Online Appendix to: Evaluating Computer Intrusion Detection Systems: A Survey of Common Practices

ALEKSANDAR MILENKOSKI, University of Würzburg
MARCO VIEIRA, University of Coimbra
SAMUEL KOUNEV, University of Würzburg
ALBERTO AVRITZER, Siemens Corporation, Corporate Technology
BRYAN D. PAYNE, Netflix, Inc.

A. EVALUATION OF INTRUSION DETECTION SYSTEMS: HISTORICAL OVERVIEW

In Figure 11, we depict chronologically ordered dates that mark major developments in the area of intrusion detection system (IDS) evaluation from its inception until the present date.

The earliest effort on evaluating IDSes in a systematic manner is the work of Puketza et al. [1996, 1997]. These authors presented an approach for evaluating IDSes based on principles of the field of software systems testing. They were the first to develop a framework for evaluating IDSes, which they describe in detail in their work from 1997. They used the framework to evaluate a network-based IDS in terms of attack detection accuracy, resource consumption, and performance under stress.

The years of 1998, 1999, and 2000 mark a major accomplishment in the area of IDS evaluation. The Lincoln Laboratory at the Massachusetts Institute of Technology, sponsored by the Defense Advanced Research Projects Agency (DARPA), evaluated multiple IDSes using generated trace files that contain host and network activities of benign and malicious nature. The latter are commonly known as the DARPA datasets (see Section 2.1). Cunningham et al. [1999] describe the approach taken to generate the DARPA datasets in detail. The DARPA datasets are still extensively used in IDS evaluation studies.

Debar et al. [1998] from the IBM Zurich Research Laboratory developed a workbench for evaluating IDSes. The workbench enabled the execution of attack scripts stored in a database maintained internally at IBM and the generation of regular workloads for training anomaly based IDSes. Debar et al. demonstrated the use of the workbench by evaluating multiple host-based IDSes.

A recent effort to support the rigorous evaluation of IDSes is being driven by Symantec. Dumitras and Shou [2011] presented Symantec's Worldwide Intelligence Network Environment (WINE) datasets,[45] which contain local and remote attacks (see Table I). They also presented an evaluation platform that makes use of the datasets and is available for use by researchers for evaluating security mechanisms. However, since the datasets are captured from real network infrastructures and systems, and therefore contain private user data, they can only be accessed on-site at Symantec to avoid legal issues. The large scale of this project is indicated by the fact that Symantec continuously monitors and records malicious activities using more than 240,000 sensors deployed in 200 countries.

In addition to attacks, which can be used for evaluating IDSes, the WINE datasets contain samples of malware (i.e., malicious software like trojans or viruses), which

[45] http://www.symantec.com/about/profile/universityresearch/sharing.jsp.

DOI: http://dx.doi.org/10.1145/2808691

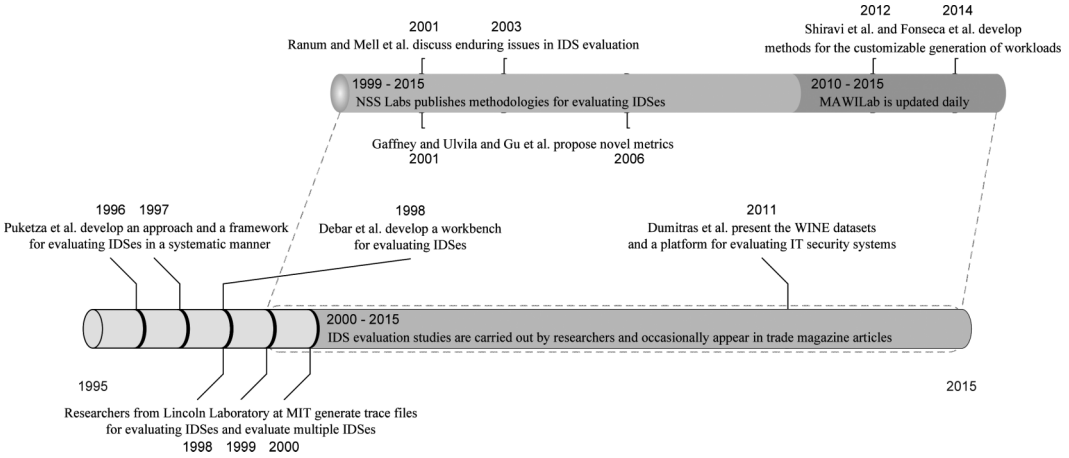

Fig. 11. Timelime showing dates that mark major developments in the area of IDS evaluation.

can be used for evaluating malware detection systems (e.g., antivirus systems). In contrast to IDSes, which are designed to detect ongoing attacks (see Section 1.1), malware detection systems are designed to detect malware running on a given host, whose installation normally takes place after an intrusion (i.e., a successful attack) has occurred. Evaluation of malware detection systems is outside the scope of this work.

There have been many small-scale IDS evaluation efforts between 2000 and today. Articles reviewing and comparing IDSes occasionally appear in trade magazines, such as the IDS evaluation study presented in the *SC* magazine in 2011.[46] Following the rising interest of researchers in intrusion detection since 2000, many IDS evaluation studies have been presented as part of publications proposing novel intrusion detection techniques or IDS evaluation methods.

Several works published between 2000 and today have had long-term impact on the IDS evaluation area: Ranum [2001] and Mell et al. [2003] proposed approaches and gave recommendations toward addressing enduring issues in IDS evaluation (e.g., the use of faulty or unrepresentative workloads, inaccurate interpretation of results from IDS evaluation studies); Gaffney and Ulvila [2001] and Gu et al. [2006] were the first to propose metrics for quantifying IDS attack detection accuracy that use specific measurement methods to address issues in using the conventional metrics at the time, such as the receiver operating characteristic curve (see Section 2.2); and focusing on the issue of using unrepresentative workloads, Shiravi et al. [2012] and Fonseca et al. [2014] developed methods for the customizable generation of IDS evaluation workloads that closely resemble real-world workloads at the time they are generated (see Section 2.1).

In 2010, the Measurement and Analysis on the WIDE Internet (MAWI) Working Group of the Widely Integrated Distributed Environment (WIDE) project announced MAWILab, a repository of publicly available traces intended for use in IDS evaluation studies [Fontugne et al. 2010].[47] This is a significant effort to enable the representative evaluation of modern network-based IDSes. The trace files in MAWILab contain network traffic captured from a trans-Pacific 150Mbps link between Japan and the United States. They contain regular network traffic as well as attacks, which are labeled before

[46]http://www.scmagazine.com/idsips/grouptest/241/#.
[47]http://www.fukuda-lab.org/mawilab/index.html.

the public release of the traces using a variety of attack labeling methods. MAWILab has been updated daily since its release until the present date.

In 1999, NSS Labs, an information security research and testing organization, pioneered third-party testing of IDSes with the publication of the first systematic, criteria-driven methodology for IDS testing. From 1999 until the present date, NSS Labs has been continuously supplying methodologies for testing IDSes to the public following trends in IDS design.[48] These methodologies may serve as guidelines for the rigorous testing of IDSes. For instance, in 2014, NSS Labs published a methodology for testing next-generation IDSes[49]—that is, IDSes designed to detect novel threats, such as advanced persistent threats and social media threats.

[48]Recent methodologies published by NSS Labs can be found at https://www.nsslabs.com/reports/categories/methodologies.

[49]https://www.nsslabs.com/reports/next-generation-intrusion-prevention-systems-ngips-test-methodology-v10.