and by promoting rigorous experimental methods. WINE allows researchers to define reference data sets, for validating new techniques or for conducting empirical studies, and provides the metadata needed for understanding the results. WINE archives these reference data sets in order to facilitate repeatable experiments and to enable meaningful comparisons against the prior art. Moreover, the field data included in WINE will likely provide key insights across a broad spectrum of disciplines, such as software reliability, computer security, machine learning, networking, economics, or visual analytics.

1. Introduction

We have not yet been able to follow a software defect over the course of its entire life—from a programming bug that evades testing and introduces a latent vulnerability, through the stealth exploitation of this vulnerability in zero-day attacks, its discovery and description in a public advisory, the release of anti-virus signatures and of patches for the vulnerability, the automatic generation of exploits based on the patch, and to the final race between these attacks and the remediation measures introduced by the security community. WINE aims to fill these gaps with data that covers the entire *lifecycle of software defects* and that is representative of the impact that atomicity violations, data races, *etc.* have in the real world.

Symantec has established some of the most comprehensive sources of Internet threat data in the world. More than 240,000 sensors in over 200 countries monitor attack activity through a combination of Symantec products and services such as Symantec DeepSight Threat Management System, Symantec Managed Security Services and Norton consumer products, as well as additional third-party data sources. Symantec also gathers malicious code intelligence from more than 130 million client, server, and gateway systems that have deployed its antivirus products. Additionally, Symantec's distributed honeypot network collects data from around the globe, capturing previously unseen threats and attacks and providing valuable insight into attacker methods. Spam and phishing data is captured through a variety of sources including the Symantec Probe Network, a system of more than 2.5 million decoy accounts; MessageLabs Intelligence, a respected source of data and analysis for messaging security issues, trends and statistics; and other Symantec technologies. Data is collected in more than 86 countries. Over 8 billion email messages and over 1 billion Web requests are processed per day across 16 major data centers. These resources give Symantec's analysts unparalleled sources of data with which to identify, analyze, and provide informed commentary on emerging trends in attacks, malicious code activity, phishing, and spam.

Operational model. To protect the sensitive information included in the datasets (*e.g.* hosts that have been compromised), we require researchers to access the WINE system at one of our Culver City, CA or Herndon, VA locations. Researchers will have access to the raw data collected. Symantec Research Labs will accept proposals that briefly explain the research question investigated and the data needed for conducting the research. A snapshot of the data requested will be archived, for future reference, and all the analysis and experimentation will be conducted on the infrastructure provided by Symantec Research Labs. The researchers will retain all right, title and interest to the research results.

More information on accessing WINE is available at `http://www.symantec.com/WINE`.

2. Examples of data included in WINE

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. However, the access to WINE is not restricted to researchers in the field of computer security. We believe that the WINE data is interesting from other perspectives as well, *e.g.*, for gaining a deeper understanding of software reliability or as a test case for machine learning techniques. Moreover, we aim to aggregate the data feeds collected at Symantec in order to enable broad experimental research.

For example, the WINE data sets include:[1]

- *Binary reputation*: 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 Symantec's 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*: Records occurrences of known threats, for which Symantec has created signatures and which are detected by our anti-virus products. This data set includes intrusion-detection telemetry.
- *Email spam*: Samples of phishing and spam emails, collected by Symantec's enterprise-grade systems for filtering spam.
- *URL reputation*: 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*: A collection of both packed and unpacked malware samples (viruses, worms, bots, etc.), used for creating Symantec's A/V signatures.

3. Scientific impact

The WINE data provides the opportunity to investigate a number of emerging research topics. While we expect that the most compelling contributions will result from the use of WINE by external researchers, we provide a few examples of topics, from a domain that we are familiar with.

Measuring software reliability. Software reliability is difficult to assess because, unlike performance, it cannot be measured directly. Reliability estimations must be based on a large sample of empirical observations, taken from multiple perspectives, in order to ensure that the results are representative of real-world system behaviors. For example, the introduction of concurrency bugs (*e.g.*, race conditions, atomicity violations) was studied by analyzing the bug reports and revision logs of large open-source projects [2]. However, these findings do not *discern the programming bugs that yield exploited vulnerabilities and that help malware propagate* in the wild, which emphasizes a fundamental shortcoming in our assessment of software quality. By correlating data from open-source software repositories with the historical information provided by WINE, we have the opportunity to gain a deeper understanding of security vulnerabilities. This will allow us to minimize the impact of vulnerabilities by focusing on the programming bugs that matter.

Benchmarking computer security. A benchmark for computer security must necessarily include sensitive code and data. For example, the IP addresses of hosts initiating network-based attacks could point to personal computers that have been infected with malware, while binary samples of malware cannot be made freely available on the Internet. Because these artifacts could damage computer systems or reveal personally identifiable information about the users affected by cyber attacks, publicly disseminating such a benchmark raises scientific and ethical challenges. Building on lessons learned from benchmarks in operating systems or computer architecture, some efforts in the past have addressed privacy concerns by generating synthetic input traces from statistical distributions of the raw data. However, because the cyberthreat landscape changes frequently, the results of such benchmarking techniques are difficult to relate to the real-world performance of the systems-under-test [3]. We tackle the key challenges for security benchmarking by monitoring the access to the raw data in WINE, by establishing a predictable process for data collection and by including the *metadata* required for distinguishing meaningful conclusions from artifacts. WINE provides many of the data sets currently needed in the security research community [1], and it also includes unique data sets that have not been discussed before (*e.g.*, historical information on malicious executables extending *before* the threat identification).

Building a platform for repeatable experimentation. For WINE, the combined size of ingress data feeds is approximately 0.5 PB. Unlike other architectures for data-intensive computing, such as MapReduce or parallel databases, our experimental platform has a key requirement of ensuring the *reproducibility and comparability* of results. In addition to archiving snapshots of the data sets used in each experiment, the platform that we are building for WINE aims to provide integrated tools to help researchers record their hypotheses and experimental procedures. This will allow us to maintain a *lab book*, which is a common practice in other experimental fields (*e.g.*, applied physics, cell biology) and is essential for ensuring the reproducibility of the results.

Scaling machine learning algorithms. The WINE data sets were shown to be amenable to machine learning techniques such as belief propagation, support vector machines or temporal sequence matching. For example, the binary-reputation data is well suited for the evaluation of parallel algorithms operating on very large graphs.

References

[1] J. Camp et al. Data for cybersecurity research: Process and "wish list". `http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf`, Jun 2009.

[2] S. Lu et al. Learning from mistakes: A comprehensive study on real world concurrency bug characteristics. In *ASPLOS*, pages 329–339, Seattle, WA, USA, 2008.

[3] 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*, 3(4):262–294, 2000.

[1] This does not represent an exhaustive catalog of the WINE data.